An Adroit Stratified Unrelated Question Randomized Response Model using Neyman Allocation

Housila P. Singh and Tanveer A. Tarray*
School of Studies in Statistics,
Vikram University, Ujjain – 456010 – India.
Corresponding Author : tanveerstat@gmail.com

Received: 23, January 2014 / Revised: 19, February, 2014 / Accepted: 7, May 2014

ABSTRACT

In this paper we consider the use of stratified random sampling with Neyman allocation to Singh et al. (1994) unrelated question randomized response strategy for completely truthful reporting. It has been shown that for the prior information given, our new model is more efficient in terms of variance (in the case of completely truthful reporting) than Kim and Elam's (2007) model. Numerical illustrations and graphs are also given in support of the present study.

Keywords: Randomized response technique, Stratified random sampling, Estimation of proportion, Neyman allocation.

1. Introduction

Warner (1965) was the first to suggest an ingenious method to estimate the proportion of sensitive / stigmatizing character like induced abortion, drug abuse, homosexual activities, tax evasion etc. through a randomization device such as deck of cards, spinner etc. such that the respondent's privacy should be protected. Other developments in this technique are due to Greenberg et al. (1969). The unrelated question randomized response data – gathering device to procure trustworthy information on stigmatized characters was introduced by Horvitz et al. (1967). They avow better cooperation from the respondents as compared to Warner's (1965) original model. While developing theory for this model, Greenberg et al. (1969), considered both the situations when π_y, the proportion innocuous character (say) Y in population is known and when it is unknown, we shall call this model as UY- model. Some modifications in the randomized response (RR) model has been suggested by Chaudhuri and Mukerjee (1988, 2011), Mangat et al. (1992), Mangat and Singh (1990), Mangat

(1994), Grewal et al. (2005-2006), Singh and Mathur (2004), Perri (2008) and Singh and Tarray (2012, 2013,2014 a,b).

Hong et al. (1994) envisaged 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 country in India through a proportional sampling method may be very difficult compared to the researcher's time, effort and money.

To overcome this problem, Kim and Warde (2004) and Kim and Elam (2005, 2007) suggested 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 group or income group, for instance, see Kim and Elam (2007, p. 216).

A primary focus of this paper is the implementation of unrelated Stratified RR technique using Singh et al. (1994) unrelated question RR Strategy. In Section 2 we present our suggested model in the case where the proportion of respondents with the non sensitive trait in a stratum is known. In the subsection 2.2, we demonstrate the findings of empirical studies. The empirical studies show that, for the prior information given, the proposed model is more efficient in terms of variance than Kim and Elam's (2007) model. In Section 3 we offer some concluding remarks.

2. Suggested Model

In the proposed model, the population is partitioned into strata, and a sample is selected by simple random sampling with replacement (SRSWR) from each stratum. To get the full benefit from stratification, it is assumed 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 (deck of cards) R_i which consists of three types of cards bearing three statements:

(i) I belong to sensitive group A.
(ii) I belong to non - sensitive group Y, and
(iii) Draw one more card,

The statements are represented with proportions P_i, P_{1i} and P_{2i} respectively. The respondent is required to draw one card randomly from the above deck of cards and give answer in terms of "Yes" or "No" according to his / her actual status when the statement, (i) or (ii) drawn. However, if the statement (iii) is drawn, he / she is required to repeat the above process without replacing that card. If the statement (iii) is drawn in the second phase, he is directed to report "No" see Singh et al. (1994, p. 235). 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 $\sum_{i=1}^{k} n_i = n$. Under the assumption that these "Yes" and "No" reports are made truthfully and P_i is set by the researcher. If m_i is the total number of cards in the proposed deck in stratum i, then the probability X_i of a "Yes" answer in stratum i for this procedure is:

$$X_i = \left[\pi_{Si} P_i + P_{1i}\pi_{yi}\right]\left[1 + P_{2i}\left(m_i/(m_i - 1)\right)\right] \qquad \text{for } i = 1,2,\cdots,k; \qquad (1)$$

where π_{Si} is the proportion of people with sensitive traits in stratum i and π_{yi} is the proportion of people with the non-sensitive traits in stratum i.

Under the condition that π_{yi} is known, the unbiased estimator $\hat{\pi}_{Si}$ of π_{Si} is:

$$\hat{\pi}_{Si} = \frac{1}{P_i}\left[\frac{\hat{X}_i}{\left\{1 + P_{2i}\left(m_i/(m_i - 1)\right)\right\}} - P_{1i}\pi_{yi}\right] \qquad \text{for } i = 1,2,\cdots,k; \qquad (2)$$

where $\hat{X}_i$ is the proportion of "Yes" answer in the sample from stratum for i.

Since each $\hat{X}_i$ follows a binomial distribution $B(n_i, X_i)$, therefore the variance of the estimator $\hat{\pi}_{Si}$ is

$$V\left(\hat{\pi}_{Si} \mid \pi_{yi}\right) = \frac{X_i(1 - X_i)}{n_i\left\{1 + P_{2i}\left(m_i/(m_i - 1)\right)\right\}^2}. \qquad (3)$$

Since the selections in different strata are made independently, the estimators for individual strata can be added together to obtain an estimator for the entire population. Thus the unbiased estimator of π_S is

$$\hat{\pi}_S = \sum_{i=1}^{k} w_i \hat{\pi}_{Si} = \sum_{i=1}^{k} w_i \frac{1}{P_i}\left[\frac{\hat{X}_i}{\left\{1 + P_{2i}\left(m_i/(m_i - 1)\right)\right\}} - P_{1i}\pi_{yi}\right] \qquad (4)$$

The variance of the unbiased estimator $\hat{\pi}_S$ given π_{yi} is:

$$V\left(\hat{\pi}_S \mid \pi_{yi}\right)=\sum_{i=1}^{k} w_i^2 \frac{X_i\left(1-X_i\right)}{n_i\left\{1+P_{2i}\left(m_i/(m_i-1)\right)\right\}^2}. \tag{5}$$

Information on π_{Si} and π_{yi} are usually unavailable. But if prior information on π_{Si} and π_{yi} are available from the past experience then it helps to derive the following Neyman allocation formula.

Theorem 1

The Neyman allocation n to $n_1, n_2, \cdots, n_{k-1}$, and n_k to derive the minimum variance of $\hat{\pi}_S$ subject $n=\sum_{i=1}^{k} n_i$ is approximately given by

$$\frac{n_i}{n}=\frac{\dfrac{w_i\sqrt{X_i\left(1-X_i\right)}}{\left\{1+P_{2i}\left(m_i/(m_i-1)\right)\right\}}}{\sum_{i=1}^{k}\dfrac{w_i\sqrt{X_i\left(1-X_i\right)}}{\left\{1+P_{2i}\left(m_i/(m_i-1)\right)\right\}}}. \tag{6}$$

Proof: - Follows, for example, from section 5.5 of Cochran (1977) .

Putting (6) in (5) we get the minimum variance of the estimator $\hat{\pi}_S$ given π_{yi} is given by:

$$V\left(\hat{\pi}_S \mid \pi_{yi}\right)=\frac{1}{n}\left[\sum_{i=1}^{k}\frac{w_i\sqrt{X_i\left(1-X_i\right)}}{\left\{1+P_{2i}\left(m_i/(m_i-1)\right)\right\}}\right]^2. \tag{7}$$

An unbiased estimator of variance in (5) can be obtained by replacing n_i by $(n_i$-1).

Table 1: The percent relative efficiency of $\hat{\pi}_{ke}$ with respect to $\hat{\pi}_S$ when π_{yi} known, n= 1000, $P_2=P_{12}=P_{22}$ and $P_1=P_{11}=P_{21}$.

π_{S1}	π_{S2}	w_1	w_2	m_1	m_2	π_y	P_1= 0.3	P_1=0.4	P_1=0.5	P_1=0.6
							P_2 =0.1	P_2 =0.2	P_2 =0.3	P_2 =0.4
0.48	0.53	0.70	0.30	0.10	0.70	0.95	599.27	397.00	332.72	415.08
0.48	0.53	0.70	0.30	0.10	0.70	0.93	652.66	417.38	340.56	410.50
0.48	0.53	0.70	0.30	0.10	0.70	0.91	704.10	436.95	348.01	406.39
0.48	0.53	0.60	0.40	0.11	0.69	0.95	572.48	381.98	318.32	379.78
0.48	0.53	0.60	0.40	0.11	0.69	0.93	629.89	404.41	327.98	379.57
0.48	0.53	0.60	0.40	0.11	0.69	0.91	685.25	425.99	337.21	379.47
0.48	0.53	0.40	0.60	0.12	0.68	0.95	523.49	357.32	298.75	350.22
0.48	0.53	0.40	0.60	0.12	0.68	0.93	588.87	383.86	311.85	356.27
0.48	0.53	0.40	0.60	0.12	0.68	0.91	652.13	409.53	324.49	362.12
0.48	0.53	0.30	0.70	0.13	0.67	0.95	495.45	340.11	280.51	308.22
0.48	0.53	0.30	0.70	0.13	0.67	0.93	564.03	368.15	294.74	316.46
0.48	0.53	0.30	0.70	0.13	0.67	0.91	630.56	395.36	308.55	324.46
0.58	0.63	0.70	0.30	0.14	0.66	0.95	499.08	325.29	265.32	284.35
0.58	0.63	0.70	0.30	0.14	0.66	0.93	551.30	345.82	273.80	283.42
0.58	0.63	0.70	0.30	0.14	0.66	0.91	601.59	365.50	281.83	282.66
0.58	0.63	0.60	0.40	0.15	0.65	0.95	476.38	309.82	248.30	253.08
0.58	0.63	0.60	0.40	0.15	0.65	0.93	532.02	331.94	258.13	255.08
0.58	0.63	0.60	0.40	0.15	0.65	0.91	585.62	353.19	267.49	257.01
0.58	0.63	0.40	0.60	0.16	0.64	0.95	435.44	284.26	223.13	213.96
0.58	0.63	0.40	0.60	0.16	0.64	0.93	497.87	309.53	235.32	219.71
0.58	0.63	0.40	0.60	0.16	0.64	0.91	558.20	333.92	247.05	225.24
0.58	0.63	0.30	0.70	0.17	0.63	0.95	412.31	268.66	207.13	188.99
0.58	0.63	0.30	0.70	0.17	0.63	0.93	477.43	295.02	220.03	195.82
0.58	0.63	0.30	0.70	0.17	0.63	0.91	540.53	320.54	232.50	202.44
0.68	0.73	0.70	0.30	0.18	0.62	0.95	403.84	257.62	206.45	211.23
0.68	0.73	0.70	0.30	0.18	0.62	0.93	455.07	278.49	215.78	212.44
0.68	0.73	0.70	0.30	0.18	0.62	0.91	504.42	298.46	224.58	213.61
0.68	0.73	0.60	0.40	0.19	0.61	0.95	385.94	243.51	189.95	182.01
0.68	0.73	0.60	0.40	0.19	0.61	0.93	440.00	265.57	200.21	185.37
0.68	0.73	0.60	0.40	0.19	0.61	0.91	492.07	286.71	209.96	188.55
0.68	0.73	0.40	0.60	0.20	0.60	0.95	354.26	220.52	165.67	145.55
0.68	0.73	0.40	0.60	0.20	0.60	0.93	414.04	244.95	177.52	151.34
0.68	0.73	0.40	0.60	0.20	0.60	0.91	471.79	268.46	188.88	156.87
0.68	0.73	0.30	0.70	0.21	0.59	0.95	336.31	207.33	152.30	127.35
0.68	0.73	0.30	0.70	0.21	0.59	0.93	398.30	232.50	164.54	133.70
0.68	0.73	0.30	0.70	0.21	0.59	0.91	458.34	256.80	176.33	139.81

3. Efficiency comparison

In section 3, to have a tangible idea about the performance of the proposed RR model relative to Kim and Elam (2007) RR model, we have computed the percent relative efficiency of the proposed estimator $\hat{\pi}_S$ with respect to Kim and Elam (2007) estimator $\hat{\pi}_{ke}$ in the case when π_{yi} is known. The percent relative efficiency of the proposed estimator $\hat{\pi}_S$ with respect to Kim and Elam (2007) estimator $\hat{\pi}_{ke}$ is given by

$$PRE(\hat{\pi}_S, \hat{\pi}_{ke}) = \frac{V(\hat{\pi}_{ke})}{V(\hat{\pi}_S)} \times 100 \tag{8}$$

To look at the magnitude of the percent relative efficiency, we choose n = 1000, $k = 2$, $\pi_y = \pi_{y1} = \pi_{y2}$, $P_1 = P_{11} = P_{21}, P_2 = P_{12} = P_{22}$, $P_1+P_2=1$, $P_1 \neq P_2$ $m_1 \neq m_2$ and findings are shown in Table 1, where $V(\hat{\pi}_{ke})$ and $V(\hat{\pi}_S)$ are respectively given by (7) and Kim and Elam (2007, equation (3.7), p.223). Substantial gain in efficiency is observed when π_y is small. This fact is also depicted in Fig. 1. Since all the PRE values in Table 1 are greater than 100, our stratified unrelated question RR model using Neyman allocation is more efficient in terms of variance than Kim and Elam (2007) stratified RR model under the assumptions given and the prior information used.

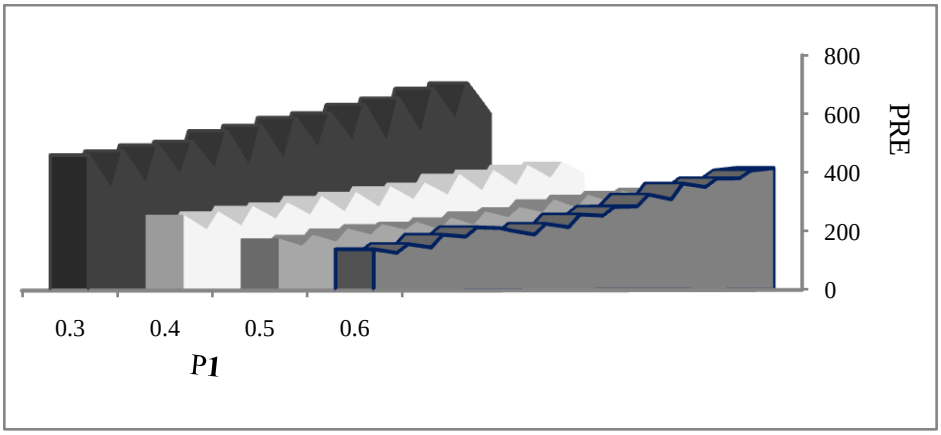

Figure 1: The percent relative efficiency of $\hat{\pi}_{ke}$ with respect to $\hat{\pi}_S$ $\pi_y (= \pi_{y1} = \pi_{y2})$ is known

4. Conclusion

The purpose of this paper is to estimate the proportion π_S of the population belonging to a sensitive group using RR technique in stratified unrelated question randomized response sampling. A stratified unrelated question RR model using Singh et al. (1994) procedure, an improved unrelated question RR model for completely truthful reporting has been proposed. It has been shown that for the prior information given, the proposed stratified unrelated question RR model using Neyman allocation is more efficient in terms of variance than Kim and Elam (2007) unrelated question stratified RR model. A notable point in this study is that the proposed model is better than the one earlier considered by Kim and Elam (2007).

Acknowledgements

The authors wish to thank the Associate Editor / Referee for their valuable suggestions / comments which helped to improve the presentation of the paper.

References

1. Chaudhuri A. and Mukerjee R. (1988). Randomized Response: Theory and Techniques. Marcel- Dekker, New York, USA.
2. Cochran W.G. (1977). Sampling Technique, 3rd Edition. New York: John Wiley and Sons, USA.
3. Greenberg B., Abul- Ela A., Simmons W.R. and Horvitz D.G. (1969). The unreleased question randomized response: Theoretical framework. *Jour. Amer. Statist. Assoc., 64,529-539.* DOI: 10.1080/01621459.1969.10500991
4. GrewalI.S, Bansal M.L. and Sidhu S.S. (2005–2006). Population mean corresponding to Horvitz–Thompson's estimator for multi-characteristics using randomized response technique. *Model Assist. Statist. Appl. 1, 215–220.*
5. Horvitz D.G. Shah B.V. and Simmons W.R. (1967). The unrelated question randomized response model. Proc. of Social Statistics Section. *Jour. Amer.Statist. Assoc., 65-72.*
6. Hong K., Yum J. and Lee H. (1994). A stratified randomized response technique. *Korean Jour. Appl. Statist., 7, 141-147.*
7. Kim J.M. and Elam M.E. (2005). A two–stage stratified Warner's randomized response model using optimal allocation. *Metrika., 61, 1-7.* DOI: 10.1007/s001840400319
8. Kim J.M. and Elam M.E. (2007). A stratified unrelated randomized response model, *Statistical Papers, 48, 215-233.* DOI: 10.1007/s00362-006-0327-6
9. Kim J.M. and Warde W.D. (2004). A stratified Warner randomized response model. *Jour. Statist. Plan. Inference, 120, 155-165.* DOI: 10.1016/S0378-3758(02)00500-1
10. Mangat N.S. Singh R. and Singh S. (1992). An improved unrelated question randomized response strategy. *Calcuta Statist. Assoc. Bull. 42, 277-81.*

11. Mangat, N.S. (1994). An improved randomized response strategy. *Jour. Roy. Statist. Soc., B, 56 (1), 93-95.*

12. Mangat, N.S. and Singh R. (1990). An alternative randomized procedure. *Biometrika,77, 439-442.* DOI: 10.1093/biomet/77.2.439

13. Perri P.F. (2008): Modified randomized devices for Simmons' model. *Model Assist. Statist. Appl., 3(3), 233-239.*

14. Singh S., Singh R., Mangat N.S. and Tracy D.S. (1994). An alternative device for randomized responses. *Statistica, anno, 54(2), 233-243.*

15. Singh H.P. and Mathur N. (2004). Unknown repeated trails in the unrelated question randomized response model. *Biometrical Jour., 46(3), 375-378.* DOI: 10.1002/bimj.200210032

16. Singh H.P. and Tarray T.A. (2012). A Stratified Unknown repeated trials in randomized response sampling. *Commun. Korean Statist. Soc., 19, (6), 751-759.* DOI: 10.5351/CKSS.2012.19.6.751

17. Singh H.P. and Tarray T.A. (2013). An alternative to Kim and Warde's mixed randomized response model. *Statist. Oper. Res. Trans. , 37 (2), 189-210.*

18. Singh H.P. and Tarray T.A. (2014 a). An alternative to stratified Kim and Warde's randomized response model using optimal (Neyman) allocation. *Model Assist.Statist. Appl., 9, 37-62.*

19. Singh H.P. and Tarray T.A. (2014 b). A dexterous randomized response model for estimating a rare sensitive attribute using Poisson distribution. *Statist. Prob. Lett.,90,42-45.* DOI: 10.1016/j.spl.2014.03.019

20. Warner S.L. (1965): Randomized response: A survey technique for eliminating evasive answer bias. *Jour. Amer. Statist. Assoc., 60, 63-69.* DOI: 10.1080/01621459.1965.10480775