World Applied Sciences Journal 38 (4): 324-328, 2020
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2020.324.328

A Proficient Optional Randomized Response Model for Estimating a Rare Sensitive Attribute Using Poisson Distribution

Tanveer A. Tarray, Zahoor Ahmad Ganie and Baziga Youssuf

Islamic University of Science and Technology - J & K, India - 192122

Abstract: The nitty-gritty of this paper is to estimating the mean of the number of persons possessing a rare sensitive attribute based on randomization device by utilizing the Poisson distribution in survey sampling. It is shown that the proposed model is more efficient than existing estimators when the proportion of persons possessing a rare unrelated attribute is known. Properties of the proposed randomized response model have been studied along with recommendations. Numerical illustrations and graphs are also given in support of the present study.

Key words: Randomized response technique · Estimation of proportion · Rare sensitive characteristics
AMS Subject Classification: 62D05.

INTRODUCTION

The growth of the internet makes it easy to perform data collection on a large scale. However, accompanying such benefits are concerns about information privacy. Because of these concerns, some people might decide to give false information in fear of privacy problem, or they might simply refuse to divulge any information at all. Data-analysis and knowledge-discovery tools using these data might therefore produce results with low accuracy or, even worse, produce false knowledge. Some people might be willing to selectively divulge information if they can get benefit in return. Examples of the benefit provided include discount of purchase, useful recommendations and information filtering. However, a significant number of people are not willing to divulge their information because of privacy concerns. [1] suggested an ingenious method of collecting information on sensitive characters. According to the method, for estimating the population proportion π possessing the sensitive character "A", a simple random with replacement sample of n persons is drawn from the population. Each interviewee in the sample is furnished an identical randomization device where the outcome "I possess character A" occurs with probability P while its complement "I do not possess character A" occurs with probability (1-P). The respondent answers "Yes" if the outcome of the randomization device tallies with his actual status otherwise he answers "No". Some modifications in the model has been suggested in [2-32].

[22] provided theoretical framework for a modification to the Warner's model proposed by [7]. The proposed method consisted in modifying the randomization device where the second outcome "I do not possess the character A" was replaced by the outcome "I possess the character Y"where "Y" was unrelated to character "A". This modified model is now known as 'unrelated question model, or U- model'.

In this paper we consider the problem where the number of persons possessing a rare sensitive attribute is very small and huge sample size is required to estimate this number. The capacity of our communication systems is increasing rapidly; so it should soon be possible to conduct such large randomized response surveys over the internet, by telephone, etc. We have discussed the situation when, the proportion of persons possessing a rare unrelated attributes is known in sectionsii. Properties of the proposed randomized response model have been studied along with recommendations. Efficiency comparison is worked out to investigate the performance of the suggested procedures. Numerical studies and graphical representations are worked out to demonstrate the superiority of the suggested model.

Estimation of a Rare Sensitive Attribute in Sampling-Known Rare Unrelated Attributes: Let π_1 be the true proportion of the rare sensitive attribute A_1 in the population Θ. For example, the proportion of AIDS/ HIV patients who continue having affairs with strangers; the

Corresponding Author: Tanveer A. Tarray, Islamic University of Science and Technology - J & K, India - 192122.

proportion of persons who have witnessed a murder; the proportion of persons who are told by the doctors that they will not survive long due to a ghastly disease. Consider selecting a large sample of n persons from the population such that as $n \to \infty$ and $\pi_1 \to 0$ then $n\,\pi_1 = \delta_1$ (finite). Let π_2 be the true proportion of the population having the rare unrelated attribute A_2 such that as $n \to \infty$ and $\pi_2 \to 0$ then $\pi_1 \to \delta_2$ (finite and known). For example, π_2 might be the proportion of persons who are born exactly at 12:00 o'clock; the proportion of babies born blind.

In the proposed procedure, a sample of size n is selected by simple random sampling with replacement from the population. In the first stage of the survey interview, an individual respondent in the sample is instructed to use the randomization device R_1 which consists of a rare sensitive question (A_1) card with probability P_1 and a non-sensitive question (A_2) card with probability 1-P_1. The respondent in the sample is free to answer the randomization question in terms of "Yes" and "No" either by using randomization device R_1 or without using it.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. Under the supposition that the "Yes" and "No" reports are made truthfully and T and P are set by the investigator, the probability of a "Yes" answer for this procedure is:

$$\begin{aligned}\theta_0 &= \pi_1 T + (1-T)[\pi_1 P + (1-P)\pi_2] \\ \theta_0 &= \pi_1\{T + (1-T)P\} + (1-T)(1-P)\pi_2 \end{aligned} \quad (1)$$

Note that both attributes A_1 and A_2 are very rare in population. As before, assuming that, as $n \to \infty$ and $\theta_0 \to 0$ such that $N\theta_0 = \delta_0$ (finite),
i.e.

$$\begin{aligned}\delta_0 &= \{T + (1-T)P\}\delta_1 + (1-T)(1-P)\delta_2 \\ &= T_1^*\delta_1 + T_2^*\delta_2\end{aligned}$$

where $T_1^* = T + (1-T)P \quad and \quad T_2^* = (1-T)(1-P)$

Let $y_1, y_2, \ldots, y_n$ be a random sample of n observations from the Poisson distribution with parameter δ_0. The likelihood function of the random sample of n observations is given by

$$L = \prod_{i=1}^{n} \frac{e^{-\delta_0}\delta_0^{y_i}}{y_i!}. \quad (2)$$

$$= \left(e^{-n\,\delta_0}\right)\prod_{i=1}^{n}\delta_0^{y_i}\prod_{i=1}^{n}\frac{1}{y_i!} = \left(e^{-n\,\delta_0}\right)\ \delta_0^{\sum_{i=1}^{n} y_i}\ \prod_{i=1}^{n}\frac{1}{y_i!}$$

Taking natural logarithm on both sides of (2) we have

$$L_n = (-n\delta_0) + (\sum_{i=1}^{n} y_i)\log\ \delta_0 + \sum_{i=1}^{n}\ \log\ \frac{1}{y_i!}$$

$$L_n = -n\{T_1^*\delta_1 + T_2^*\delta_2\} + (\sum_{i=1}^{n} y_i)\log\{T_1^*\delta_1 + T_2^*\delta_2\} - \sum_{i=1}^{n}\ \log y_i!$$

On putting $\frac{\partial L}{\partial \delta_1} = 0$

The maximum-likelihood estimator of δ_1 is given by

$$\hat{\delta}_1\ = \frac{1}{T_1^*}\left[\frac{1}{n}\sum_{i=1}^{n} y_i - T_2^*\delta_2\right]$$

Thus, we have the following theorem.

Theorem 2.1: The estimator $\hat{\delta}_1$ is an unbiased estimator of the parameter $\hat{\delta}_1$.

Proof: Since $y_i \sim P(\delta_0)$, that is, y_i follows a Poisson distribution with parameter $\delta_0 = T_1^*\delta_1 + T_2^*\delta_2$, we have

$$E(\hat{\delta}_1) = \frac{1}{T_1^*}\left[\frac{1}{n}\sum_{i=1}^{n} E(y_i) - T_2^*\delta_2\right] = \frac{1}{T_1^*}\left[\frac{1}{n}\sum_{i=1}^{n}\delta_0 - T_2^*\delta_2\right]$$

$$= \frac{1}{T_1^*}\left[\delta_0 - T_2^*\delta_2\right] = \frac{1}{T_1^*}\left[T_1^*\delta_1 + T_2^*\delta_2 - T_2^*\delta_2\right] = \delta_1$$

which proves the theorem.

Theorem 2.2: The variance of the estimator $\hat{\delta}_1$ is given by

$$V(\hat{\delta}_1) = \frac{\delta_1}{nT_1^*} + \frac{T_2^*\delta_2}{nT_1^{*2}}.$$

Proof: Since $y_i \sim P(\delta_0)$, that is, y_i follows a Poisson distribution with parameter $\delta_0 = T_1^*\delta_1 + T_2^*\delta_2$, we have

$$V(\hat{\delta}_1) = V\left\{\frac{1}{T_1^*}\left[\frac{1}{n}\sum^{n}(y_i)\right]\right\} = \frac{1}{T_1^{*2}}\left[\frac{1}{n^2}\sum_{i=1}^{n} V(y_i)\right] = \frac{1}{T_1^{*2}}\left[\frac{1}{n^2}\sum_{i=1}^{n}\delta_0\right]$$

$$= \frac{\delta_1 T_1^* + T_2^*\delta_2}{nT_1^{*2}} = \frac{\delta_1}{nT_1^*} + \frac{T_2^*\delta_2}{nT_1^{*2}}.$$

Hence the theorem.

Table 1: The percent relative efficiency of the proposed estimator $\hat{\delta}_1$ with respect to the estimator $\hat{\delta}_L$

			P				
π_1	π_2	T	0.4	0.45	0.5	0.55	0.6
0.50	0.50	0.60	361.00	300.44	256.00	222.28	196.00
0.48	0.52	0.60	371.62	308.56	262.30	227.21	199.87
0.46	0.54	0.60	382.70	317.05	268.91	232.40	203.96
0.44	0.56	0.60	394.27	325.95	275.86	237.88	208.29
0.42	0.58	0.60	406.36	335.29	283.19	243.68	212.89
0.40	0.60	0.60	419.02	345.11	290.91	249.81	217.78
0.38	0.62	0.60	432.28	355.42	299.07	256.32	222.98
0.36	0.64	0.60	446.18	366.29	307.69	263.23	228.54
0.34	0.66	0.80	680.47	537.95	435.48	359.26	300.96
0.32	0.68	0.80	714.27	563.46	455.06	374.43	312.76
0.30	0.70	0.80	751.03	591.29	476.47	391.07	325.76
0.28	0.72	0.80	791.15	621.76	500.00	409.43	340.14
0.26	0.74	0.80	835.11	655.27	525.97	429.77	356.13
0.24	0.76	0.80	883.49	692.30	554.79	452.44	374.04
0.22	0.78	0.80	936.99	733.43	586.96	477.85	394.22
0.20	0.80	0.80	996.47	779.38	623.08	506.55	417.13

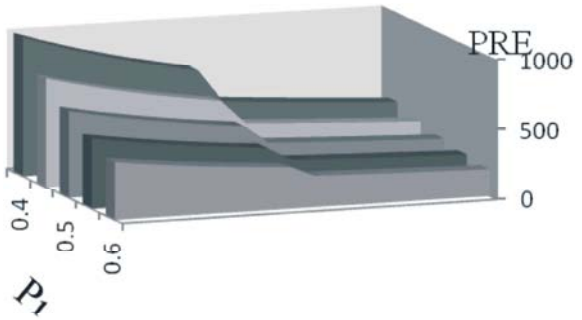

Fig. 1: The percent relative efficiency of the proposed estimator

Theorem 2.3: An unbiased estimator of the variance of the estimator $\hat{\delta}_1$ is

$$\hat{\nu}(\hat{\delta}_1) = \frac{1}{n^2 T_1^{*2}} \sum_{i=1}^{n} (y_i) \tag{3}$$

Proof: Taking expectation of both sides of (3), we have

$$E[\,\hat{\nu}(\hat{\delta}_1)\,] = \frac{1}{n^2 T_1^{*2}} E\left[\sum_{i=1}^{n} (y_i)\right] = \frac{1}{n^2 T_1^{*2}} \left[\sum_{i=1}^{n} E\,(y_i)\right]$$

$$= \frac{1}{n^2 T_1^{*2}} \left[\sum_{i=1}^{n} \delta_0\right] = \frac{\delta_1 T_1^{*} + T_2^{*}\delta_2}{n T_1^{*2}} = \frac{\delta_1}{n T_1^{*}} + \frac{T_2^{*}\delta_2}{n T_1^{*2}}.$$

which proves the theorem.

Relative Efficiency: The percent relative efficiency of the proposed estimator $\hat{\delta}_1$ with respect to the estimator $\hat{\delta}_L$ is given by

$$PRE(\hat{\delta}_1, \hat{\delta}_L) = \frac{V(\hat{\delta}_L)}{V(\hat{\delta}_1)} = \frac{[P\pi_1 + (1-P)\pi_2]T_1^{*2}}{[T_1^{*}\pi_1 + T_2^{*}\pi_2]P^2} \times 100\,, \tag{4}$$

where

$V(\hat{\delta}_L)$ with $n\pi_1 = \delta_1$ and $n\pi_2 = \delta_2$.

From Equation (4), it is clear that the percent relative efficiency of the proposed estimator is free from the sample size n. To look at the magnitude of the percent relative efficiency, we choose P from 0.4 to 0.6 and T = 0.6, 0.8 and compute the percent relative efficiency using the formula (4) and findings are shown in Table 1. Table-1 exhibits that the percent efficiency is greater than 100 which follows that the proposed procedure is better. Substantial gain in efficiency is observed when P is very small. This fact is also depicted in Fig. 1.

CONCLUSION

This paper premeditated the problem where the number of persons possessing a rare sensitive attribute is very small and huge sample size is required to estimate. We have discussed the situation when the proportion of persons possessing a rare unrelated attributes is known. Properties of the proposed randomized response model have been studied along with recommendations. Efficiency comparison is worked out to investigate the

performance of the suggested procedures. It is interesting to mention that the proposed procedure is superior to the one recently envisaged estimator.

ACKNOWLEDGEMENTS

The authors are thankful to the Editor-in- Chief and the learned referee for his valuable suggestions regarding improvement of the paper.

REFERENCE

1. Warner, S.L., 1965. Randomized response: A survey technique for eliminating evasive answer bias. Jour. Amer. Statist. Assoc., 60: 63-69.
2. Zaizai Y., W. Jingyu, L. Junfeng and W. Hua, 2008. Ratio imputation method for handling item-nonresponse in Eichhorn model. Model Assist. Stat. Appl., 3(2): 89-98.
3. Cochran, W.G., 1977. Sampling Technique, 3rd Edition. New York: John Wiley and Sons, USA.
4. Fox, J.A. and P.E. Tracy, 1986. Randomized Response: A method of sensitive surveys. Newbury Park, CA: SEGE Publications.
5. Grewal, I.S., M.L. Bansal and S.S. Sidhu, 2005-2006. Population mean corresponding to Horvitz-Thompson's estimator for multi- characteristics using randomized response technique. Model Assist. Stat. Appl., 1: 215-220.
6. Hong, Z., 2005-2006. Estimation of mean in randomized response surveys when answers are incompletely truthful. Model Assist. Stat. Appl., 1: 221- 230.
7. Harvitz, D.G., B.V. Shah and W.R. Simmons, 1969. The unrelated question randomized response model. Proceedings of social statistics section. J. Amer. Statist. Assoc., pp: 65-72.
8. Lanke, J., 1975. On the choice of the unrelated question in simmonsversion of randomized response. Jour. Amer. Statist. Assoc., 70: 80-83.
9. Mahmood, M., S. Singh and S. Horn, 1998. On the confidentiality guaranteed under randomized response sampling: a comparison with several new techniques. Biom. Jour., 40: 237-242.
10. Mangat, N.S., 1991. An optional randomized response sampling technique using non-stigmatized attribute. Statistica, Anno, 51(4): 595-602.
11. Mahajan, P.K., 2005-2006. Optimum stratification for scrambled response with ratio and regression methods of estimation. Model Assist. Stat. Appl., 1: 17-22.
12. Mahajan, P.K., P. Sharma and R.K. Gupta, 2007. Optimum stratification for allocation proportional to strata totals for scrambled response. Model Assist. Stat. Appl., 2(2): 81-88.
13. Sidhu, S.S. and M.L. Bansal, 2008. Estimator of population total using Rao, Hartley and Cochran's scheme using optional randomized response technique in multi- character surveys. Model Assist. Stat. Appl., 3(3): 259-267.
14. Mangat, N.S., R. Singh and S. Singh, 1992. An improved unrelated question randomized response strategy. Cal.Statist. Assoc. Bull., pp: 277-281.
15. Singh, S., R. Singh, N.S. Mangat and D.S. Tracy, 1994. An alternative device for randomized responses. Statistica, Anno., 54(2): 233-243.
16. Singh, H.P., M.K. Srivastava, N. Srivastava, T.A. Tarray, V. Singh and S. Dixit, 2015. Chain regression-type estimator using multiple auxiliary information in successive sampling. Hacettepe Journal of Mathematics and Statistics, 44(5): 1247-1256.
17. Tarray, T.A. and H.P. Singh, 2015. A randomized response model for estimating a rare sensitive attribute in stratified sampling using Poisson distribution. Model Assist. Statist. Appl., 10: 345-360.
18. Tarray, T.A. and H.P. Singh, 2018. Missing data in clinical trials: stratified Singh and Grewal's randomized response model using geometric distribution. Trends in Bioinformatics, 11(1): 44-55.
19. Tarray, T.A., H.P. Sigh and Saadia Masood, 2019. An Endowed Randomized Response Model for Estimating a Rare Sensitive Attribute Using Poisson Distribution. Trends in Applied Sciences Research, 12: 1-6.
20. Singh, R. and N.S. Mangat, 1996. Elements of Survey Sampling, Kluwer Academic Publishers, Dordrecht, The Netherlands.
21. Singh, S., 2003. Advanced sampling theory with applications. Kluwer Academic Publishers, Dordrecht.
22. Mangat, N.S., 1994. An improved randomized response strategy. Jour. Roy. Statist. Soc., B, 56(1): 93-95.
23. Mangat, N.S. and R. Singh, 1990. An alternative randomized procedure. Biometrika, 77: 439-442.
24. Moors, J.A., 1971. Optimization of the unrelated question randomized response model. Jour. Amer. Statist. Assoc., 66: 627-629.
25. Singh, H.P. and T.A. Tarray, 2012. A Stratified Unknown repeated trials in randomized response sampling. Common. Korean Statist. Soc., 19(6): 751-759.

26. Singh, H.P. and T.A. Tarray, 2013. An alternative to Kim and Warde'smixed randomized response model. Statist. Oper. Res. Trans.(SORT), 37(2): 189-210.
27. Singh, H.P. and T.A. Tarray, 2014. An Improvement Over Kim and Elam Stratified Unrelated Question Randomized Response Model Using Neyman Allocation. Sankhya-B, The Ind. Jour. Statist., DOI 10.1007/s13571-014-0088-5.
28. Tarray, T.A. and H.P. Singh, 2015. Some improved additive randomized response models utilizing higher order moments ratios of scrambling variable. Model Assist. Stat. Appl., 10(4): 361-383.
29. Tarray, T.A. and H.P. Singh, 2017. A stratified unrelated question randomized response model using Neyman allocation, Comm. Stat. - Theory and Methods, 46(1): 17-27.
30. Tarray, T.A., H.P. Singh and Y. Zaizai, 2015. A stratified optional randomized response model. Sociological Methods and Research. DOI: 10.1177/0049124115605332, 1-15.
31. Singh, S., S. Horn, R. Singh and N.S. Mangat, 2003. On the use of modified randomization device for estimating the prevalence of a sensitive attribute. Statist. Trans., 6(4): 515-522.
32. Tracy, D.S. and N.S. Mangat, 1996. Some developments in randomized response sampling during the last decade-A follow up of review by Chaudhuri and Mukherjee. Jour. Appl. Statist. Sci., 4(2/3): 147-158.