Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666
Vol. 8, Issue 2 (2015): 69-76

AN ADROIT SINGH AND MATHUR'S RANDOMIZATION DEVICE FOR ESTIMATING A RARE SENSITIVE ATTRIBUTE USING POISSON DISTRIBUTION

Tanveer Ahmad Tarray[1] and Housila Prasad Singh[2]
[1]Department of Computer Science and Engineering , Islamic University of Science and Technology, Awantipora, Pulwama, Kashmir, India
[2]School of Studies in Statistics, Vikram University, Ujjain, India.
E Mail: [1] tanveerstat@gmail.com; [2] hpsujn@gmail.com

Received February 04, 2015
Modified November 09, 2015
Accepted November 25, 2015

Abstract

This paper presents the problem of estimating the mean of the number of persons possessing a rare sensitive attribute based on Singh and Mathur (2004) randomization device by utilizing the Poisson distribution in survey sampling. Properties of the proposed randomized response model have been studied. It is also shown that the proposed model is more efficient than Land et al. (2011) when the proportion of persons possessing a rare unrelated attribute is known. Numerical illustration is also given in support of the present study.

Key Words: Randomized Response Technique. Estimation Of Proportion, Rare Sensitive Attributes, Variance, Simple Random Sampling.

1. Introduction

The randomized response technique to procure trustworthy data for estimating the proportion of a population possessing a sensitive attribute "A" (say) was first introduced by Warner (1965). This model considers simple random sampling design. It requires the interviewee to give a "Yes" or "No" answers either to the sensitive question or to its negative depending on the outcome of a randomizing device not reported to the interviewer. This pioneering work of Warner's (1965) led to modifications and developments in various directions. Greenberg et al. (1969) felt that, to protect the privacy of respondents, it is desirable that the two questions be unrelated and suggested an unrelated question randomized response model. In Greenberg et al.'s (1969) unrelated question model, the data – gathering randomization device consists of two questions: (i) Are you a member of group "A"? (ii) Are you a member of group "Y"?, where the characteristic "Y" or its complement or innocuous and unrelated to "A". For instance in estimating the proportion of persons having extramarital relations in a certain community, the two questions may be : (a) Are you having extramarital relations? (b) Were you born in the month of March? Evidently, the second question has nothing to do with extramarital relations. Greenberg et al. (1969) in their theoretical development, dealt with two situations involving π_y (the proportion of persons with unrelated character, Y): that where it is known and that where it is unknown. Greenberg et al. (1969) suggested that one of optimal choices close to zero or one according as $\pi_y < 0.5$or $\pi_y > 0.5$. Since the work by Warner (1965), a large amount of literature has emerged on the use and construction of various randomization devices to estimate the population proportion of a sensitive attribute in survey sampling. For example one could refer to Fox and Tracy (1986), Tracy and Mangat (1995), Kim and Elam (2005), Singh and Tarray (2012, 2014 a,b,c,d,e,f,g,h, 2015), the papers by Tarray and Singh (2014, 2015) and Tarray et al. (2015). For the sake of completeness and convenience to the readers, we have given the descriptions of Land et al. (2011) model.

1.1 Land et al. (2011) randomized response model

Land et al. (2011) envisaged an estimation of problem where the number of persons possessing a rare sensitive attribute is small and large sample size is needed to estimate this number. Suppose π_1 is the true proportion of the rare sensitive attribute A_1 in the populationΩ. Consider selecting a large sample of n persons from the population such that as $n \rightarrow \infty$ and $\pi_1 \rightarrow 0$ 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 $n \rightarrow \infty$ and $\pi_2 \rightarrow 0$ then $n\pi_2 = \delta_2$ (finite and known).

Each respondent selected in the sample is instructed to say "Yes" if he belongs to the rare sensitive attribute A_1 and if he is not in group A_1 then he / she is requested to rotate a spinner bearing two types of statements:

(a) Do you possess the rare sensitive attribute A_1?

and

(b) Do you possess the rare unrelated attribute A_2?

with probabilities P_1 and $(1-P_1)$ respectively. Thus, , the probability of a "Yes" answer is given by

$$\theta_0 = P_1\pi_1 + (1-P_1)\,\pi_2 \tag{1}$$

Note that both attributes A_1 and A_2 are very rare in the population. Assuming that as $n \rightarrow \infty$ and $\theta_0 \rightarrow 0$ such that $n\theta_0 = \delta_0$(finite). Thus

$$\delta_0 = P_1\delta_1 + (1-P_1)\,\delta_2 \tag{2}$$

Based on a random sample of size n from the Poisson distribution with parameterδ_0 , Land et al. (2011) obtained the maximum likelihood estimator of the parameter δ_1 as:

$$\hat{\delta}_L = \frac{1}{P_1}\left[\frac{1}{n}\sum_{i=1}^{n} y_i - (1-P_1)\delta_2\right], \tag{3}$$

whose variance is given by

$$V(\hat{\delta}_L) = \frac{\delta_1}{nP_1} + \frac{(1-P_1)\delta_2}{nP_1^2}. \tag{4}$$

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. The study is carried out when the proportion of persons possessing a rare unrelated attributes is known in sections 2. Properties of the proposed randomized response model have been studied. In section 2.1, the efficiency comparison is worked out to investigate the performance of the suggested procedures.

2. Suggested estimator 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 Ω. For example, the proportion of AIDS patients who continue having affairs with strangers; the 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, for more examples see Land et al. (2011). Consider selecting a large sample of n persons from the population such that as $n \rightarrow \infty$ and $\pi_1 \rightarrow 0$ 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 $n \rightarrow \infty$ and $\pi_2 \rightarrow 0$ then $n\pi_2 = \delta_2$ (finite and known). For example, π_2 might be the proportion of persons who are born between 12:00 and 12:01 or 12:05 O'clock; the proportion of babies born blind; see Land et al. (2011). If a respondent belongs to the rare sensitive attribute A_1 , then he /she is requested to repeat the trial in the Greenberg et al. (1696) randomization device (i.e. U-model) if in the first trial he /she doesn't get the statement according to his /her status. The rest of the procedure remains the same. The repetition of the trial is known to the interviewee but remains unknown to the interviewer, see Singh and Mathur (2004). The privacy of the respondents possessing the sensitive attribute is protected in the proposed procedure. Assuming completely truthful reporting by the respondents, the probability of "Yes" answer is given by

$$\theta_0^* = [\pi_1\{P_1 + (1-P_1)P_1\} + (1-P_1)\,\pi_2] \tag{5}$$

Note that both attributes A_1 and A_2 are very rare in population. Letting that, as $n \rightarrow \infty$ and $\theta_0^* \rightarrow 0$ such that $n\theta_0^* = \delta_0^*$ (finite),

i.e.

$$\begin{aligned}\delta_0^* &= [\delta_1\{P_1 + (1-P_1)P_1\} + (1-P_1)\,\delta_2]\\ &= T_1^*\delta_1 + T_2^*\delta_2,\end{aligned}$$

where $T_1^* = \{P_1 + (1-P_1)P_1\}$ and $T_2^* = (1-P)$.

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

$$L \;=\; \prod_{i=1}^{n} \frac{e^{-\delta_0^*}\delta_0^{*y_i}}{y_i!}. \tag{6}$$

$$= \left(e^{-n\,\delta_0^*}\right)\prod_{i=1}^{n}\delta_0{}^{*y_i}\prod_{i=1}^{n}\frac{1}{y_i!} \;=\left(e^{-n\,\delta_0^*}\right)\,\delta_0^{*\sum\limits_{i=1}^{n} y_i}\;\prod_{i=1}^{n}\frac{1}{y_i!}.$$

Taking natural logarithm on both sides of (6) we have

$$\text{Log L} = \left(-n\delta_0^*\right) + \left(\sum_{i=1}^{n} y_i\right)\log\delta_0^* \;+\sum_{i=1}^{n}\log\frac{1}{y_i!}$$

or

$$\text{Log L} = -\,n\{T_1^*\delta_1 + T_2^*\delta_2\} + \left(\sum_{i=1}^{n} y_i\right)\log\{T_1^*\delta_1 + T_2^*\delta_2\} - \sum_{i=1}^{n}\log y_i! \tag{7}$$

Differentiating (7) partially with respect to δ_1 and equating to zero, we get the maximum – likelihood estimator of δ_1 as

$$\hat{\delta}_1 \;= \frac{1}{T_1^*}\left[\frac{1}{n}\sum_{i=1}^{n} y_i - T_2^*\delta_2\right] \tag{8}$$

Thus, we have the following theorem.

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

$$V(\hat{\delta}_1) = \frac{\delta_1}{nT_1^*} + \frac{T_2^*\delta_2}{nT_1^{*2}}. \tag{9}$$

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

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

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

Hence the theorem.

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

$$\hat{v}(\hat{\delta}_1) = \frac{1}{n^2 T_1^{*2}}\sum_{i=1}^{n} y_i \tag{10}$$

Proof. Taking expectation of both sides of (10), we have

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

$$= \frac{1}{n^2 T_1^{*2}}\left[\sum_{i=1}^{n}\delta_0^*\right] = \frac{\delta_1 T_1^* + T_2^*\delta_2}{nT_1^{*2}} = \frac{\delta_1}{nT_1^*} + \frac{T_2^*\delta_2}{nT_1^{*2}}. \tag{11}$$

which proves the theorem.

2.1 Comparison with Land et al. (2011) estimator

From (3) and (7), we have

$$V(\hat{\delta}_L) - V(\hat{\delta}_1) = \frac{\delta_1}{nP_1} + \frac{(1-P_1)\delta_2}{nP_1^2} - \frac{\delta_1}{nT_1^*} - \frac{T_2^*\delta_2}{nT_1^{*2}}.$$

$$= \frac{\delta_1}{n}\left[\left(\frac{1}{P_1} - \frac{1}{T_1^*}\right) + \frac{\delta_2}{n}\left(\frac{(1-P_1)}{P_1^2} - \frac{T_2^*}{T_1^{*2}}\right)\right]$$

$$= \frac{(1-P_1)}{nP_1}\left[\frac{\delta_1}{(2-P_1)} + \frac{(1-P_1)(3-P_1)\delta_2}{(2-P_1)^2}\right] > 0 \tag{12}$$

which shows that proposed estimator $\hat{\delta}_1$ is always better than Land et al. (2011) estimator $\hat{\delta}_L$.

2.2 Relative Efficiency

The percent relative efficiency of the proposed estimator $\hat{\delta}_1$ with respect to the Land et al. (2011) estimator $\hat{\delta}_L$ is given by

$$\mathrm{PRE}(\hat{\delta}_1,\hat{\delta}_L)=\frac{V(\hat{\delta}_L)}{V(\hat{\delta}_1)}=\frac{[P_1\delta_1+(1-P_1)\delta_2]T_1^{*^2}}{[T_1^*\delta_1+T_2^*\delta_2]P_1^2}\times 100\,, \tag{13}$$

From Equation (13), 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 chose different values of P_1. Table 1 exhibits that the percent relative efficiency is greater than 100 which follow that the proposed procedure is better than that of Land et al. (2011). Substantial gain in efficiency is observed when P_1 is very small.

δ_1	δ_2		P_1						
		T_1	0.1	0.15	0.2	0.25	0.3	0.35	0.4
0.50	0.50	0.60	950.00	616.67	450.00	350.00	283.33	235.71	200.00
0.48	0.52	0.60	980.40	633.93	460.80	357.00	287.87	238.54	201.60
0.46	0.54	0.60	1010.80	651.20	471.60	364.00	292.40	241.37	203.20
0.44	0.56	0.60	1041.20	668.47	482.40	371.00	296.93	244.20	204.80
0.42	0.58	0.60	1071.60	685.73	493.20	378.00	301.47	247.03	206.40
0.40	0.60	0.60	1102.00	703.00	504.00	385.00	306.00	249.86	208.00
0.38	0.62	0.60	1132.40	720.27	514.80	392.00	310.53	252.69	209.60
0.36	0.64	0.60	1162.80	737.53	525.60	399.00	315.07	255.51	211.20
0.34	0.66	0.80	1193.20	754.80	536.40	406.00	319.60	258.34	212.80
0.32	0.68	0.80	1223.60	772.07	547.20	413.00	324.13	261.17	214.40
0.30	0.70	0.80	1254.00	789.33	558.00	420.00	328.67	264.00	216.00
0.28	0.72	0.80	1284.40	806.60	568.80	427.00	333.20	266.83	217.60
0.26	0.74	0.80	1314.80	823.87	579.60	434.00	337.73	269.66	219.20
0.24	0.76	0.80	1345.20	841.13	590.40	441.00	342.27	272.49	220.80
0.22	0.78	0.80	1375.60	858.40	601.20	448.00	346.80	275.31	222.40
0.20	0.80	0.80	1406.00	875.67	612.00	455.00	351.33	278.14	224.00

Table 1: The percent relative efficiency of the proposed estimator $\hat{\delta}_1$ with respect to Land et al. (2011) estimator $\hat{\delta}_L$.

5. Conclusion

This paper discusses 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 developed a method to estimate the mean of the number of persons possessing a rare sensitive attribute utilizing the Poisson distribution in survey sampling when the proportion of persons possessing a rare unrelated attributes is known. Properties of the proposed randomized response model have been studied. The proposed procedure has been compared with that of Land et al. (2011) both theoretically and empirically. It is interesting to mention that the proposed procedure using Poisson distribution is superior to the one recently envisaged by Land et al. (2011) both theoretically and empirically.

Acknowledgement

Authors are thankful to the Editor–in-Chief and the unknown learned referees for their valuable suggestions regarding improvement of the paper.

References

1. Fox , J.A. and Tracy, P.E. (1986). Randomized Response: A Method of Sensitive Surveys, SEGE Publications Newbury Park, CA.
2. 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, p. 529-539.
3. Kim, J.M. and Elam, M.E. (2005). A two – stage stratified Warner's randomized response model using Neyman allocation, Metrika, 61, p. 1-7.
4. Land, M., Singh, S. and Sedory, S.A. (2011). Estimation of a rare attribute using Poisson distribution, Statistics, 46(3), p. 351-360.
5. Singh, H.P. and Mathur, N. (2004). Unknown repeated trials in the unrelated question randomized response model, Biometrical Jour., 46(3), p. 375-378.
6. Singh, H.P. and Tarray, T.A. (2012). A stratified unknown repeated trials in randomized response sampling, Comm. Kor. Statist. Soc., 19(6), p. 751-759**.**
7. 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, p. 37-62.
8. Singh, H.P. and Tarray, T.A. (2014 b). An improvement over Kim and Elam stratified unrelated question randomized response model using Neyman allocation, Sankhya – B, DOI : 10.1007/s13571-014-0088-5.
9. Singh, H.P. and Tarray, T.A. (2014 c). An adroit stratified unrelated question randomized response model using Neyman allocation. Sri. Jour. Appl. Statist., 15(2), p. 83-90.
10. Singh, H.P. and Tarray, T.A. (2014 d). An alternative to Kim and Warde's mixed randomized response model. Statist, Oper. Res. Trans., 37 (2), p. 189-210.
11. Singh, H.P. and Tarray, T.A. (2014 e). A dexterous randomized response model for estimating a rare sensitive attribute using Poisson distribution, Statist. Prob. Lett., 90, p. 42-45.
12. Singh, H.P. and Tarray, T.A. (2014 f). An efficient alternative mixed randomized response procedure, Soc. Meth. Res. ,DOI: 10.1177/ 0049124114553309.
13. Singh, H.P. and Tarray, T.A.. (2014 g). An alternative to Kim and Warde's mixed randomized response technique, Statistica, Anno, 73(3), p. 379 - 402.
14. Singh, H.P. and Tarray, T.A. (2014 h). An alternative estimator in stratified RR strategies, Jour. Reli. Statist. Stud., 7(2), p. 105-118.

15. Singh, H.P. and Tarray, T.A. (2015). A revisit to the Singh, Horn, Singh and Mangat's randomization device for estimating a rare sensitive attribute using Poisson distribution, Model Assist. Statist. Appl., 10, p. 129-138.
16. Tarray, T.A. and Singh H.P. (2014). A Proficient randomized response model. Istatistika: Jour. Turkey Statist. Assoc.,7(3), p. 87-98.
17. Tarray, T.A. and Singh, H.P. (2015). A general procedure for estimating the mean of a sensitive variable using auxiliary information, Investigacion Operacionel, 36(3), p. 249-262.
18. Tarray, T.A., Singh, H.P. and Zaizai, Y. (2015). A stratified optional randomized response model, Socio. Meth. Res. DOI: 10.1177/ 0049124115605332, 1-15.
19. Tracy, D.S. and Mangat, N.S. (1995): A partial randomized response strategy, Test, 4(2), p. 315 – 321.
20. Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. Jour, Ameri. Statist. Assoc., 60, p. 63-69.