Open Journal of Statistics, 2016, 6, 85-95
Published Online February 2016 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2016.61011

Improved Estimation of Rare Sensitive Attribute in a Stratified Sampling Using Poisson Distribution

Abdul Wakeel, Masood Anwar

Department of Mathematics, COMSATS Institute of Information Technology, Islamabad, Pakistan
Email: chabdulwakeel@gmail.com

Received 21 December 2015; accepted 20 February 2016; published 23 February 2016

Abstract

In this study, we propose a two stage randomized response model. Improved unbiased estimators of the mean number of persons possessing a rare sensitive attribute under two different situations are proposed. The proposed estimators are evaluated using a relative efficiency comparison. It is shown that our estimators are efficient as compared to existing estimators when the parameter of rare unrelated attribute is known and in unknown case, depending on the probability of selecting a question.

Keywords

Poisson Distribution, Rare Sensitive Attribute, Rare Unrelated Attribute, Stratified Sampling

1. Introduction

The collection of data through direct questioning on rare sensitive issues such as extramarital affairs, family disturbances and declaring religious affiliation in extremism condition is far-reaching issue. Warner [1] introduced the randomized response procedure to procure trustworthy data for estimating π, the proportion of respondents in the population belonging to the sensitive group. Greenberg *et al*. [2] suggested an unrelated question randomized response model in which each individual selected in the samples was asked to reply "*yes*" or "*no*" to one of two statements: (a) Do you belong to Group A? (b) Do you belong to Group Y? with respective probabilities P and $(1-P)$. Second question asked in the sampling does not have any effect on the first question. Greenberg *et al*. [2] considered π_A and π_Y the proportion of persons possessing sensitive and unrelated characteristic respectively and discussed both the cases when π_Y was known and unknown. The probability of *yes*

How to cite this paper: Wakeel, A. and Anwar, M. (2016) Improved Estimation of Rare Sensitive Attribute in a Stratified Sampling Using Poisson Distribution. *Open Journal of Statistics*, **6**, 85-95. http://dx.doi.org/10.4236/ojs.2016.61011

responses θ_0, defined by them is $\theta_0 = P\pi_A + (1-P)\pi_Y$. Mangat and Singh [3] proposed a two stage randomized response procedure which required the use of two randomization devices. The random device R_1 consists of two statements namely (a) I belong to the sensitive group, and (b) Go to random device R_2, with probabilities T and $(1-T)$ respectively. The random device R_2 which uses two statements (a) I belong to the sensitive group, and (b) I do not belong to the sensitive group with known probabilities P and $(1-P)$ respectively. Then θ_0, the probability of *yes* responses is $\theta_0 = T\pi + (1-T)\{P\pi + (1-P)(1-\pi)\}$.

Later on, different modifications have been made to improve the methodology for collection of information. Some of them are Lee *et al*. [4], Chaudhuri and Mukerjee [5], Mahmood *et al*. [6], Land *et al*. [7], Bhargava and Singh [8].

Land *et al*. [7] proposed the estimators for the mean number of persons possessing the rare sensitive attribute using the unrelated question randomized response model by utilizing a Poisson distribution. Recently, Lee *et al*. [4] extended the Land *et al*.'s [7] study to stratify sampling and propose the estimators when the parameter of rare unrelated attribute is known and unknown.

In this study, we propose improved estimators for the mean and its variance of the number of persons possessing a rare sensitive attribute based on stratified sampling by using Poisson distribution. The estimators are proposed when the parameter of the rare unrelated attribute is known and unknown. The proposed estimators are evaluated using a relative efficiency comparing the variances of the estimators reported in Lee *et al*. [4].

2. Improved Estimation of a Rare Sensitive Attribute in Stratified Sampling-Known Rare Unrelated Attributes

Consider the population of size N individuals which is divided into L subpopulations (strata) of sizes $N_h\left(h = 1, 2, \cdots, L\right)$. All the subpopulations are disjoint and together comprise the whole population. In stratum h, n_h respondent are selected by simple random sampling with replacement (SRSWR) and asked to use the pair of randomization devices R_{h1} and R_{h2}, each consisting of the two statements. The randomization device R_{h1} is constructed as:

(i) "I possessrare sensitive attribute *A*"

(ii) "Go to randomization device R_{h2}"

with respective probabilities P_{1h} and $(1-P_{1h})$.

The randomization device R_{h2} consists of two statements:

(i) "I possess rare sensitive attribute *A*"

(ii) "I possess rare unrelated attribute *Y*"

with probabilities P_{2h} and $(1-P_{2h})$ respectively.

By this randomized device, the probability of a *yes* response in stratum *h* is given by

$$\theta_{h0} = P_{h1}\pi_{hA} + (1-P_{h1})\{P_{h2}\pi_{hA} + (1-P_{h2})\pi_{hY}\}, \tag{1}$$

where π_{hA} and π_{hY} are the population proportions of individuals possessing rare sensitive and rare unrelated attributes in the h^{th} stratum, respectively. Here π_{hY} is assumed to be known. Since *A* and *Y* are very rare attributes, $n_h\theta_{h0} = \lambda_{h0}$ is finite, assuming $n_h \to \infty$ and $\theta_{h0} \to 0$.

Let $x_{h1}, x_{h2}, \cdots, x_{hn_h}$ be an n_h random sample in stratum *h* from a Poisson distribution with parameter λ_{h0}. Then the maximum likelihood estimator for the mean number of persons who have the rare sensitive attribute in stratum *h*, $\lambda_{hA}\left(= n_h\pi_{hA}\right)$, is given by

$$\hat{\lambda}_{hA} = \frac{1}{\{P_{h1} + (1-P_{h1})P_{h2}\}}\left[\frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi} - (1-P_{h1})(1-P_{h2})\lambda_{hY}\right], \tag{2}$$

where $\lambda_{hY} = n_h\pi_{hY}$ is (known) mean of persons who have rare unrelated attribute in stratum *h*. The parameter λ_A, is the mean number of persons possessing rare sensitive attribute *A*, in a population of size *N* and its estimator $\hat{\lambda}_A$ is given by

$$\hat{\lambda}_A = \sum_{h=1}^{L} W_h\hat{\lambda}_{hA} = \sum_{h=1}^{L} W_h \frac{1}{\{P_{h1} + (1-P_{h1})P_{h2}\}}\left[\frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi} - (1-P_{h1})(1-P_{h2})\lambda_{hY}\right], \tag{3}$$

where $W_h = N_h/N$.

The variance of the estimator $\hat{\lambda}_{hA}$ in each stratum is given by

$$V\left(\hat{\lambda}_{hA}\right)=\frac{A_h}{n_h}, \tag{4}$$

where

$$A_h=\frac{\lambda_{hA}}{\left[P_{h1}+\left(1-P_{h1}\right)P_{h2}\right]}+\frac{\left(1-P_{h1}\right)\left(1-P_{h2}\right)\lambda_{hY}}{\left[P_{h1}+\left(1-P_{h1}\right)P_{h2}\right]^2}.$$

Thus, the variance expression of the estimator $\hat{\lambda}_A$ may be derived as

$$V\left(\hat{\lambda}_A\right)=V\left[\sum_{h=1}^{L}W_h\hat{\lambda}_{hA}\right]=\sum_{h=1}^{L}\frac{W_h^2A_h}{n_h}. \tag{5}$$

THEOREM 1. *$\hat{\lambda}_A$ is an unbiased estimator of λ_A.*
Proof. From (3), we have

$$E\left(\hat{\lambda}_A\right)=\sum_{h=1}^{L}\frac{W_h}{\left[P_{h1}+\left(1-P_{h1}\right)P_{h2}\right]}\left\{\frac{1}{n_h}\sum_{i=1}^{n_h}E\left(x_{hi}\right)-\left(1-P_{h1}\right)\left(1-P_{h2}\right)\lambda_{hY}\right\},$$

$$=\sum_{h=1}^{L}W_h\left\{\frac{\lambda_{h0}-\left(1-P_{h1}\right)\left(1-P_{h2}\right)\lambda_{hY}}{P_{h1}+\left(1-P_{h1}\right)P_{h2}}\right\}=\sum_{h=1}^{L}W_h\lambda_{hA}=\lambda_A.$$

THEOREM 2. *The unbiased estimator for $V\left(\hat{\lambda}_A\right)$ is given by*

$$\hat{V}\left(\hat{\lambda}_A\right)=\sum_{h=1}^{L}\frac{W_h^2}{n_h^2\left[P_{h1}+\left(1-P_{h1}\right)P_{h2}\right]^2}\left\{\sum_{i=1}^{n_h}x_{hi}\right\}. \tag{6}$$

Proof.

$$E\left\{\hat{V}\left(\hat{\lambda}_A\right)\right\}=\sum_{h=1}^{L}\frac{W_h^2}{n_h^2\left[P_{h1}+\left(1-P_{h1}\right)P_{h2}\right]^2}\left\{\sum_{i=1}^{n_h}E\left(x_{hi}\right)\right\}$$

$$=\sum_{h=1}^{L}\frac{W_h^2}{n_h\left[P_{h1}+\left(1-P_{h1}\right)P_{h2}\right]^2}\left\{\lambda_{h0}\right\}=V\left(\hat{\lambda}_A\right)$$

Now, we consider the proportional and optimal allocations of the total sample size n into different strata. The method of proportional allocation is used to define sample sizes in each stratum depending on each stratum size. Since the sample size in each stratum is defined as $n_h=nN_h/N$, the variance of the estimator $\hat{\lambda}_A$, under proportional allocation of sample size is given by

$$V\left(\hat{\lambda}_A\right)_{prop}=\frac{1}{n}\left[\sum_{h=1}^{L}W_hA_h\right]. \tag{7}$$

However, the optimal allocation is a technique to define sample size to minimize variance for a given cost or to minimize the cost for a specified variance. The n_h is proportionate to the standard deviation, S_h of the variable. In stratified sampling, let cost function is defined as $C=c_0+\sum_{h=1}^{L}c_hn_h$, where c_0 is the fixed cost and c_h is the cost for the each individual stratum. Within each stratum the cost is proportional to the size of sample, but the cost c_h may vary from stratum to stratum. For fixed cost, using the Cauchy Schwarz inequality, the sample size n_h to minimize $V\left(\hat{\lambda}_A\right)$ is given by

$$n_h=n\times\frac{W_h\sqrt{A_h/c_h}}{\sum_{h=1}^{L}W_h\sqrt{A_h/c_h}}. \tag{8}$$

So the minimum variance of the estimator for the specified cost C under the optimum allocation of sample size is given by

$$V\left(\hat{\lambda}_A\right)_{opt} = \frac{1}{n}\left[\sum_{h=1}^{L} W_h \sqrt{A_h \times c_h}\right] \times \left[\sum_{h=1}^{L} W_h \sqrt{A_h / c_h}\right]. \tag{9}$$

3. Improved Estimation of a Rare Sensitive Attribute in Stratified Sampling-Unknown Rare Unrelated Attributes

In this section, the estimators for the mean number of rare sensitive attribute are proposed under the assumptions that the sizes of stratum are known; however, $\lambda_{hY} = n_h \pi_{hY}$, the mean of the rare unrelated attribute is unknown. In this case each selected respondent from stratum h is asked to use the sequential pair of randomization devices. That in the h^{th} stratum, n_h, respondents are asked to use the randomization devices R_{h1} and R_{h2} consisting of two statements. The device R_{h1} consists of two statements:

(i) "I possess a sensitive group A"

(ii) "Go to randomization device R_{h2}"

The statements occur with respective probabilities P_{1h} and $(1-P_{1h})$.

The two statements of the randomization device R_{h2} are:

(i) "I possess a sensitive attribute A"

(ii) "I possess unrelated attribute Y"

represented with respective probabilities P_{2h} and $(1-P_{2h})$. After using the first pair of randomized devices, respondent is asked to use the same pair of devices R_{h1} and R_{h2} but with probabilities T_{1h}, $(1-T_{1h})$ and T_{2h}, $(1-T_{2h})$, respectively.

The probabilities of the *yes* responses for the first and second use of pair of randomization devices are respectively given by

$$\theta_{h1} = P_{h1}\pi_{hA} + (1-P_{h1})\left[P_{h1}\pi_{hA} + (1-P_{h1})\pi_{hY}\right] \tag{10}$$

and

$$\theta_{h2} = T_{h1}\pi_{hA} + (1-T_{h1})\left[T_{h1}\pi_{hA} + (1-T_{h1})\pi_{hY}\right], \tag{11}$$

where π_{hA} and π_{hY} are the respective population proportions of rare sensitive and rare unrelated attribute in the stratum h. As n_h is large and $(\pi_{hA}, \pi_{hY}) \to 0$, therefore $(\theta_{h1}, \theta_{h2}) \to 0$. Now, obviously $\lambda_{h1} = n_h\theta_{h1}$, $\lambda_{h2} = n_h\theta_{h2}$. Let x_{h1i} and x_{h2i} ($h = 1, 2, \cdots, L,\quad i = 1, 2, \cdots, n_h$) be the pair of responses from the ith respondent selected in h^{th} stratum. We have

$$Var\left(x_{h1i}\right) = E\left(x_{h1i}\right) = \lambda_{h1} = P_{h1}\lambda_{hA} + (1-P_{h1})\left\{P_{h2}\lambda_{hA} + (1-P_{h2})\lambda_{hY}\right\} \tag{12}$$

$$Var\left(x_{h2i}\right) = E\left(x_{h2i}\right) = \lambda_{h2} = T_{h1}\lambda_{hA} + (1-T_{h1})\left\{T_{h2}\lambda_{hA} + (1-T_{h2})\lambda_{hY}\right\} \tag{13}$$

$$\begin{aligned} Cov\left(x_{h1i}, x_{h2i}\right) &= E\left(x_{h1i}x_{h2i}\right) - E\left(x_{h1i}\right)E\left(x_{h2i}\right) \\ &= \left\{P_{h1} + (1-P_{h1})P_{h2}\right\}\left\{T_{h1} + (1-T_{h1})T_{h2}\right\}\lambda_{hA} + (1-P_{h1})(1-P_{h2})(1-T_{h1})(1-T_{h2})\lambda_{hY}, \end{aligned} \tag{14}$$

Following the expression given in Equations (12) and (13), we have the sample means for both set of responses as

$$\frac{1}{n_h}\sum_{i=1}^{n_h} x_{h1i} = P_{h1}\hat{\lambda}_{hA} + (1-P_{h1})\left[P_{h2}\hat{\lambda}_{hA} + (1-P_{h2})\hat{\lambda}_{hY}\right] \tag{15}$$

and

$$\frac{1}{n_h}\sum_{i=1}^{n_h} x_{h2i} = T_{h1}\hat{\lambda}_{hA} + (1-T_{h1})\left[T_{h2}\hat{\lambda}_{hA} + (1-T_{h2})\hat{\lambda}_{hY}\right]. \tag{16}$$

By solving (15) and (16), we get estimators of λ_{hA} and λ_{hY} as

$$\hat{\lambda}_{hA} = \frac{1}{n_h B_h}\left[(1-T_{1h})(1-T_{2h})\sum_{i=1}^{n_h} x_{h1i} - (1-P_{1h})(1-P_{2h})\sum_{i=1}^{n_h} x_{h2i}\right] \tag{17}$$

$$\hat{\lambda}_{hY} = \frac{1}{n_h D_h}\left[\left\{T_{1h} + \left(1 - T_{1h}\right)T_{2h}\right\}\sum_{i=1}^{n_h} x_{h1i} - \left\{P_{1h} + \left(1 - P_{1h}\right)P_{2h}\right\}\sum_{i=1}^{n_h} x_{h2i}\right] \tag{18}$$

where

$B_h = \left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\} - \left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\}$ and $D_h = \left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\} - \left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}$.

$$\begin{aligned} V\left(\hat{\lambda}_{hA}\right) &= \frac{1}{\left[n_h B_h\right]^2} V\left[\left(1 - T_{h1}\right)\left(1 - T_{h2}\right)\sum_{i=1}^{n_h} x_{h1i} - \left(1 - P_{h1}\right)\left(1 - P_{h2}\right)\sum_{i=1}^{n_h} x_{h2i}\right], \\ &= \frac{1}{\left[n_h B_h\right]^2}\left[\left(1 - T_{1h}\right)^2\left(1 - T_{2h}\right)^2\sum_{i=1}^{n_h} V\left(x_{h1i}\right) + \left(1 - P_{1h}\right)^2\left(1 - P_{2h}\right)^2\sum_{i=1}^{n_h} V\left(x_{h2i}\right)\right. \\ &\quad \left. - 2\left(1 - T_{1h}\right)\left(1 - T_{2h}\right)\left(1 - P_{1h}\right)\left(1 - P_{2h}\right)\sum_{i=1}^{n_h} Cov\left(x_{h1i}, x_{h2i}\right)\right] \end{aligned} \tag{19}$$

Puttinng (12), (13) and (14) in (19) we get

$$V\left(\hat{\lambda}_{hA}\right) = \sum_{h=1}^{L}\frac{\left[A_{h1} + A_{h2}\right]}{n_h B_h^2}, \tag{20}$$

where

$$\begin{aligned} A_{h1} &= \left[\left(1 - T_{h1}\right)^2\left(1 - T_{h2}\right)^2\left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\} + \left(1 - P_{h1}\right)^2\left(1 - P_{h2}\right)^2\left\{T_{h1} + \left(1 - T_{h1}\right)T_2\right\}\right. \\ &\quad \left. - 2\left(1 - T_{h1}\right)\left(1 - T_{h2}\right)\left(1 - P_{h1}\right)\left(1 - P_{h2}\right)\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\}\left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}\right]\lambda_{hA}, \\ A_{h2} &= \left[\left(1 - T_{h1}\right)\left(1 - T_{h2}\right)\left(1 - P_{h1}\right)\left(1 - P_{h2}\right)\left\{2 - \left(T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right) - \left(P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right)\right\}\right. \\ &\quad \left. - 2\left\{\left(1 - T_{h1}\right)\left(1 - T_{h2}\right)\left(1 - P_{h1}\right)\left(1 - P_{h2}\right)\right\}^2\right]\lambda_{hY}. \end{aligned}$$

The stratified estimators of λ_A and λ_Y are defined as

$$\hat{\lambda}_A = \sum_{h=1}^{L} W_h\hat{\lambda}_{hA}, \text{ and } \hat{\lambda}_Y = \sum_{h=1}^{L} W_h\hat{\lambda}_{hY}. \tag{21}$$

THEOREM 3. $\hat{\lambda}_A$ *is an unbiased estimator for* λ_A.

Proof.

$$\begin{aligned} E\left(\hat{\lambda}_A\right) &= E\left[\sum_{h=1}^{L} W_h\hat{\lambda}_{hA}\right] = \sum_{h=1}^{L}\frac{W_h}{n_h B_h}\left[\left(1 - T_{1h}\right)\left(1 - T_{2h}\right)\sum_{i=1}^{n_h} E\left(x_{h1i}\right) - \left(1 - P_{1h}\right)\left(1 - P_{2h}\right)\sum_{i=1}^{n} E\left(x_{h2i}\right)\right] \\ &= \sum_{h=1}^{L}\frac{W_h}{B_h}\left[\left(1 - T_{1h}\right)\left(1 - T_{2h}\right)\lambda_{h1} - \left(1 - P_{1h}\right)\left(1 - P_{2h}\right)\lambda_{h2}\right]. \end{aligned} \tag{22}$$

Putting the values of λ_{h1} and λ_{h2} in Equation (22), we get the result.

THEOREM 4. *The variance of* $\hat{\lambda}_A$ *is given by*

$$V\left(\hat{\lambda}_A\right) = \sum_{h=1}^{L}\frac{W_h^2}{n_h B_h^2}\left[A_{h1} + A_{h2}\right], \tag{23}$$

where

$$\begin{aligned} A_{h1} &= \left[\left(1 - T_{h1}\right)^2\left(1 - T_{h2}\right)^2\left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\} + \left(1 - P_{h1}\right)^2\left(1 - P_{h2}\right)^2\left\{T_{h1} + \left(1 - T_{h1}\right)T_2\right\}\right. \\ &\quad \left. - 2\left(1 - P_{h1}\right)\left(1 - P_{h2}\right)\left(1 - T_{h1}\right)\left(1 - T_{h2}\right)\left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\}\right]\lambda_{hA}, \\ A_{h2} &= \left[\left(1 - P_{h1}\right)\left(1 - P_{h2}\right)\left(1 - T_{h1}\right)\left(1 - T_{h2}\right)\left\{2 - \left(P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right) - \left(T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right)\right\}\right. \\ &\quad \left. - 2\left\{\left(1 - P_{h1}\right)\left(1 - P_{h2}\right)\left(1 - T_{h1}\right)\left(1 - T_{h2}\right)\right\}^2\right]\lambda_{hY}. \end{aligned}$$

Proof. Since $\hat{\lambda}_A = \sum_{h=1}^{L} W_h \hat{\lambda}_{hA}$, we have

$$V\left(\hat{\lambda}_A\right) = V\left[\sum_{h=1}^{L} W_h \hat{\lambda}_{hA}\right] = \sum_{h=1}^{L} W_h^2 V\left(\hat{\lambda}_{hA}\right) \tag{24}$$

On putting (20) in (24) we have the theorem.

Corollary 1: An unbiased estimator for the variance of rare sensitive attribute is given by

$$\hat{V}\left(\hat{\lambda}_A\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h B_h^2}\left[\hat{A}_{h1} + \hat{A}_{h2}\right] \tag{25}$$

It can be proved easily.

THEOREM 5. $\hat{\lambda}_Y$ is an unbiased estimator of λ_Y.

Proof. From (18), we have

$$\begin{aligned}
&E\left(\hat{\lambda}_{hY}\right) \\
&= \frac{1}{n_h\left[\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\} - \left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}\right]}\left[\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\}\sum_{i=1}^{n_h} E\left(x_{h1i}\right) - \left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}\sum_{i=1}^{n_h}\left(x_{h2i}\right)\right] \\
&= \frac{1}{\left[\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\} - \left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}\right]}\left[\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\}\lambda_{h1} - \left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}\lambda_{h2}\right] \\
&= \sum_{h=1}^{L} W_h E\left(\hat{\lambda}_{hY}\right) = \sum_{h=1}^{L} W_h \lambda_{hY} = \lambda_Y.
\end{aligned}$$

Corollary 2: An unbiased estimator for $V\left(\hat{\lambda}_Y\right)$ is given by

$$\hat{V}\left(\hat{\lambda}_Y\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h D_h^2}\left[C_{h1}\hat{\lambda}_{hA} + C_{h2}\hat{\lambda}_{hY}\right] \tag{26}$$

where

$$\begin{aligned}
C_{h1} = &\left[\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\}^2\left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\} + \left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}^2\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\}\right. \\
&\left. - 2\left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}^2\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\}^2\right],
\end{aligned}$$

$$\begin{aligned}
C_{h2} = &\left[\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\}^2\left(1 - P_{h1}\right)\left(1 - P_{h2}\right) + \left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}^2\left(1 - T_{h1}\right)\left(1 - T_{h2}\right)\right. \\
&\left. - 2\left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}\left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\}\left(1 - P_{h1}\right)\left(1 - P_{h2}\right)\left(1 - T_{h1}\right)\left(1 - T_{h2}\right)\right],
\end{aligned}$$

$$D_h = \left\{T_{h1} + \left(1 - T_{h1}\right)T_{h2}\right\} - \left\{P_{h1} + \left(1 - P_{h1}\right)P_{h2}\right\}.$$

Now under proportional allocation of sample size, the variance of $\hat{\lambda}_A$ is given by

$$V\left(\hat{\lambda}_A\right)_{prop} = \frac{1}{n}\sum_{h=1}^{L} \frac{W_h}{B_h^2}\left[A_{h1} + A_{h2}\right].$$

However, in optimum allocation, the sample size in stratum h is

$$n_h = n \times \left[\frac{W_h}{B_h}\sqrt{\left(A_{h1} + A_{h2}\right)/c_h}\right] \div \left[\sum_{h=1}^{L} \frac{W_h}{B_h}\sqrt{\left(A_{h1} + A_{h2}\right)/c_h}\right]$$

and the variance of $\hat{\lambda}_A$ is given by

$$V\left(\hat{\lambda}_A\right)_{opt} = \frac{1}{n}\left[\sum_{h=1}^{L} \frac{W_h}{B_h}\sqrt{\left(A_{h1} + A_{h2}\right)c_h}\right] \times \left[\sum_{h=1}^{L} \frac{W_h}{B_h}\sqrt{\left(A_{h1} + A_{h2}\right)/c_h}\right].$$

4. Relative Efficiency

Lee *et al.* [4] proposed variance of $\hat{\lambda}_A$ for rare sensitive attribute based on Poisson distribution when the rare unrelated attribute known and unknown respectively is:

$$V_{L1}\left(\hat{\lambda}_A\right)=\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\frac{\lambda_{hA}}{P_{h1}}+\frac{\left(1-P_{h1}\right)\lambda_{hy}}{P_{h1}^2}\right], \tag{27}$$

$$V_{L2}\left(\hat{\lambda}_A\right)=\sum_{h=1}^{L}\frac{W_h^2}{n_h\left(P_{h1}-T_{h1}\right)^2}\left[\Lambda_{h1}+\Lambda_{h2}\right], \tag{28}$$

where

$$\Lambda_{h1}=\left\{P_{h1}\left(1-T_{h1}\right)^2+T_{h1}\left(1-P_{h1}\right)^2-2P_{h1}T_{h1}\left(1-P_{h1}\right)\left(1-T_{h1}\right)\right\}\lambda_{hA}$$

$$\Lambda_{h2}=\left\{\left(1-P_{h1}\right)\left(1-T_{h1}\right)\left(2-P_{h1}-T_{h1}\right)-2\left(1-P_{h1}\right)^2\left(1-T_{h1}\right)^2\right\}\lambda_{hY}.$$

For comparison of the proposed estimator with $V_L\left(\hat{\lambda}_A\right)$, the relative efficiency is given by

$$RE=\frac{V_L\left(\hat{\lambda}_A\right)}{V\left(\hat{\lambda}_A\right)}.$$

Large samples are required to estimate the means of rare sensitive attribute. So we consider a large hypothetical population, in order to study the relative efficiency, setting $n=10000$ with two strata having $n_1=4000$ and $n_2=6000$. We choose values of the parameters $\left(\lambda_{1A},\lambda_{1Y}\right)$, $\left(\lambda_{2A},\lambda_{2Y}\right)$ as $(0.5,1.5),(1.5,0.5),(1.5,1.5)$ and $(0.5,0.5)$, and we let the value P_{12} range from 0.3 to 0.7, and let that of P_{11} range from 0.6 to 0.9 when the weights $W_1=0.4$ (and $W_2=0.6$) and $W_1=0.6$ (and $W_2=0.4$) which is proportional allocation. Also, let ($\lambda_{1A}=\lambda_{2A}$) and ($\lambda_{1Y}=\lambda_{2Y}$).

4.1. Relative Efficiency When Rare Unrelated Attribute Is Known

Let $V_1\left(\hat{\lambda}_A\right)$ be the variance of the proposed estimator $\hat{\lambda}_A$ for the rare sensitive attribute when the parameter of rare unrelated attribute is known. The relative efficiency of proposed estimator with respect to $V\left(\hat{\lambda}_A\right)_{L1}$ estimator is defined as

$$RE_1=\frac{V\left(\hat{\lambda}_A\right)_{L1}}{V_1\left(\hat{\lambda}_A\right)}=\frac{\left[\sum_{h=1}^{2}W_h\left\{\frac{\lambda_{hA}}{P_h}+\frac{\left(1-P_h\right)\lambda_{hY}}{P_h^2}\right\}\right]}{\left[\sum_{h=1}^{2}W_h\left\{\frac{\lambda_{hA}}{\left[P_{h1}+\left(1-P_{h1}\right)P_{h2}\right]}+\frac{\left(1-P_{h1}\right)\left(1-P_{h2}\right)\lambda_{hY}}{\left[P_{h1}+\left(1-P_{h1}\right)P_{h2}\right]^2}\right\}\right]}. \tag{29}$$

From Equation (29) it evident that the relative efficiency of proposed estimator is free from the sample size *n*. We set the design probabilities as $P_{11}=P_{21}$ and $P_{12}=P_{22}$. In **Table 1**, the relative efficiencies are given with parameter values $\left(\lambda_{1A},\lambda_{1Y}\right)$, $\left(\lambda_{2A},\lambda_{2Y}\right)$ as $(0.5,1.5),(1.5,0.5),(1.5,1.5)$ and $(0.5,0.5)$, P_{12} varies from 0.3 to 0.7, and P_{11} from 0.6 to 0.9 having weights $W_1=0.4,0.6$ $\left(W_1+W_2=1\right)$. It is evident that the proposed estimator has efficiency greater than 1 in all cases, and is always better than the $V\left(\hat{\lambda}_A\right)_{L1}$ estimator. A study of **Figure 1** confirms this.

4.2. Relative Efficiency When Rare Unrelated Attribute Is Unknown

Let $V_2\left(\hat{\lambda}_A\right)$ be the variance of the proposed estimator $\hat{\lambda}_A$ for the rare sensitive attribute when the parameter of rare unrelated attribute is unknown. The relative efficiency of proposed estimator with respect to $V\left(\hat{\lambda}_A\right)_{L2}$ estimator is defined as

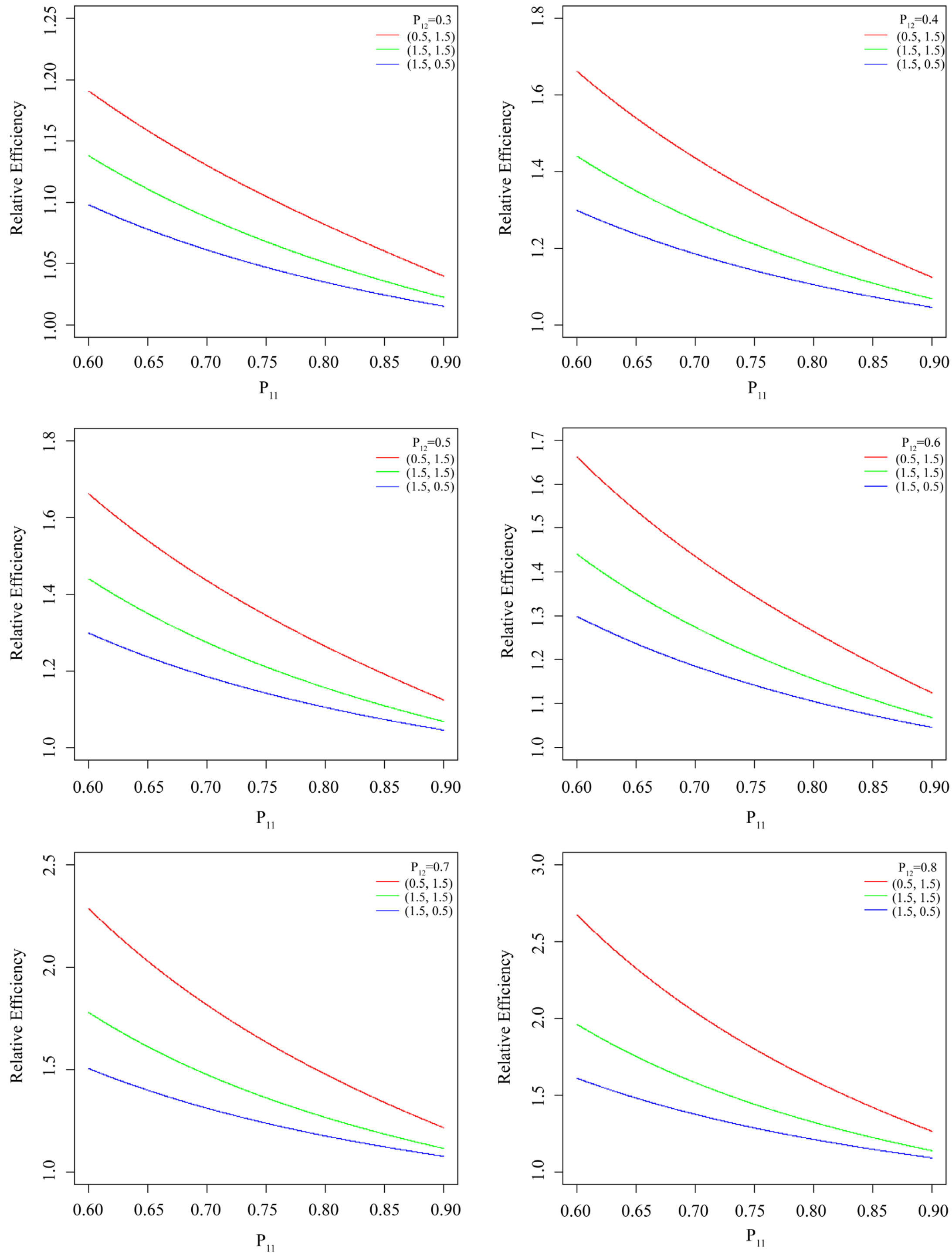

Figure 1. Relative Efficiency (RE) of the proposed model with respect to Lee *et al.* [4] for $W_1 = 0.4$ and $P_{12} = 0.3$ to 0.8.

$$RE_2 = \frac{V\left(\hat{\lambda}_A\right)_{L2}}{V_2\left(\hat{\lambda}_A\right)} = \frac{\sum_{h=1}^{2} W_h \frac{\left[A'_{h1} + A'_{h2}\right]}{\left[P_{h1} - T_{h1}\right]^2}}{\sum_{h=1}^{2} W_h \frac{\left[A_{h1} + A_{h2}\right]}{\left[\left\{P_{h1} + \left(1 - P_{h1}\right) P_{h2}\right\} - \left\{T_{h1} + \left(1 - T_{h1}\right) T_{h2}\right\}\right]^2}}. \quad (30)$$

The relative efficiency of proposed estimator is free from the sample size *n*. For the analysis, the design probabilities are fixed as $P_{11} = P_{21}$, $P_{12} = P_{22}$, $T_{11} = T_{21}$, $T_{12} = T_{22}$. Setting $\lambda_{1A} = \lambda_{2A}$, $\lambda_{1Y} = \lambda_{2Y}$ with parameter values of $(\lambda_{1A}, \lambda_{1Y})$, $(\lambda_{2A}, \lambda_{2Y})$ as $(0.5,1.5),(1.5,0.5),(1.5,1.5)$ and $P_{11} = 0.6$, $T_{11} = 0.3, 0.4$, T_{12} = 0.2, 0.3, 0.4, 0.5 and $W_1 = 0.4, 0.5$ $(W_1 + W_2 = 1)$. The relative efficiencies are given in **Table 2** depict that the proposed estimator outer perform than $V(\hat{\lambda}_A)_{L2}$ estimator having efficiency greater than 1 if we set the probabilities as $P_{12} \geq T_{12}$. However the relative efficiency starts decreasing as we take $P_{12} < T_{12}$. A study of **Figure 2** confirms this. Also, when W_1 increasesthe relative efficiency of proposed estimator increases.

Table 1. Relative efficiency of the proposed estimator with Lee *et al.* (2013).

			$W_1 = 0.4$				$W_1 = 0.6$			
P_{12}	λ_{1Y}	λ_{1A}	$P_{11} = 0.6$	0.7	0.8	0.9	$P_{11} = 0.6$	0.7	0.8	0.9
0.3	0.5	1.5	1.7346	1.5829	1.4758	1.3966	1.5630	1.4264	1.3299	1.2585
	1.5	1.5	1.9238	1.7016	1.5439	1.4266	1.7336	1.5334	1.3912	1.2855
	1.5	0.5	2.2198	1.9173	1.6887	1.5016	2.0003	1.7277	1.5217	1.3531
0.4	0.5	1.5	1.8713	1.6667	1.5228	1.4169	1.6863	1.5018	1.3723	1.2768
	1.5	1.5	2.1435	1.8333	1.6166	1.4574	1.9316	1.6520	1.4567	1.3133
	1.5	0.5	2.6070	2.1568	1.8251	1.5615	2.3492	1.9436	1.6447	1.4071
0.5	0.5	1.5	2.0097	1.7510	1.5701	1.4372	1.8109	1.5779	1.4148	1.2951
	1.5	1.5	2.3751	1.9699	1.6908	1.4885	2.2100 402	1.7751	1.5327	1.3413
	1.5	0.5	3.0537	2.4245	1.9727	1.6238	2.7517	2.1848	1.7776	1.4633
0.6	0.5	1.5	1.6090	1.01489	1.2107	1.0910	1.9370	1.6545	1.4576	1.3135
	1.5	1.5	1.9600	1.4204	1.3225	1.1377	2.3596	1.9026	1.5921	1.3698
	1.5	0.5	2.6727	1.6326	1.5961	1.2642	3.2177	2.4550	1.9215	1.5219
0.7	0.5	1.5	1.7147	1.4383	1.2464	1.1063	2.0642	1.7315	1.5005	1.3318
	1.5	1.5	2.1511	1.6900	1.3806	1.1616	2.5897	2.0346	1.6621	1.3984
	1.5	0.5	3.1223	2.2915	1.7258	1.3150	3.7592	2.7587	2.0776	1.5831

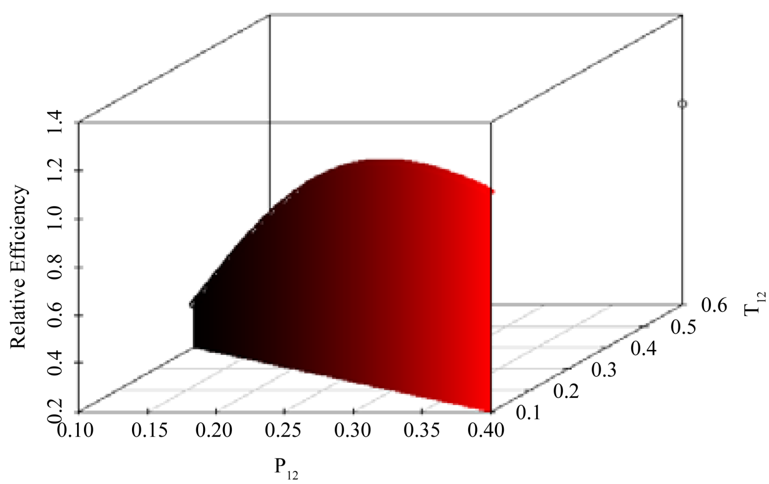

Figure 2. Relative Efficiency (RE) of the proposed model with respect to Lee *et al.* [4] for indicated values.

Table 2. Relative efficiency of the proposed estimator with Lee *et al*. (2013), $W_1 = 0.4$, and $W_1 = 0.5$.

$P_{11} = P_{21}$	$P_{12} = P_{22}$	$T_{11} = T_{21}$	$T_{12} = T_{22}$	$\lambda_{1A} = \lambda_{2A}$	$\lambda_{1Y} = \lambda_{2Y}$	RE ($W_1 = 0.4$)	RE ($W_1 = 0.5$)
0.6	0.6	0.3	0.2	1.5	0.5	12.5971	15.7464
				1.5	1.5	16.9517	21.1896
				0.5	1.5	10.0051	12.5064
			0.3	1.5	0.5	10.3926	12.9908
				1.5	1.5	13.9851	17.4814
				0.5	1.5	8.2542	10.3178
			0.4	1.5	0.5	8.1881	10.2352
				1.5	1.5	11.0186	13.7732
				0.5	1.5	6.5033	8.1292
			0.5	1.5	0.5	5.9836	7.4795
				1.5	1.5	8.0520	10.0651
				0.5	1.5	4.7524	5.9405
0.6	0.6	0.4	0.2	1.5	0.5	3.1703	3.9629
				1.5	1.5	4.4483	5.5603
				0.5	1.5	2.7607	2.4509
			0.3	1.5	0.5	2.5759	3.2198
				1.5	1.5	3.6142	4.5178
				0.5	1.5	2.2431	2.8038
			0.4	1.5	0.5	1.9814	2.4768
				1.5	1.5	2.7801	3.4752
				0.5	1.5	1.7254	2.1568
			0.5	1.5	0.5	1.3870	1.7338
				1.5	1.5	1.9461	2.4326
				0.5	1.5	1.2078	1.5098

5. Conclusion

In this study, a two stage randomized response model is proposed with improved estimators for the mean and its variance of the number of persons possessing a rare sensitive attribute based on stratified sampling by using Poisson distribution. It is shown that our proposed method have better efficiencies than the existing randomized response model, when the parameter of rare unrelated attribute is known and in unknown case, depending on the probability of selecting a question. For future work, we can obtain more sensitive information from respondents by using stratified double sampling with the proposed model.

References

[1] Warner, S.L. (1965) Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of*

Computational and Graphical Statistics, **60**, 63-66. http://dx.doi.org/10.1080/01621459.1965.10480775

[2] Greenberg, B.G., Abul-Ela, A.L.A., Simmons, W.R. and Horvitz, D.G. (1969) The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association*, **64**, 520-539. http://dx.doi.org/10.1080/01621459.1969.10500991

[3] Mangat, N.S. and Singh, R. (1990) On the Confidentiality Guaranteed under Randomized Response Sampling: A Comparison with Several New Techniques. *Biometrical Journal*, **40**, 237-242.

[4] Lee, G.S., Uhm, D. and Kim, J.M. (2013) Estimation of a Rare Sensitive Attribute in a Stratified Sample Using Poisson Distribution. *Statistics*, **47**, 685-709. http://dx.doi.org/10.1080/02331888.2011.625503

[5] Chaudhuri, A. and Mukerjee, R. (1988) Randomized Response: Theory and Techniques. Marcel Dekker, New York.

[6] Mahmood, M., Singh, S. and Horn, S. (1998) On the Confidentiality Guaranteed under Randomized Response Sampling: A Comparison with Several New Techniques. *Biometrical Journal*, **40**, 237-242. http://dx.doi.org/10.1002/(SICI)1521-4036(199806)40:2<237::AID-BIMJ237>3.0.CO;2-N

[7] Land, M., Singh, S. and Sedory, S.A. (2012) Estimation of a Rare Sensitive Attribute Using Poisson Distribution. *Statistics*, **46**, 351-360. http://dx.doi.org/10.1080/02331888.2010.524300

[8] Bhargava, M. and Singh, R. (2000) A Modified Randomization Device for Warner's Model. *Statistica*, **60**, 315-321.