A NEW OPTIMAL ESTIMATOR OF POPULATION PROPORTION IN RANDOMIZED RESPONSE SAMPLING

Oluseun Odumade
Research and Development Division
Educational Testing Service
Princeton, NJ 08541
E-mail: oodumade@ets.org

Sarjinder Singh
Department of Mathematics
Texas A&M University-Kingsville
Kingsville, TX 78363
E-mail: sarjinder@yahoo.com

ABSTRACT

A new optimal estimator of population proportion of potentially sensitive attributes in survey sampling is proposed and investigated. The proposed estimator makes use of known average values and known common variance of two scrambling variables at the data collection and estimation stages; more cooperation is expected from the respondents than in the Franklin (1989) model. The variance of the proposed estimator is minimized to determine the value of a constant which leads to an optimal estimator of the population proportion. The resulting optimal estimator has been found to be more efficient than the Franklin (1989) estimator, and the Singh and Chen (2009) estimator which suggest utilizing higher order moments of scrambling variables at the estimation stage.

Keywords: Randomized response sampling; Estimation of proportion; Two scrambling variables; optimal estimator.

1. INTRODUCTION

The collection of data through personal interview surveys on sensitive issues such as induced abortions, drug abuse, and family income is a serious issue; see for example Fox and Tracy (1986) and Kerkvliet (1994). Warner (1965) considered a case in which the respondents in a population can be divided into two mutually exclusive groups: one group with stigmatizing/sensitive characteristic *A* and the other group without it. For estimating π, the proportion of respondents in the population belonging to the sensitive group *A*, a simple random and with replacement sample (SRSWR) of n respondents is selected from the population. For collecting information on a sensitive characteristic, Warner (1965) made use of a randomization device. One such device could be a deck of cards with each card having one of the following two statements: (i) "I belong to group A" (ii) "I do not belong to group A." The statements occur with relative frequencies, P and $(1-P)$, respectively, in the deck of cards. Each respondent in the sample is asked to select a card at random from the well-shuffled deck. Without showing the card to the interviewer, the interviewee answers the question, "Is the statement true for you?" The number of respondents n_1 that answer "Yes" is binomially distributed with parameters n and $P\pi + (1-P)(1-\pi)$. The maximum likelihood estimator of π exists for $P \neq 0.5$ and is given by:

$$\hat{\pi}_w = \frac{(n_1/n)-(1-P)}{2P-1} \tag{1.1}$$

The above estimator is unbiased with variance:

$$V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2} \tag{1.2}$$

Kuk (1990) and Franklin (1989) suggested using two different independent randomization devices, depending upon the status of a person selected in the sample. If a person selected in the sample belongs to the sensitive group A, then he/she is requested to report a random number drawn from a scrambling variable S_1; otherwise if he/she belongs to the non-sensitive group A^c, then he/she is requested to report a random number drawn from another independent scrambling variable S_2.

Thus the distribution of the observed response Z_i is given by:

$$Z_i = \begin{cases} S_1 \text{ with probability } \pi \\ S_2 \text{ with probability } (1-\pi) \end{cases} \tag{1.3}$$

Assuming that $E(S_1) = \theta_1$ and $E(S_2) = \theta_2$ are known, Kuk (1990) and Franklin (1989) suggested an unbiased estimator of π as:

$$\hat{\pi}_f = \frac{\frac{1}{n}\sum_{i=1}^{n} Z_i - \theta_2}{(\theta_1 - \theta_2)} \tag{1.4}$$

with variance:

$$V(\hat{\pi}_f) = \frac{\pi(1-\pi)}{n} + \frac{\pi\gamma_{20} + (1-\pi)\gamma_{02}}{n(\theta_1 - \theta_2)^2} \tag{1.5}$$

where $\gamma_{20} = V(S_1)$ and $\gamma_{02} = V(S_2)$ are the variances of the scrambling variables S_1 and S_2 respectively. Singh and Chen (2009) have suggested computing the distribution of squares of the scrambled responses, from (1.3), at no extra cost of doing a survey:

$$Z_i^2 = \begin{cases} S_1^2 \text{ with probability } \pi \\ S_2^2 \text{ with probability } (1-\pi) \end{cases} \tag{1.6}$$

An unbiased estimator of the population proportion π is given by:

$$\hat{\pi}_h = \frac{\frac{1}{n}\sum_{i=1}^{n} Z_i^2 - (\gamma_{02} + \theta_2^2)}{(\gamma_{20} + \theta_1^2) - (\gamma_{02} + \theta_2^2)} \tag{1.7}$$

with variance given by:

$$V(\hat{\pi}_h) = \frac{\pi(1-\pi)}{n} + \frac{\pi\{\gamma_{40} + 4\gamma_{30}\theta_1 + 4\gamma_{20}\theta_1^2 - \gamma_{20}^2\} + (1-\pi)\{\gamma_{04} + 4\gamma_{03}\theta_2 + 4\gamma_{02}\theta_2^2 - \gamma_{02}^2\}}{n\{(\gamma_{20} + \theta_1^2) - (\gamma_{02} + \theta_2^2)\}^2} \tag{1.8}$$

where $\gamma_{ab} = E(S_1 - \theta_1)^a (S_2 - \theta_2)^b$, which provides higher order moments of the scrambling variables, for non-negative integer values of a and b such that $a + b \le 4$.

Singh and Chen (2009) proposed an unbiased estimator of the population proportion π as a linear combination of $\hat{\pi}_f$ and $\hat{\pi}_h$ as follows:

$$\hat{\pi}_{sc} = \alpha\,\hat{\pi}_h + (1-\alpha)\,\hat{\pi}_f \tag{1.9}$$

where α is a constant such that the variance of the estimator $\hat{\pi}_{sc}$ is minimized. If $\alpha = 1$ then the estimator $\hat{\pi}_{sc}$ reduces to the estimator $\hat{\pi}_h$, and if $\alpha = 0$ then the estimator $\hat{\pi}_{sc}$ reduces to the estimator $\hat{\pi}_f$, thus Singh and Chen (2009) focus only to study of properties of the estimator $\hat{\pi}_{sc}$ which is more general case than other estimators. The variance of the estimator $\hat{\pi}_{sc}$ is given by:

$$V(\hat{\pi}_{sc}) = \alpha^2 V(\hat{\pi}_h) + (1-\alpha)^2 V(\hat{\pi}_f) + 2\alpha(1-\alpha)Cov(\hat{\pi}_f, \hat{\pi}_h) \tag{1.10}$$

where

$$Cov(\hat{\pi}_f, \hat{\pi}_h) = \frac{\pi(1-\pi)}{n} + \frac{\pi(\gamma_{30} + 2\theta_1\gamma_{20}) + (1-\pi)(\gamma_{03} + 2\theta_2\gamma_{02})}{n\{(\gamma_{20} + \theta_1^2) - (\gamma_{02} + \theta_2^2)\}(\theta_1 - \theta_2)} \quad (1.11)$$

The optimal value of α which minimizes the variance of the estimator $\hat{\pi}_{sc}$ is given by

$$\alpha = \frac{V(\hat{\pi}_f) - Cov(\hat{\pi}_f, \hat{\pi}_h)}{V(\hat{\pi}_f) + V(\hat{\pi}_h) - 2Cov(\hat{\pi}_f, \hat{\pi}_h)} \quad (1.12)$$

The minimum variance of the Singh and Chen (2009 estimator $\hat{\pi}_{sc}$ is given by:

$$\text{Min.V}(\hat{\pi}_{sc}) = V(\hat{\pi}_f) - \frac{\{V(\hat{\pi}_f) - Cov(\hat{\pi}_f, \hat{\pi}_h)\}^2}{V(\hat{\pi}_f) + V(\hat{\pi}_h) - 2Cov(\hat{\pi}_f, \hat{\pi}_h)} \quad (1.13)$$

The Mangat (1994), Mangat and Singh (1990) and Gjestvang and Singh (2006) models are special cases of the Kuk (1990) and the Frankling (1989) models. Thus, it is worth to work further on Kuk (1990) and Franklin (1989) type models. Odumade and Singh (2010) have recently introduced an alternative to the Bar-Lev *et al.* (2004) randomized response model and the relative efficiency of the proposed model is studied model under various situations.

In (1.3), it is very difficult to find two scrambling variables X and Y which are practical and logically acceptable by the respondents. The respondents fears that they could get caught whether they belong to A and A^c because of the large difference between the mean values θ_1 and θ_2. Note that the larger the difference between θ_1 and θ_2, the more efficient is the estimator due to Franklin (1989) and Kuk (1990). In this paper, we propose an optimal randomized response technique that increases respondents' cooperation.

2. AN OPTIMAL RANDOMIZED RESPONSE MODEL

In the proposed randomization device, if a respondent belongs to sensitive group A, then he/she is assumed to draw r random values of the scrambling variable X as x_i, $i = 1,2,..,r$. Then he/she is assumed to compute the sample mean $\bar{x}$ and the sample variance s_x^2 unobserved by the interviewer. The value of a constant k is provided by the interviewer to the respondent. The respondent is assumed to compute his/her response as $\bar{x} + ks_x^2$ unobserved by the interviewer. Also if a respondent belongs to non-sensitive group A^c, then he/she is assumed to draw r random values of the scrambling variable Y as y_i, $i = 1,2,..,r$. Then he/she is assumed to compute the sample mean $\bar{y}$ and the sample variance s_y^2 unobserved by the interviewer. The value of a constant k is provided by the interviewer to the respondent. The respondent is assumed to compute his/her response as $\bar{y} + ks_y^2$ unobserved by the interviewer. Further let the scrambling variables

X and Y are independent random variables and assume $\sigma_x^2 = \sigma_y^2$. The distribution of the observed responses Z_i is given by:

$$Z_i = \begin{cases} \left(\bar{x} + ks_x^2\right) & \textit{with probability } \pi \\ \left(\bar{y} + ks_y^2\right) & \textit{with probability } (1-\pi) \end{cases} \tag{2.1}$$

Letting E_2 denote the expected values over the proposed randomization device, then

$$\begin{aligned} E_2(Z_i) &= \left(\mu_x + k\sigma_x^2\right)\pi + \left(\mu_y + k\sigma_y^2\right)(1-\pi) \\ &= \mu_x\pi + k\pi\sigma_x^2 + \mu_y + k\sigma_y^2 - \pi\mu_y - k\pi\sigma_y^2 \text{ (By assuming } \sigma_x = \sigma_y = \sigma) \\ &= \mu_x\pi + \mu_y + k\sigma_y^2 - \pi\mu_y \\ &= \left(\mu_x - \mu_y\right)\pi + \mu_y + k\sigma_y^2 \end{aligned} \tag{2.2}$$

Theorem 2.1. The estimator $\hat{\pi}_{new}$ is an unbiased estimator of π.

$$\hat{\pi}_{new} = \frac{\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu_y - k\sigma_y^2}{\left(\mu_x - \mu_y\right)} \tag{2.3}$$

Proof: Let E_1 denote the expected value over the possible sample. Then

$$E\left(\hat{\pi}_{new}\right) = E_1E_2\left[\hat{\pi}_{new}\right] = E_1E_2\left[\frac{\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu_y - k\sigma_y^2}{\left(\mu_x - \mu_y\right)}\right]$$

$$= E_1\left[\frac{\frac{1}{n}\sum_{i=1}^{n} E_2(Z_i) - \mu_y - k\sigma_y^2}{\left(\mu_x - \mu_y\right)}\right] = E_1\left[\frac{\frac{1}{n}\sum_{i=1}^{n}\left[\left(\mu_x - \mu_y\right)\pi + \mu_y + k\sigma_y^2\right] - \mu_y - k\sigma_y^2}{\left(\mu_x - \mu_y\right)}\right]$$

$$= \pi$$

Theorem 2.2.The variance of the estimator $\hat{\pi}_{new}$ is given by

$$\min.V\left(\hat{\pi}_{new}\right) = \frac{\pi(1-\pi)}{n} + \frac{\pi\sigma_x^2(1-\pi)\sigma_y^2}{rn\left(\mu_x - \mu_y\right)^2}$$

$$-\frac{\left\{\pi\mu_{30}+(1-\pi)\mu_{03}-\pi^2\mu_x\sigma_x^2-(1-\pi)^2\mu_y\sigma_y^2-\pi(1-\pi)\{\mu_y\sigma_x^2+\mu_x\sigma_y^2\}\right\}^2}{rn\left\{\pi(\mu_{40}-\sigma_x^4)+(1-\pi)(\mu_{04}-\sigma_y^4)-\{\pi\sigma_x^2+(1-\pi)\sigma_y^2\}^2\right\}(\mu_x-\mu_y)^2} \quad (2.4)$$

where $\mu_{ab}=E(X-E(X))^a(Y-E(Y))^b$ for a and b being non-negative inters, $\sigma_x^2=\mu_{20}$ and $\sigma_y^2=\mu_{02}$.

Proof: Let V_1 and V_2 denote the variance over the possible sample and over the randomization device, we have

$$V\left(\hat{\pi}_{new}\right)=E_1V_2\left(\hat{\pi}_{new}\right)+V_1E_2\left(\hat{\pi}_{new}\right)$$

$$=E_1V_2\left[\frac{\frac{1}{n}\sum_{i=1}^{n}Z_i-\mu_y-k\sigma_y^2}{(\mu_x-\mu_y)}\right]+V_1E_2\left[\frac{\frac{1}{n}\sum_{i=1}^{n}Z_i-\mu_y-k\sigma_y^2}{(\mu_x-\mu_y)}\right]$$

$$=E_1\left[\frac{\frac{1}{n^2}\sum_{i=1}^{n}V_2\left(Z_i\right)}{\left(\mu_x-\mu_y\right)^2}\right]+V_1\left(\pi\right)=\frac{\sigma_2^2}{n\left(\mu_x-\mu_y\right)^2} \quad (2.5)$$

Now, we have

$$\sigma_2^2=E_2\left(Z_i^2\right)-\left(E_2\left(Z_i\right)\right)^2$$

$$=\pi E_2\left(\bar{x}^2+ks_x^2\right)^2+\left(1-\pi\right)E_2\left(\bar{y}^2+ks_y^2\right)^2-\left[\pi\left(\mu_x+k\sigma_x^2\right)+\left(1-\pi\right)\left(\mu_y+k\sigma_y^2\right)\right]^2$$

$$=\pi E_2\left[\bar{x}^2+k^2s_x^4+2k\bar{x}s_x^2\right]+\left(1-\pi\right)E_2\left[\bar{y}^2+k^2s_y^4+2k\bar{y}s_y^2\right]$$
$$-\left[\pi\mu_x+\left(1-\pi\right)\mu_y+k\left\{\pi\sigma_x^2+\left(1-\pi\right)\sigma_y^2\right\}\right]$$

$$=\pi\left[\left(\mu_x^2+\frac{\sigma_x^2}{r}\right)+\frac{k^2}{r}\left(\mu_{40}-\sigma_x^4\right)+\frac{2k}{r}\mu_{30}\right]+\left(1-\pi\right)\left[\left(\mu_y^2+\frac{\sigma_y^2}{r}\right)+\frac{k^2}{r}\left(\mu_{04}-\sigma_y^4\right)+\frac{2k}{r}\mu_{03}\right]$$

$$-\left[\left\{\pi\mu_x+\left(1-\pi\right)\mu_y\right\}^2+k^2\left\{\pi\sigma_x^2+\left(1-\pi\right)\sigma_y^2\right\}^2+2k\left\{\pi\mu_x+\left(1-\pi\right)\mu_y\right\}\left\{\pi\sigma_x^2+\left(1-\pi\right)\sigma_y^2\right\}\right]$$

or

$$\sigma_2^2 = \pi\mu_x^2 + (1-\pi)\mu_y^2 - \{\pi\mu_x + (1-\pi)\mu_y\}^2 + \frac{\pi\sigma_x^2}{r} + \frac{(1-\pi)}{r}\sigma_y^2$$
$$+ \frac{k^2}{r}\left[\pi(\mu_{40} - \sigma_x^4) + (1-\pi)(\mu_{04} - \sigma_y^4) - \{\pi\sigma_x^2 + (1-\pi)\sigma_y^2\}^2\right]$$
$$+ \frac{2k}{r}\left[\pi\mu_{30} + (1-\pi)\mu_{03} - \{\pi\mu_x + (1-\pi)\mu_y\}\{\pi\sigma_x^2 + (1-\pi)\sigma_y^2\}\right]$$

$$= \pi(1-\pi)(\mu_x - \mu_y)^2 + \frac{\pi\sigma_x^2}{r} + \frac{(1-\pi)}{r}\sigma_y^2$$

$$+ \frac{k^2}{r}\left[\pi\mu_{40} - \pi\sigma_x^4 + (1-\pi)\mu_{40} - (1-\pi)\sigma_y^4 - \pi^2\sigma_x^4 - (1-\pi)^2\sigma_y^4 - 2\pi(1-\pi)\sigma_x^2\sigma_y\right]$$

$$+ \frac{2k}{r}\left[\pi\mu_{30} + (1-\pi)\mu_{03} - \{\pi^2\mu_x\sigma_x^2 + \pi(1-\pi)\mu_y\sigma_x^2 + \pi(1-\pi)\mu_x\sigma_y^2 + (1-\pi)^2\mu_y\sigma_y^2\}\right]$$

$$= \pi(1-\pi)(\mu_x - \mu_y)^2 + \frac{\pi\sigma_x^2}{r} + \frac{(1-\pi)}{r}\sigma_y^2$$

$$+ \frac{k^2}{r}\left[\pi(\mu_{40} - \sigma_x^4) + (1-\pi)(\mu_{40} - \sigma_y^4) - \{\pi\sigma_x^2 + (1-\pi)\sigma_y^2\}^2\right]$$

$$+ \frac{2k}{r}\left[\pi\mu_{30} + (1-\pi)\mu_{03} - \pi^2\mu_x\sigma_x^2 - \pi(1-\pi)\{\mu_y\sigma_x^2 + \mu_x\sigma_y^2\} - (1-\pi)^2\mu_y\sigma_y^2\right] \quad (2.6)$$

On substituting (2.6) into (2.5), the variance of the proposed estimator $\hat{\pi}_{new}$ is given by:

$$V(\hat{\pi}_{new}) = \frac{\pi(1-\pi)}{n} + \frac{\pi\sigma_x^2 + (1-\pi)\sigma_y^2}{rn(\mu_x - \mu_y)^2} + \frac{k^2\left[\pi(\mu_{40} - \sigma_x^4) + (1-\pi)(\mu_{04} - \sigma_y^4) - \{\pi\sigma_x^2 + (1-\pi)\sigma_y^2\}^2\right]}{rn(\mu_x - \mu_y)^2}$$

$$+ \frac{2k\left[\pi\mu_{30} + (1-\pi)\mu_{03} - \pi^2\mu_x\sigma_x^2 - (1-\pi)^2\mu_y\sigma_y^2 - \pi(1-\pi)\{\mu_y\sigma_x^2 + \mu_x\sigma_y^2\}\right]}{rn\left\{\pi(\mu_{40} - \sigma_x^4) + (1-\pi)(\mu_{04} - \sigma_y^4) - \{\pi\sigma_x^2 + (1-\pi)\sigma_y^2\}^2\right\}(\mu_x - \mu_y)^2} \quad (2.7)$$

On setting the first derivative of $V(\hat{\pi}_{new})$ with respect to k equal to zero, the variance of the proposed estimated is minimized if:

$$k = \frac{\pi\mu_{30} + (1-\pi)\mu_{03} - \pi^2\mu_x\sigma_x^2 - (1-\pi)^2\mu_y\sigma_y^2 - \pi(1-\pi)\{\mu_y\sigma_x^2 + \mu_x\sigma_y^2\}}{\pi(\mu_{40} - \sigma_x^4) + (1-\pi)(\mu_{04} - \sigma_y^4) - \{\pi\sigma_x^2 + (1-\pi)\sigma_y^2\}^2} \quad (2.8)$$

On substituting (2.8) in (2.7), we get the theorem.

Note that if $k = 0$ (or $r = 1$), the proposed estimator has the same variance as that of the Franklin's (1989) estimator.

3. RELATIVE EFFICIENCY

The proposed estimator $\hat{\pi}_{new}$ will be more efficient than the Franklin's (1989) estimator $\hat{\pi}_f$ if

$$V(\hat{\pi}_{new}) < V(\hat{\pi}_f) \quad (3.1)$$

Assuming $\gamma_{20} = \gamma_{02} = \sigma_x^2 = \sigma_y^2$, and substituting (1.5) and (2.4) into (3.1), we have

$$\frac{\left\{\pi\mu_{30} + (1-\pi)\mu_{03} - \pi^2\mu_x\sigma_x^2 - (1-\pi)\mu_y\sigma_y^2 - \pi(1-\pi)\{\mu_y\sigma_x^2 + \mu_x\sigma_y^2\}\right\}^2}{\left\{\pi(\mu_{40} - \sigma_x^4) + (1-\pi)(\mu_{04} - \sigma_y^4) - \{\pi\sigma_x^2 + (1-\pi)\sigma_y^2\}^2\right\}(\mu_x - \mu_y)^2} > 0 \quad (3.2)$$

which will always be true. Thus, the proposed estimator $\hat{\pi}_{new}$ is likely to perform better than the estimator $\hat{\pi}_f$ at equal cooperation of the respondents. In order to see the magnitude of the relative efficiency of the proposed estimator $\hat{\pi}_{new}$ with respect to four different randomized response models, we define the percent relative efficiencies in (a) through (d) as follows:

(a) The percent relative efficiency, RE(1), of the proposed estimator $\hat{\pi}_{new}$ with respect to the Franklin's estimator:

$$RE(1) = \frac{V(\hat{\pi}_f)}{V(\hat{\pi}_{new})} \times 100\% \quad (3.3)$$

(b) The percent relative efficiency, RE(2), of the proposed estimator $\hat{\pi}_{new}$ with respect to the Franklin's estimator with r trials per respondent:

$$RE(2) = \frac{V(\hat{\pi}_{f(r)})}{V(\hat{\pi}_{new})} \times 100\% \quad (3.4)$$

where $V(\hat{\pi}_{f(r)}) = \dfrac{\pi(1-\pi)}{n} + \dfrac{\pi\gamma_{20} + (1-\pi)\gamma_{02}}{nr(\theta_1 - \theta_2)^2}$. (3.5)

(c) The percent relative efficiency, RE(3), of the proposed estimator $\hat{\pi}_{new}$ with respect to the Warner (1965) estimator:

$$RE(3) = \frac{V(\hat{\pi}_w)}{V(\hat{\pi}_{new})} \times 100\% \tag{3.6}$$

(d) The percent relative efficiency, RE(4), of the proposed estimator $\hat{\pi}_{new}$ with respect to the Singh and Chen (2009) estimator:

$$RE(4) = \frac{V(\hat{\pi}_{sc})}{V(\hat{\pi}_{new})} \times 100\% \tag{3.7}$$

(e) The percent relative efficiency, RE (5), of the Singh and Chen (2009) estimator with respect to the Franklin (1989) estimator:

$$RE(5) = \frac{V(\hat{\pi}_f)}{V(\hat{\pi}_{sc})} \times 100\% \tag{3.8}$$

The results obtained for different values of parameters of a randomization device are presented in Table 1. The optimum value of α which minimizes the variance of the Singh and Chen (2009) estimator is also given in Table 1.

Discussion of the results: We decided to let $P = 0.8$ (which is a very reasonable and practical choice) in the Warner (1965) model while considering the problem of estimation of π with the proposed estimators $\hat{\pi}_{new}$. We suggest a privacy protection criterion as:

$$\lambda_{Z_k,i} = \frac{f(Z_k \mid k \in A)}{f(Z_k \mid k \notin A)} \tag{3.9}$$

which refers to the privacy protection with respect to the response Z_k for a respondent k being a member of A. For these measures $0 \le \lambda_{Z_k,i} < \infty$ applies with $\lambda_{Z_k,1} = 1$ indicating total protected data privacy for unit k being a member of group A. This means that the value Z_k contains absolutely no information on the variable of interest. The more the λ-measure differs from unity the more information on the variable under study is contained in the response, hence, less privacy protection. The maximum $\lambda_{Z_k,i} = \infty$ (or 0) describes a situation where one can conclude from the answer Z_k directly to the membership or the non-membership of A. A respondent would answer untruthfully or not answer at all in such a case.

In consideration to the proposed privacy protection criterion in (3.9), we decided to make a very practical choice of the known parameters of the scrambling variables as: $\theta_1 = 3.5$, $\theta_2 = 2$, $\gamma_{20} = 0.5$, and $\gamma_{02} = 1.5$ in the Franklin's model. By the one sigma empirical rule, most of the values of the scrambling variables S_1, and S_2 could,

respectively, be any real numbers in the ranges: $\left(3.5-\sqrt{0.5}, 3.5+\sqrt{0.5}\right)$ = (2.29, 3.71) and $\left(2-\sqrt{0.5}, 2+\sqrt{0.5}\right)$ = (1.29, 2.71), but the values are not 100% bounded to these intervals. Due to an overlap between these three intervals, although it will be hard to guess about the status of the respondents based on their reported responses, but still a smaller value close to 4.0 is likely to come from a person belonging to the sensitive group. To overcome this difficult, we have now developed a new optimal randomized response device in which each respondent is requested to draw $r=3$ random numbers, compute their mean and variance are report the response as $\bar{x}+ks_x^2$ if he/she belongs to group A, and report the response $\bar{y}+ks_y^2$. Further we assumed that $\sigma_y^2=\sigma_x^2=1.5$ which will make more difficult to guess if a respondent belongs to the group A or A^c, and hence more cooperation is expected from the respondents. As said earlier the value of k is provided by the interviewer based on his/her past experience related to the survey under investigation, thus it will be remain helpful to conduct to pilot survey of do some simulation study to guess the value of k or will remain more useful in repeated surveys over time. The optimum values of k which could be used in real practice are also given in Table 1. To investigate the relative efficiency values, we wrote a SAS code. The value of π was allowed to change as 0.1, 0.3, and 0.5, because the results corresponding to 0.7 and 0.9 can be obtained by the symmetry of results. The values of the third moments μ_{30} (and μ_{03}) were allowed to change between -3 to +3 with a step of 3; and the values of the fourth moment μ_{40} (and μ_{04}) were allowed to change between 2.5 to 28.5 with a step of 10. Also it was made sure that $(\mu_{40}-\sigma_x^4)>0$ and $(\mu_{04}-\sigma_y^4)>0$ while computing the variance of the proposed estimator. The proposed estimator shows efficient results for many combinations, thus only limited results were printed then the proposed estimator was showing efficiency more than hundred percent but less than a thousand percent. The results so stored in the file are presented in table 1, which we further summarize as follows: For $\pi=0.1$, if $\mu_{30}=\mu_{03}=-3$, $\mu_{40}=2.5$ and $\mu_{04}=23.5$, then the RE(1) value is 441.7%, RE(2) value is 184.4%, RE(3) value is 331.4%, RE(4) value is 254.2% which indicate the proposed estimator $\hat{\pi}_{new}$ remains more efficient the rest of four estimators for optimum values of $k=-0.3683$. The value of RE(5) remains 173.7% which shows Singh and Chen (2009) estimator remains more efficient than the Franklin's estimator By keeping the same values of μ_{30} and μ_{03}, as soon as the values of the fourth moments of the scrambling variables become $\mu_{40}=13.0$ and $\mu_{04}=23.5$, the values of RE(1), RE(2), RE(3) and RE(4) respectively become 418.8%, 174.9%, 314.2%, and 248.6% for the optimum value of $k=-0.3468$.The value of RE(5) becomes 168.5% for the optimum value of $\alpha=0.7837$. For the choice of parameters of the scrambling variables considered in this paper, for $\pi=0.1$, the value of RE(1) lies between 251.1% and 441.7% with median efficiency of 307.5%; the value of RE(2) lies between 105.3% and 184.4% with median efficiency of 128.4%; the value of RE(3) lies between 189.1% to 331.4% with median efficiency of 230.7%; the RE(4) value lies between 248.7% to 254.2% with a median value of 291.1% and the optimum value of k lies between -0.1539 to -0.3926 with a median value of -0.3276. Thus based on a good guess about the value of π close to 0.1,

the interviewer can choose the value of k between -0.39 to -0.15 depending upon other parameters of the scrambling variables as shown in table 1. Also note that the value of RE(5) lies between 102.3% to 173.7% with a median value of 106.0%, and the optimum value of α lies between 0.1448 to 0.8184 with a median value of 0.4054. Similar values of α are reported in Singh and Chen (2009) estimator while suggesting to utilize higher order moments of the scrambling variables. In the same way, the results for other values of π listed in table 1 can be interpreted. Note that the proposed estimator $\hat{\pi}_{new}$ remains more efficient than the other four estimators for the value of π being close to zero, which is more practical because the proportion of a sensitive character is expected to be close to zero in real surveys.

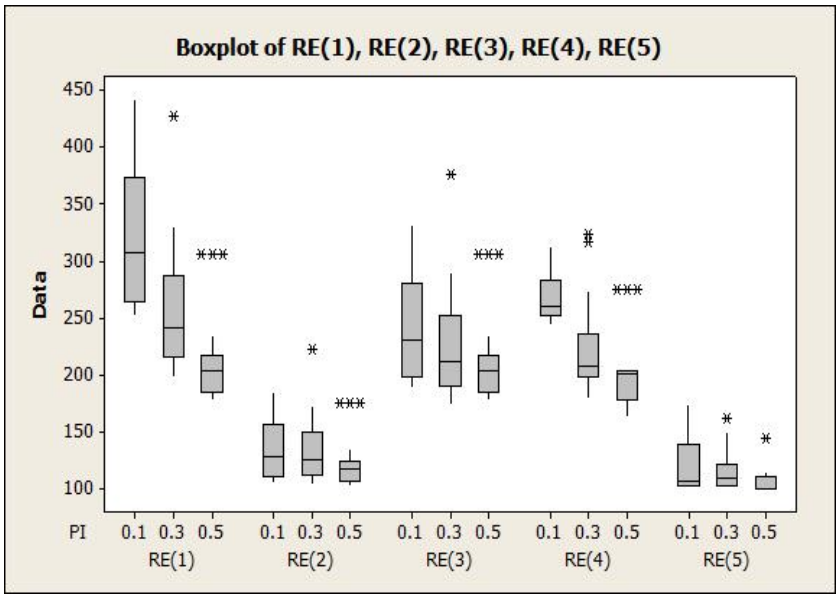

Fig. 1. Graphical representation of the percent relative efficiencies

From the box plots in Figure 1, it is clear median of that RE(1) value remains higher than the median values of the corresponding other relative efficiency values for different values of π between 0.1, 0.3 and 0.5 with a step of 0.2. The box plots for RE(1), RE(2) and RE(3) values show one outlier value for $\pi = 0.3$ and two outliers for $\pi = 0.5$, however the box plot for RE(4) values shows one extra outlier for $\pi = 0.3$. The box plot for RE(5) shows only two outlier values. It is worth noting that the outliers are on the upper inner fences of the box plots showing that in certain cases the proposed estimator remains extremely more efficient than the competitors. The outliers for RE(5) values show that the Singh and Chen (2009) estimator also shows sometimes extremely better results than the Franklin's estimators.

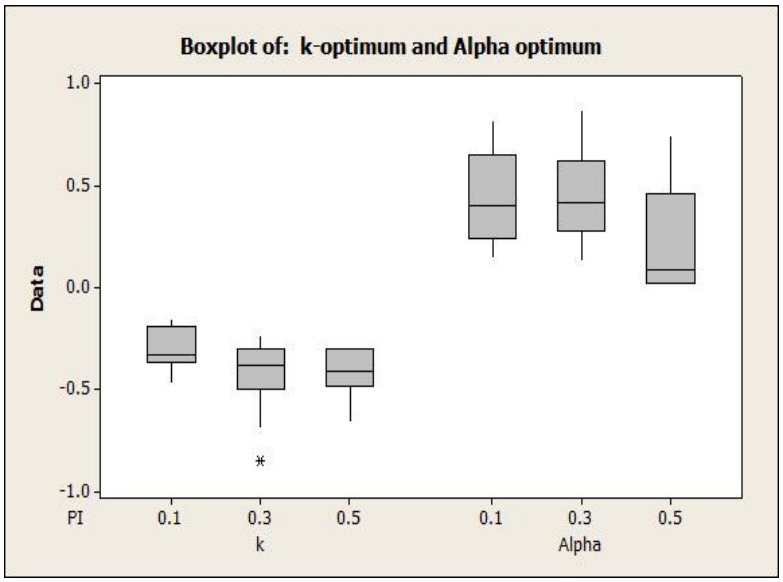

Fig. 2. Graphical representation of the optimum values of k and α

Figure 2 shows the box plots of the optimum values of k to be used in the proposed estimator and optimum value of α to be used in the Chen and Singh (2009) estimator. The box plots for the optimum values of α are skewed to the right indicating that most of the values of α remain close to one as indicated by Singh and Chen (2009). The box plots for the optimum values of k are also skewed to the right, but whiskers are longer to the opposite side of the skewness indicates that the optimum value of k remains close to negative of 0.5. Thus, we conclude that the new proposed optimal method can be used to obtain better estimate of the population proportion by making use of an appropriate choice of scrambling variables.

References

Fox, J.A. and Tracy, P.E. (1986). *Randomized response: A method for sensitive surveys*. SAGE Publications.

Franklin, L.A. (1989). A comparison of estimators for randomized response sampling with continuous distributions from a dichotomous population. *Commun. Statist.- Theory Meth.*, 18 :489-505.

Gjestvang, C. R. and Singh, S. (2006). A new randomized response model. *Journal of the Royal Statistical Society , B,* 68 :523-530.

Kerkvliet, J. (1994). Estimating a logit model with randomized data: The case of cocaine use. *Austral. J. Statist.*,36 :9-20.

Kuk, A.Y.C. (1990). Asking sensitive questions indirectly. *Biomerika*, 77(2) :436-438.

Mangat, N.S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society*, B, 56 :93-95.

Mangat, N.S. and Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, 77(2) :439-442.

Odumade, O. and Singh, S. (2010). An alternative to the Bar-Lev, Bobovitch and Boukai randomized response model. *Sociological Methods & Research*, 39(2), 206-221.

Singh, S. and Chen, C.C. (2009). Utilization of higher order moments of scrambling variables in randomized response sampling. *Journal of Statistical Planning and Inference,*139, 3377-3380.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60:63-69

Table 1. Comparison of the five estimators for different values of the parameters.

π	μ_{30}	μ_{03}	μ_{40}	μ_{04}	RE(1)	RE(2)	RE(3)	RE(4)	RE(5)	k	α
0.1	-3.0	-3.0	2.5	23.5	441.7	184.4	331.4	254.2	173.7	-0.3683	0.8184
			13.0	23.5	418.8	174.9	314.2	248.6	168.5	-0.3468	0.7837
			23.5	23.5	400.3	167.2	300.4	244.2	163.9	-0.3276	0.7518
		0.0	2.5	13.0	351.9	146.9	264.1	312.9	112.5	-0.4732	0.6512
			2.5	23.5	270.0	112.7	202.6	254.8	106.0	-0.2086	0.3311
			13.0	13.0	329.8	137.7	247.5	296.8	111.1	-0.4147	0.5880
			13.0	23.5	267.1	111.5	200.4	252.9	105.6	-0.1964	0.3139
			23.5	13.0	314.4	131.3	235.9	285.7	110.0	-0.3691	0.5360
			23.5	23.5	264.6	110.5	198.6	251.2	105.3	-0.1855	0.2985
	0.0	-3.0	2.5	23.5	405.9	169.5	304.6	266.1	152.6	-0.3506	0.7179
			13.0	23.5	388.2	162.1	291.3	259.9	149.4	-0.3301	0.6890
			23.5	23.5	373.7	156.1	280.5	255.0	146.6	-0.3118	0.6623

		0.0	2.5	13.0	323.3	135.0	242.6	304.5	106.2	-0.4329	0.4438
			2.5	23.5	262.1	109.5	196.7	253.9	103.2	-0.1908	0.2394
			13.0	13.0	307.5	128.4	230.7	291.1	105.6	-0.3794	0.4054
			13.0	23.5	259.9	108.5	195.0	252.1	103.1	-0.1797	0.2278
			23.5	13.0	296.1	123.7	222.2	281.6	105.2	-0.3377	0.3730
			23.5	23.5	257.9	107.7	193.5	250.5	102.9	-0.1697	0.2172
	3.0	-3.0	2.5	23.5	376.9	157.4	282.9	272.6	138.3	-0.3328	0.6274
			13.0	23.5	363.1	151.6	272.5	266.5	136.3	-0.3134	0.6033
			23.5	23.5	351.6	146.8	263.8	261.5	134.5	-0.2961	0.5810
		0.0	2.5	13.0	301.0	125.7	225.9	293.2	102.7	-0.3926	0.2800
			2.5	23.5	255.4	106.6	191.6	251.6	101.5	-0.1731	0.1587
			13.0	13.0	289.6	120.9	217.3	282.7	102.5	-0.3441	0.2581
			13.0	23.5	253.6	105.9	190.3	250.1	101.4	-0.1630	0.1514
			23.5	13.0	281.3	117.4	211.1	275.0	102.3	-0.3063	0.2393
			23.5	23.5	252.1	105.3	189.1	248.7	101.4	-0.1539	0.1448
0.3	-3.0	-3.0	13.0	23.5	300.8	156.9	264.8	185.9	161.8	-0.4211	0.8695
			23.5	23.5	267.6	139.6	235.6	179.5	149.1	-0.3513	0.7498
		0.0	2.5	23.5	229.6	119.8	202.2	207.6	110.6	-0.3602	0.5357
			13.0	13.0	275.2	143.6	242.3	236.3	116.5	-0.5382	0.7889
			13.0	23.5	215.3	112.3	189.5	198.6	108.4	-0.2886	0.4317
			23.5	13.0	236.8	123.5	208.5	212.1	111.7	-0.3927	0.5824
			23.5	23.5	206.7	107.8	182.0	193.3	106.9	-0.2408	0.3616
	0.0	-3.0	2.5	23.5	286.7	149.5	252.4	227.9	125.8	-0.4547	0.6253
			13.0	13.0	427.5	223.0	376.4	317.3	134.8	-0.6794	0.7866
			13.0	23.5	253.1	132.0	222.8	208.1	121.6	-0.3644	0.5420
			23.5	13.0	305.0	159.1	268.5	239.1	127.6	-0.4957	0.6591
			23.5	23.5	234.8	122.5	206.7	198.0	118.6	-0.3039	0.4782
		0.0	2.5	13.0	279.4	145.8	246.0	274.2	101.9	-0.6869	0.2760
			2.5	23.5	205.2	107.1	180.7	203.0	101.1	-0.2894	0.1565
			13.0	13.0	226.9	118.4	199.8	223.7	101.4	-0.4324	0.2080
			13.0	23.5	197.6	103.1	174.0	195.9	100.9	-0.2319	0.1320
			23.5	2.5	329.8	172.0	290.3	322.8	102.2	-0.8547	0.3097
			23.5	13.0	208.9	109.0	183.9	206.5	101.2	-0.3155	0.1668
	3.0	-3.0	2.5	23.5	240.6	125.5	211.8	222.5	108.1	-0.3839	0.3469
			13.0	13.0	299.6	156.3	263.8	272.4	110.0	-0.5735	0.4191
			13.0	23.5	222.9	116.3	196.2	208.1	107.1	-0.3076	0.3072
			23.5	13.0	249.5	130.2	219.7	229.9	108.5	-0.4185	0.3625
			23.5	23.5	212.5	110.8	187.1	199.8	106.4	-0.2566	0.2757
0.5	-3.0	-3.0	23.5	23.5	234.4	134.4	234.4	162.9	143.9	-0.3750	0.7472
		0.0	13.0	23.5	216.5	124.2	216.5	188.1	115.2	-0.4091	0.6276
			23.5	13.0	216.5	124.2	216.5	188.1	115.2	-0.4091	0.6276

			23.5	23.5	191.5	109.8	191.5	173.9	110.1	-0.2961	0.4387
		3.0	2.5	23.5	204.0	117.0	204.0	203.8	100.1	-0.4853	0.0906
			13.0	13.0	204.0	117.0	204.0	203.8	100.1	-0.4853	0.0906
			13.0	23.5	177.7	101.9	177.7	177.6	100.1	-0.3000	0.0415
			23.5	2.5	204.0	117.0	204.0	203.8	100.1	-0.4853	0.0906
			23.5	13.0	177.7	101.9	177.7	177.6	100.1	-0.3000	0.0415
	0.0	-3.0	2.5	23.5	305.8	175.3	305.8	275.0	111.2	-0.6618	0.4800
			13.0	13.0	305.8	175.3	305.8	275.0	111.2	-0.6618	0.4800
			13.0	23.5	216.5	124.2	216.5	200.1	108.2	-0.4091	0.3611
			23.5	2.5	305.8	175.3	305.8	275.0	111.2	-0.6618	0.4800
			23.5	13.0	216.5	124.2	216.5	200.1	108.2	-0.4091	0.3611
			23.5	23.5	191.5	109.8	191.5	179.9	106.5	-0.2961	0.2894
		0.0	2.5	23.5	204.0	117.0	204.0	203.9	100.0	-0.4853	0.0300
			13.0	13.0	204.0	117.0	204.0	203.9	100.0	-0.4853	0.0300
			13.0	23.5	177.7	101.9	177.7	177.7	100.0	-0.3000	0.0215
			23.5	2.5	204.0	117.0	204.0	203.9	100.0	-0.4853	0.0300
			23.5	13.0	177.7	101.9	177.7	177.7	100.0	-0.3000	0.0215
	3.0	-3.0	2.5	23.5	204.0	117.0	204.0	204.0	100.0	-0.4853	0.0179
			13.0	13.0	204.0	117.0	204.0	204.0	100.0	-0.4853	0.0179
			13.0	23.5	177.7	101.9	177.7	177.7	100.0	-0.3000	0.0145
			23.5	2.5	204.0	117.0	204.0	204.0	100.0	-0.4853	0.0179
			23.5	13.0	177.7	101.9	177.7	177.7	100.0	-0.3000	0.0145

Remark: Results for $\pi = 0.7$, and $\pi = 0.9$ can be obtained from $\pi = 0.3$ and $\pi = 0.1$ respectively.