Weighted Squared Distance for Two-Decks Randomized Response Model

Augustus Jayaraj[1]

Oluseun Odumade[2]

Stephen Sedory[3] and Sarjinder Singh[3]

[1]Cornell University
[2]Deloitte Consulting
[3]Department of Mathematics
Texas A&M University-Kingsville
Kingsville, TX 78363, USA

Abstract

In this paper, we propose consider a new weighted squared distance while minimizing a distance between the true proportions and the observed proportions of (Yes, Yes), (Yes, No), (No, Yes) and (No, No) answers in the set-up of Odumade and Singh (2009) model. The resultant estimator is shown to be unbiased estimator of the proportion of the sensitive attribute of interest in a population and has smaller variance than the estimator of Odumade and Singh (2009) with the same protection of the respondents.

Keywords: Unrelated question model, two-decks of cards, lower bound of variance and protection of respondents.

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 the case where 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 sample of n respondents is selected with replacement from the population. For collecting information on the 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_0 and $(1-P_0)$ 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 people n_1 that answer "Yes" is binomially distributed with parameters $P_0\pi+(1-P_0)(1-\pi)$ and n. The maximum likelihood estimator of π exists for $P_0 \neq 0.5$ and is given by:

$$\hat{\pi}_w = \frac{\frac{n_1}{n} - (1-P_0)}{2P_0 - 1} \tag{1.1}$$

The above estimator is unbiased with variance:

$$V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{P_0(1-P_0)}{n(2P_0-1)^2} \tag{1.2}$$

Mangat and Singh (1990) suggested a two-stage randomized response model. In the first stage each respondent is requested to use a randomization device, R_1^*, such as a deck of cards with each card having written one of the following two statements:

(i) "I belong to group A" (ii) "Go to the randomization device R_2^*"

The statements occur with relative frequencies T_0 and A respectively in the first device R_1^*. In the second stage, if directed by the outcome of R_1^*, the respondent is requested to use the randomization device R_2^* which is the same as the Warner (1965) device. Under the two-stage randomized response model, an unbiased estimator of the population proportion π is given by:

$$\hat{\pi}_{ms} = \frac{\frac{n_1}{n} - (1-T_0)(1-P_0)}{(2P_0-1) + 2T_0(1-P_0)} \tag{1.3}$$

with variance:

$$V(\hat{\pi}_{ms}) = \frac{\pi(1-\pi)}{n} + \frac{(1-P_0)(1-T_0)[1-(1-P_0)(1-T_0)]}{n\{(2P_0-1)+2T_0(1-P_0)\}^2} \tag{1.4}$$

Mangat (1994) considered another randomized response model where each respondent selected in the sample is requested to report “Yes” if he/she belongs to the sensitive group A otherwise he/she is instructed to use the Warner (1965) device. For this model an unbiased estimator of the population proportion π is given by:

$$\hat{\pi}_m = \frac{\frac{n_1}{n} - (1-P_0)}{P_0} \tag{1.5}$$

with variance given by:

$$V(\hat{\pi}_m) = \frac{\theta_m(1-\theta_m)}{nP_0^2} \tag{1.6}$$

where $\theta_m = \pi + (1-\pi)(1-P_0)$.

Mangat (1994) model remains more efficient than both the Warner (1965) and Mangat and Singh (1990) models. Note that the Mangat (1994) model has been improved by Gjestvang and Singh (2006).

2. Odumade and Singh (2009) Model

Each respondent in the simple random and with replacement (SRSWR) sample of n is provided with two decks of cards marked as Deck-I and Deck-II as shown in Figure 2.1.

Deck-I
$I \in A$ with probability P
$I \in A^c$ with probability $(1-P)$

Deck-II
$I \in A$ with probability T
$I \in A^c$ with probability $(1-T)$

Fig. 2.1. Two decks of cards

Each respondent is requested to draw two cards simultaneously, one card from reach deck of cards, and read the statements in order. The respondent first matches his/her status with the statement written on the first deck of cards, and then he/she matches his/her status with the statement written on the second deck of cards. Let π be the true proportion of respondents in the population that possesses the characteristic A.

Consider a situation that the selected respondent belongs to group A: Now if he/she draws first card with statement $I \in A$ with probability P from the first deck of cards and second card with statement $I \in A$ with probability T from the second deck of cards, then he/she is requested to report: (Yes, Yes).

Consider another situation that the selected respondent belongs to group A^c: Now if he/she draws first card with statement $I \in A^c$ with probability $(1-P)$ from the first deck of cards and second card with statement $I \in A^c$ with probability $(1-T)$ from the second deck of cards, then he/she is also requested to report: (Yes, Yes). Thus the response (Yes, Yes) can come from both types of respondents either belonging to the group A or A^c and hence their privacy will be maintained. Thus, the probability of getting (Yes, Yes) response is given by:

$$P(\text{Yes, Yes}) = \theta_{11} = PT\pi + (1-P)(1-T)(1-\pi) = (P+T-1)\pi + (1-P)(1-T) \tag{2.1}$$

Now consider a situation that the selected respondent belongs to group A: Now if he/she draws first card with statement $I \in A$ with probability P from the first deck of cards and second card with statement $I \in A^c$ with probability $(1-T)$ from the second deck of cards, then he/she is requested to report: (Yes, No).

Consider another situation that the selected respondent belongs to group A^c: Now if he/she draws first card with statement $I \in A^c$ with probability $(1-P)$ from the first deck of cards and second card with statement $I \in A$ with probability T from the second deck of cards, then he/she is also requested to report: (Yes, No). Thus the response (Yes, No) can come from both types of respondents either belonging to the group A or A^c and hence their privacy will not be disclosed.

Thus, the probability of getting (Yes, No) response is given by:

$$P(\text{Yes, No}) = \theta_{10} = P(1-T)\pi + (1-P)T(1-\pi) = (P-T)\pi + T(1-P) \qquad (2.2)$$

Now consider a situation that the selected respondent belongs to group A: Now if he/she draws first card with statement $I \in A^c$ with probability $(1-P)$ from the first deck of cards and second card with statement $I \in A$ with probability T from the second deck of cards, then he/she is requested to report: (No, Yes).

Consider another situation that the selected respondent belongs to group A^c: Now if he/she draws first card with statement $I \in A$ with probability P from the first deck of cards and second card with statement $I \in A^c$ with probability $(1-T)$ from the second deck of cards, then he/she is also requested to report: (No, Yes). Thus the response (No, Yes) can come from both types of respondents either belonging to the group A or A^c and hence their privacy will not be disclosed.

Thus, the probability of getting (No, Yes) response is given by:

$$P(\text{No, Yes}) = \theta_{01} = (1-P)T\pi + P(1-T)(1-\pi) = (T-P)\pi + P(1-T) \qquad (2.3)$$

Now consider a situation that the selected respondent belongs to group A: Now if he/she draws first card with statement $I \in A^c$ with probability $(1-P)$ from the first deck of cards and second card with statement $I \in A^c$ with probability $(1-T)$ from the second deck of cards, then he/she is requested to report: (No, No).

Consider another situation that the selected respondent belongs to group A^c: Now if he/she draws first card with statement $I \in A$ with probability P from the first deck of cards and second card with statement $I \in A$ with probability T from the second deck of cards, then he/she is also requested to report: (No, No). Thus the response (No, No) can come from both types of respondents either belonging to the group A or A^c and hence their privacy will not be disclosed.

Thus, the probability of getting (No, No) response is given by:

$$P(\text{No, No}) = \theta_{00} = (1-P)(1-T)\pi + PT(1-\pi) = (1-P-T)\pi + PT \qquad (2.4)$$

The responses from the n respondents can be classified in to 2×2 contingency table as shown in Table 2.1.

Responses	*Yes*	*No*
Yes	n_{11}	n_{10}
No	n_{01}	n_{00}

Table 2.1. The 2×2 contingency table.

The true probabilities of (Yes, Yes), (Yes, No), (No, Yes) and (No, No) responses in the population can be classified in a 2×2 contingency table as shown in Table 2.2.

Two Decks	True probabilities	Deck – II	
		T	$(1-T)$
Deck-I	P	θ_{11}	θ_{10}
	$(1-P)$	θ_{01}	θ_{00}

Table 2.2. The 2×2 contingency table.

where θ_{11}, θ_{10} , θ_{01} and θ_{00} are given in (2.1), (2.2), (2.3) and (2.4) respectively. Remember that our aim is to estimate the unknown population proportion π of the respondents belonging to the group A.

Let $\hat{\theta}_{11}=n_{11}/n$, $\hat{\theta}_{10}=n_{10}/n$, $\hat{\theta}_{01}=n_{01}/n$ and $\hat{\theta}_{00}=n_{00}/n$ be the observed proportions of (Yes, Yes), (Yes, No), (No, Yes) and (No, No) responses. Odumade and Singh (2009) define squared distance between the observed proportions and the true proportions as:

$$\begin{aligned} D &= \frac{1}{2}\sum_{i=0}^{1}\sum_{j=0}^{1}\left(\theta_{ij}-\hat{\theta}_{ij}\right)^2 \\ &= \frac{1}{2}\left[\theta_{11}-\hat{\theta}_{11}\right]^2+\frac{1}{2}\left[\theta_{10}-\hat{\theta}_{10}\right]^2+\frac{1}{2}\left[\theta_{01}-\hat{\theta}_{01}\right]^2+\frac{1}{2}\left[\theta_{00}-\hat{\theta}_{00}\right]^2 \\ &= \frac{1}{2}\left[(P+T-1)\pi+(1-P)(1-T)-\hat{\theta}_{11}\right]^2+\frac{1}{2}\left[(P-T)\pi+T(1-P)-\hat{\theta}_{10}\right]^2 \\ &\quad +\frac{1}{2}\left[(T-P)\pi+P(1-T)-\hat{\theta}_{01}\right]^2+\frac{1}{2}\left[(1-P-T)\pi+PT-\hat{\theta}_{00}\right]^2 \end{aligned} \tag{2.5}$$

Odumade and Singh (2009) decided to choose π such that the least square distance D is minimum. Thus, to find such a choice of π they set:

$$\frac{\partial D}{\partial \pi}=0$$

or

$$\left[(P+T-1)\pi+(1-P)(1-T)-\hat{\theta}_{11}\right](P+T-1)+\left[(P-T)\pi+T(1-P)-\hat{\theta}_{10}\right](P-T)$$
$$+\left[(T-P)\pi+P(1-T)-\hat{\theta}_{01}\right](T-P)+\left[(1-P-T)\pi+PT-\hat{\theta}_{00}\right](1-P-T)=0$$

or

$$\pi=\frac{1}{2}+\frac{(P+T-1)(\hat{\theta}_{11}-\hat{\theta}_{00})+(P-T)(\hat{\theta}_{10}-\hat{\theta}_{01})}{2\left[(P+T-1)^2+(P-T)^2\right]} \tag{2.6}$$

By the method of moments, they suggested an unbiased estimator of the population proportion π is given by:

$$\hat{\pi}_{os}=\frac{1}{2}+\frac{(P+T-1)\left[\hat{\theta}_{11}-\hat{\theta}_{00}\right]+(P-T)\left[\hat{\theta}_{10}-\hat{\theta}_{01}\right]}{2\left[(P+T-1)^2+(P-T)^2\right]} \tag{2.7}$$

and the variance of the estimator $\hat{\pi}_{os}$ is given by:

$$V(\hat{\pi}_{os})=\frac{(P+T-1)^2\{PT+(1-P)(1-T)\}+(P-T)^2\{T(1-P)+P(1-T)\}}{4n\left[(P+T-1)^2+(P-T)^2\right]^2}-\frac{(2\pi-1)^2}{4n} \tag{2.8}$$

They also suggested an unbiased estimator to estimate the variance of $\hat{\pi}_{os}$ is given by:

$$\hat{v}(\hat{\pi}_{os})=\frac{1}{4(n-1)}\left[\frac{(P+T-1)^2\{PT+(1-P)(1-T)\}+(P-T)^2\{T(1-P)+P(1-T)\}}{\{(P+T-1)^2+(P-T)^2\}}-(2\hat{\pi}_{os}-1)^2\right] \tag{2.9}$$

Note that if $T=P=P_0$ (say), then the variance of the proposed estimator $\hat{\pi}_{os}$ in (2.8) becomes:

$$V(\hat{\pi}_{os})_{P=T=P_0}=\frac{\pi(1-\pi)}{n}+\frac{P_0(1-P_0)}{2n(2P_0-1)^2}=V(\hat{\pi}_w)_{q=2} \tag{2.10}$$

which is same variance if each respondent is requested to use the Warner (1965) device twice.

Singh and Sedory (2011, 2012) suggested a new log-likelihood estimator of the population proportion π and developed a lower bound on the variance in this randomized response sampling setup. They consider the problem of maximizing the likelihood function, which is defined as:

$$L=\binom{n}{n_{11},n_{10},n_{01},n_{00}}\theta_{11}^{n_{11}}\theta_{10}^{n_{10}}\theta_{01}^{n_{01}}\theta_{00}^{n_{00}} \tag{2.11}$$

On setting $\frac{\partial \log(L)}{\partial \pi} = 0$, the maximum likelihood estimate $\hat{\pi}_{\text{mle}}$ of π is given by a solution to the following equation:

$$\frac{\hat{\theta}_{11}(P+T-1)}{\theta_{11}} + \frac{\hat{\theta}_{10}(P-T)}{\theta_{10}} + \frac{\hat{\theta}_{01}(T-P)}{\theta_{01}} + \frac{\hat{\theta}_{00}(1-P-T)}{\theta_{00}} = 0 \tag{2.12}$$

By the well known Cramer-Rao inequality, the lower bound of the variance of the maximum likelihood estimate $\hat{\pi}_{mle}$ of π is given by:

$$V(\hat{\pi}_{mle}) \geq \frac{1}{n\left\{\frac{(P+T-1)^2}{\theta_{11}} + \frac{(P-T)^2}{\theta_{10}} + \frac{(T-P)^2}{\theta_{01}} + \frac{(1-P-T)^2}{\theta_{00}}\right\}} \tag{2.13}$$

In a very short span of time, Odumade and Singh (2009) model becoming popular among the randomized response technique developed and has many citations on the Google Scholar, and a few of them are listed as: Singh and Sedory (2011, 2012), Abdelfatahm, Mazloum, and Singh (2013), Arnab, Singh, and North (2012), Lee, Sedory, and Singh (2013a), Lee, Sedory, and Singh (2013b), Su, Sedory, and Singh (2015), Bacanli and Tuncel (2014),Chen and Singh (2011), Lee, Su, Mondragon, Salinas, Zamora, Sedory, and Singh (2016), Abdelfatah and Mazloum, (2015a, 2015b), Batool and Shabbir (2016), Arnab and Shangodoying (2016), Lee, Hong, Kim, and Son (2014), Arnab and Thuto (2015), Batool, Shabbir and Hussain (2015) and Fox (2016). Chaudhuri (2015) has contributed a special issue, and several improvements in different directions on this topic can be found in this issue. This motivated to think more on the data collected with Odumade and Singh (2009) model if it can be analyzed in another efficient way!

In the next section, we suggest a new method to derive a new unbiased and efficient estimator by making use of two decks.

3. Proposed Weighted Squared Distance Based Estimator

We consider a weighted squared distance (WD) between the true proportions and observed proportions as:

$$\begin{aligned}
\text{WD} &= \frac{1}{2}\left[\left(\frac{PT}{P+T-1}\right)(\theta_{11}-\hat{\theta}_{11})^2 + \frac{P(1-T)}{(P-T)}(\theta_{10}-\hat{\theta}_{10})^2 + \frac{(1-P)T}{(T-P)}(\theta_{01}-\hat{\theta}_{01})^2 + \frac{(1-P)(1-T)}{(1-P-T)}(\theta_{00}-\hat{\theta}_{00})^2\right] \\
&= \frac{1}{2}\left[\left(\frac{PT}{P+T-1}\right)\{(P+T-1)\pi + (1-P)(1-T) - \hat{\theta}_{11}\}^2 + \frac{P(1-T)}{(P-T)}\{(P-T)\pi + T(1-P)\}^2\right. \\
&\quad \left. + \frac{(1-P)T}{(T-P)}\{(T-P)\pi + P(1-T) - \hat{\theta}_{01}\}^2 + \frac{(1-P)(1-T)}{(1-P-T)}\{(1-T-P)\pi + PT - \hat{\theta}_{00}\}^2\right]
\end{aligned} \tag{3.1}$$

On setting

$$\frac{\partial \text{WD}}{\partial \pi} = 0$$

or

$$PT\left[(P+T-1)\pi+(1-P)(1-T)-\hat{\theta}_{11}\right]+P(1-T)\left[(P-T)\pi+T(1-P)-\hat{\theta}_{10}\right]$$
$$+(1-P)T\left[(T-P)\pi+P(1-T)-\hat{\theta}_{01}\right]+(1-P)(1-T)\left[(1-P-T)\pi+PT-\hat{\theta}_{00}\right]=0$$

or

$$\pi=\frac{PT\hat{\theta}_{11}+P(1-T)\hat{\theta}_{10}+(1-P)T\hat{\theta}_{01}+(1-P)(1-T)\hat{\theta}_{00}-4PT(1-P)(1-T)}{1-2P(1-P)-2T(1-T)}$$

Now we have the following theorem:

Theorem 3.1. An unbiased estimator of population proportion π is given by

$$\hat{\pi}_p=\frac{PT\hat{\theta}_{11}+P(1-T)\hat{\theta}_{10}+(1-P)T\hat{\theta}_{01}+(1-P)(1-T)\hat{\theta}_{00}-4PT(1-P)(1-T)}{1-2P(1-P)-2T(1-T)} \qquad (3.2)$$

Proof. Note that $n_{ij}\sim B(n,\theta_{ij})$, thus $E(\hat{\theta}_{11})=\theta_{11}$, $E(\hat{\theta}_{10})=\theta_{10}$, $E(\hat{\theta}_{01})=\theta_{01}$ and $E(\hat{\theta}_{01})=\theta_{00}$. Taking expected value on both sides of (3.2), we have

$$E(\hat{\pi}_p)=E\left[\frac{PT\hat{\theta}_{11}+P(1-T)\hat{\theta}_{10}+(1-P)T\hat{\theta}_{01}+(1-P)(1-T)\hat{\theta}_{00}-4PT(1-P)(1-T)}{1-2P(1-P)-2T(1-T)}\right]$$

$$=\frac{PTE(\hat{\theta}_{11})+P(1-T)E(\hat{\theta}_{10})+(1-P)TE(\hat{\theta}_{01})+(1-P)(1-T)E(\hat{\theta}_{00})-4PT(1-P)(1-T)}{1-2P(1-P)-2T(1-T)}$$

$$=\frac{1}{\{1-2P(1-P)-2T(1-T)\}}\big[PT\{(P+T-1)\pi+(1-P)(1-T)\}+P(1-T)\{(P-T)\pi+T(1-P)\}$$

$$+(1-P)T\{(T-P)\pi+P(1-T)\}++(1-P)(1-T)\{(1-P-T)\pi+PT\}-4PT(1-P)(1-T)\big]$$

$$=\frac{1}{\{1-2P(1-P)-2T(1-T)\}}\big[\{1-2P(1-P)-2T(1-T)\}\pi+4PT(1-P)(1-T)-4PT(1-P)(1-T)\big]$$

$$=\pi$$

which proves the theorem.

Theorem 3.2. The variance of the unbiased estimator $\hat{\pi}_p$ of population proportion π is given by

$$V(\hat{\pi}_p)=\frac{\pi(1-\pi)}{n}+\frac{(2P-1)^2(2T-1)^2\{P(1-P)+T(1-T)\}\pi}{n\{1-2P(1-P)-2T(1-T)\}^2}+\frac{PT(1-P)(1-T)\{1-16PT(1-P)(1-T)\}}{n\{1-2P(1-P)-2T(1-T)\}^2} \qquad (3.3)$$

Proof**.** Because $n_{ij}\sim B(n,\theta_{ij})$, now using the concept of binomial distribution, we have:

$$V(\hat{\theta}_{11}) = \frac{\theta_{11}(1-\theta_{11})}{n}, \quad Cov(\hat{\theta}_{11}, \hat{\theta}_{10}) = \frac{-\theta_{11}\theta_{10}}{n}, \quad Cov(\hat{\theta}_{10}, \hat{\theta}_{01}) = \frac{-\theta_{10}\theta_{01}}{n}$$

$$V(\hat{\theta}_{10}) = \frac{\theta_{10}(1-\theta_{10})}{n}, \quad Cov(\hat{\theta}_{11}, \hat{\theta}_{01}) = \frac{-\theta_{11}\theta_{01}}{n}, \quad Cov(\hat{\theta}_{10}, \hat{\theta}_{00}) = \frac{-\theta_{10}\theta_{00}}{n}$$

$$V(\hat{\theta}_{01}) = \frac{\theta_{01}(1-\theta_{01})}{n}, \quad Cov(\hat{\theta}_{11}, \hat{\theta}_{00}) = \frac{-\theta_{11}\theta_{00}}{n}, \quad Cov(\hat{\theta}_{01}, \hat{\theta}_{00}) = \frac{-\theta_{01}\theta_{00}}{n}$$

and

$$V(\hat{\theta}_{00}) = \frac{\theta_{00}(1-\theta_{00})}{n}.$$

By the definition of variance, we have

$$V(\hat{\pi}_p) = V\left[\frac{PT\hat{\theta}_{11} + P(1-T)\hat{\theta}_{10} + (1-P)T\hat{\theta}_{01} + (1-P)(1-T)\hat{\theta}_{00} - 4PT(1-P)(1-T)}{1-2P(1-P)-2T(1-T)}\right]$$

$$= \frac{1}{\{1-2P(1-P)-2T(1-T)\}^2}\Big[P^2T^2V(\hat{\theta}_{11}) + P^2(1-T)^2V(\hat{\theta}_{10}) + (1-P)^2T^2V(\hat{\theta}_{01})$$
$$+ (1-P)^2(1-T)^2V(\hat{\theta}_{00})$$
$$+ 2P^2T(1-T)Cov(\hat{\theta}_{11}, \hat{\theta}_{10}) + 2PT^2(1-P)Cov(\hat{\theta}_{11}, \hat{\theta}_{01})$$
$$+ 2PT(1-P)(1-T)Cov(\hat{\theta}_{11}, \hat{\theta}_{00}) + 2P(1-T)(1-P)TCov(\hat{\theta}_{10}, \hat{\theta}_{01})$$
$$+ 2P(1-T)^2(1-P)Cov(\hat{\theta}_{10}, \hat{\theta}_{00}) + 2(1-P)^2T(1-T)Cov(\hat{\theta}_{01}, \hat{\theta}_{00})\Big]$$

$$= \frac{1}{n\{1-2P(1-P)-2T(1-T)\}^2}\Big[P^2T^2\theta_{11}(1-\theta_{11}) + P^2(1-T)^2\theta_{10}(1-\theta_{10}) + (1-P)^2T^2\theta_{01}(1-\theta_{01})$$

$$+ (1-P)^2(1-T)^2\theta_{00}(1-\theta_{00}) - 2P^2T(1-T)\theta_{11}\theta_{10} - 2PT^2(1-P)\theta_{11}\theta_{01}$$
$$-2PT(1-P)(1-T)\theta_{11}\theta_{00} - 2P(1-T)(1-P)T\theta_{10}\theta_{01}$$
$$- 2P(1-T)^2(1-P)\theta_{10}\theta_{00} - 2(1-P)^2T(1-T)\theta_{01}\theta_{00}\Big]$$

which on simplification reduces to (3.3). Hence the theorem.

Theorem 3.3. An unbiased estimator of the variance of the unbiased estimator $\hat{\pi}_p$ of population proportion π is given by

$$\hat{v}(\hat{\pi}_p) = \frac{\hat{\pi}_p(1-\hat{\pi}_p)}{n-1} + \frac{(2P-1)^2(2T-1)^2\{P(1-P)+T(1-T)\}\hat{\pi}_p}{n\{1-2P(1-P)-2T(1-T)\}^2} + \frac{PT(1-P)(1-T)\{1-16PT(1-P)(1-T)\}}{n\{1-2P(1-P)-2T(1-T)\}^2} \tag{3.4}$$

Proof. It is easy to show that

$$E\left[\hat{v}(\hat{\pi}_p)\right] = V(\hat{\pi}_p)$$

which proves the theorem.

In the next section, we suggest a new compromised optimal estimator of the population proportion π of the sensitive attribute in a population.

4. Proposed Compromised Optimal Estimator

We consider a compromised estimator of the population proportion π of a sensitive attribute in the population by combining the above two estimators as:

$$\hat{\pi}_c = \alpha\hat{\pi}_{os} + (1-\alpha)\hat{\pi}_p \tag{4.1}$$

where α is a suitable chosen constant such that the variance of the compromised estimator $\hat{\pi}_c$ is minimum. Now we have the following theorem:

Theorem 4.1. The minimum variance of the estimator $\hat{\pi}_c$ in (4.1) is given by:

$$\text{Mn.V}(\hat{\pi}_c) = V(\hat{\pi}_p) - \frac{\{Cov(\hat{\pi}_{os},\hat{\pi}_p) - V(\hat{\pi}_p)\}^2}{V(\hat{\pi}_{os}) + V(\hat{\pi}_p) - 2Cov(\hat{\pi}_{os},\hat{\pi}_p)} \tag{4.2}$$

Proof. By the definition of variance, we have

$$\begin{aligned} V(\hat{\pi}_c) &= V[\alpha\hat{\pi}_{os} + (1-\alpha)\hat{\pi}_p] \\ &= \alpha^2 V(\hat{\pi}_{os}) + (1-\alpha)^2 V(\hat{\pi}_p) + 2\alpha(1-\alpha)Cov(\hat{\pi}_{os},\hat{\pi}_p) \\ &= V(\hat{\pi}_p) + \alpha^2\left[V(\hat{\pi}_{os}) + V(\hat{\pi}_p) - 2Cov(\hat{\pi}_{os},\hat{\pi}_p)\right] + 2\alpha[Cov(\hat{\pi}_{os},\hat{\pi}_p) - V(\hat{\pi}_p)] \end{aligned} \tag{4.3}$$

On setting

$$\frac{\partial V(\hat{\pi}_c)}{\partial \alpha} = 0$$

We get the optimum value of α given by

$$\alpha = -\frac{Cov(\hat{\pi}_{os},\hat{\pi}_p) - V(\hat{\pi}_p)}{V(\hat{\pi}_{os}) + V(\hat{\pi}_p) - 2Cov(\hat{\pi}_{os},\hat{\pi}_p)} \tag{4.4}$$

On substituting (4.4) in (4.3), we have the theorem.

Now we have the following corollaries:

Corollary 4.1. The covariance between the estimators $\hat{\pi}_{os}$ and $\hat{\pi}_p$ is given by:

$$Cov\left(\hat{\pi}_{os},\hat{\pi}_p\right) = \frac{1}{2n(1-2P(1-P)-2T(1-T))\{(P+T-1)^2+(P-T)^2\}}\left[PT(P+T-1)\{\theta_{11}(1-\theta_{11})\right.$$

$$\left. -P(1-T)\theta_{10}\theta_{11} - (1-P)T\theta_{01}\theta_{11} - (1-P)(1-T)\theta_{00}\theta_{11}\right\}$$

$$+(P-T)\{P(1-T)\theta_{10}(1-\theta_{10})-PT\theta_{11}\theta_{10}-(1-P)T\theta_{10}\theta_{01}-(1-P)(1-T)\theta_{10}\theta_{00}\}$$

$$+(T-P)\{(1-P)T\theta_{01}(1-\theta_{01})-PT\theta_{11}\theta_{01}-P(1-T)\theta_{10}\theta_{01}-(1-P)(1-T)\theta_{01}\theta_{00}\}$$

$$+(1-P-T)\{(1-P)(1-T)\theta_{00}(1-\theta_{00})-PT\theta_{11}\theta_{00}-P(1-T)\theta_{10}\theta_{00}-(1-P)T\theta_{01}\theta_{00}\}\Big] \tag{4.5}$$

Corollary 4.2. An estimator of the optimum value of α given by

$$\hat{\alpha}=-\frac{\text{cov}(\hat{\pi}_{os},\hat{\pi}_p)-v(\hat{\pi}_p)}{v(\hat{\pi}_{os})+v(\hat{\pi}_p)-2\text{cov}(\hat{\pi}_{os},\hat{\pi}_p)} \tag{4.6}$$

where $\text{cov}(\hat{\pi}_{os},\hat{\pi}_p)$ can be obtained from (4.5) by replacing θ_{ij} by $\hat{\theta}_{ij}$. A practicable compromised biased estimator of the population proportion π of a sensitive attribute in the population by combining the above two estimators as:

$$\hat{\pi}_c^{*}=\hat{\alpha}\hat{\pi}_{os}+(1-\hat{\alpha})\hat{\pi}_p \tag{4.7}$$

It is easy to show that $V(\hat{\pi}_c^{*})\approx V(\hat{\pi}_c)$ to the first order of approximation.

Now in the next section, we study the percent relative efficiency of the proposed estimators over the Odumade and Singh (2009) estimator at equal protection of the respondents.

5.Relative Efficiency

The percent relative efficiency of the proposed new estimator $\hat{\pi}_p$ over the Odumade and Singh (2009) estimator is given by:

$$\text{RE}(1)=\frac{V(\hat{\pi}_{os})}{V(\hat{\pi}_p)}\times 100\% \tag{5.1}$$

The percent relative efficiency of the proposed compromised estimator $\hat{\pi}_c$ over the Odumade and Singh (2009) estimator is given by:

$$\text{RE}(2)=\frac{V(\hat{\pi}_{os})}{V(\hat{\pi}_c)}\times 100\% \tag{5.2}$$

The percent relative efficiency of the proposed compromised estimator $\hat{\pi}_c$ over the Singh and Sedory (2011, 2012) lower bound of variance is given by:

$$\text{RE}(3)=\frac{V(\hat{\pi}_{mle})}{V(\hat{\pi}_c)}\times 100\% \tag{5.2}$$

The percent relative efficiency values RE(1), RE(2) and RE(3) are presented in Table 5.1 for different values of π in the range 0.05 to 0.5 with a step of 0.05 for $P = 0.7$ and $T = 0.7$ for all estimators considered here.

Table 5.1. Percent RE values

π	RE(1)	RE(2)	RE(3)
0.05	122.3	126.3	100.0
0.10	116.8	118.4	100.0
0.15	112.3	112.8	100.0
0.20	108.7	108.7	100.0
0.25	105.5	105.7	100.0
0.30	102.7	103.5	100.0
0.35	100.2	101.9	100.0
0.40	97.9	100.8	100.0
0.50	93.7	100.0	100.0

The RE(1) values show that the proposed estimator $\hat{\pi}_p$ remains more efficient that the Odumade and Singh (2009) estimator when the value of π is close to zero with maximum relative efficiency of 122.6% at equal protection of respondents. The compromised optimal estimator $\hat{\pi}_c$ show relate efficiency of 126.3% for value of $\pi = 0.05$ and remains efficient than the Odumade and Singh (2009) model until the value of π approaches 0.5. The value of RE(3)=100% indicates that the optimal estimator $\hat{\pi}_c$ attains the lower bound of variance developed by Singh and Sedory (2011, 2012). The benefit of the optimal estimator $\hat{\pi}_c$ is that it provides a closed form of an estimator and can be used to construct confidence interval estimates, if required. The SAS code used in the analysis are given in the Appendix-A.

References

Abdelfatah, S. and Mazloum, R. (2015a). Efficient Estimation in a Two-Stage Randomized Response Model. *Mathematical Population Studies: An International Journal of Mathematical Demography,* 22(4), 234-251.

Abdelfatah, S. and Mazloum, R. (2015b).An improved stratified randomized response model using two decks of cards. *Model Assisted Statistics and Applications*, 10, 309-320.

Abdelfatahm S., Mazloum, R. and Singh, S. (2013). Efficient use of a two-stage randomized response procedure. *Braz. J. Probab. Stat.*, 27(4), 608-617.

Arnab, R. and Shangodoying, D.K. (2016). Randomized Response Techniques Using Maximum Likelihood Estimator. *Communications in Statistics - Theory and Methods*, 44: 3340–3349, 2015.

Arnab, R. and Thuto, M. (2015). Randomized response techniques: A case study of the risky behaviors' of students of a certain University. *Model Assisted Statistics and Applications*, vol. 10, no. 4, pp. 421-430, 2015

Arnab, R., Singh, S. and North, D. (2012). Use of Two Decks of Cards in Randomized Response Techniques for Complex Survey Designs. *Communications in Statistics - Theory and Methods*, 41(16-17), 3198-3210.

Bacanli, S. and Tuncel, T. (2014). A Post-Stratified Randomized Response Model for Proportion. *American Journal of Mathematics and Statistics*, 4(3): 156-161

Batool, F. and Shabbir, J. (2016). A Two Stage Design for Multivariate Estimation of Proportions. *Communications in Statistics - Theory and Methods*, DOI: 10.1080/ 03610926. 2014.942435

Batool, F., Shabbir, J. and Hussain, Z. (2015). An Improved Binary Randomized Response Model Using Six Decks of Cards. *Communications in Statistics: Theory and Methods*, DOI: 10.1080/03610918.2015.1053922

Chaudhuri, A. (2015). Special Issue: Warner's Randomized Response Model. *Model Assisted Statistics and Applications*, 10(4), pp. 277-457.

Chen, C.C. and Singh, S. (2011). Pseudo-Bayes and pseudo-empirical Bayes estimators in randomized response sampling. *Journal of Statistical Computation and Simulation*, 81(6), 779-793

Fox, J.A. (2016). *Randomized Response and Related Methods: Surveying Sensitive Data*. Second Edition, SAGE.

Fox, J.A. and Tracy, P.E.(1986). *Randomized Response: A method of Sensitive Surveys*. Newbury Park, CA: SEGE Publications.

Gjestvang, C.R. Singh, S. (2006). A new randomized response model. *J. Roy. Statist. Soc.*, B, 68, 523-530.

Lee, C.S., Sedory, S.A. and Singh, S. (2013a). Simulated Minimum Sample Size Requirements in Various Randomized Response Models. *Communications in Statistics - Simulation and Computation*, 42(4), 731-789.

Lee, C.S., Sedory, S.A. and Singh, S. (2013b). Estimating at least seven measures of qualitative variables from a single sample using randomized response technique. *Statistics and Probability Letters*, 83(1), 399-409.

Lee, Cheon-Sig, Su, Shu-Ching, Mondragon, Katrina, Zamora, Monique L., Sedory, S.A. and Singh, S. (2016). Comparison of Cramer-Rao lower bounds of variances for at least equal protection of respondents. *Statistica Neerlandica* 70 (2), 80-99.

Lee, G.S., Hong, K.H., Kim, J.M., and Son, C.K. (2014). Estimation of the proportion of a sensitive attribute based on a two-stage randomized response model with stratified unequal probability sampling. *Braz. J. Probab. Stat.*, 28 (3), 381-408.

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 question indirectly. Biometrika 77, 436-438.

Mangat, N.S., Singh, R., (1990). An alternative randomized response procedure. *Biometrika*, 77, 349-442.

Mangat, N.S. (1994). An improved Randomized response strategy. *J. Roy. Statist. Soc.* B, 56, 93-95.

Odumade, O. and Singh, S. (2009). Efficient use of two decks of cards in randomized response sampling. *Communications in Statistics-Theory and Methods*, 38, 439-446.

Singh, S. and Sedory, S.A. (2011). Cramer-Rao Lower Bound of Variance in Randomized Response. Sampling. *Sociological Methods & Research*, 40(3) 536–546.

Singh, S. and Sedory, S.A. (2012). A true simulation study of three estimators at equal protection of respondents in randomized response sampling. *Statistica Neerlandica*, 66(4), 442-451.

Su, S.C., Sedory, S.A., and Singh, S. (2015). Kuk's Model Adjusted for Protection and Efficiency. *Sociological Methods and Research*, Vol. 44(3) 534-551.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*, 60, 63-69

Appendix-A

```
DATA DATA1;
INPUT PI;
CARDS;
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.4
0.5
;
DATA DATA2;
SET DATA1;
P = 0.7;
T = 0.7;
A=1;
B=1;
C=1;
D=1;
TH11 = (P+T-1)*PI + (1-P)*(1-T);
TH10 = (P-T)*PI + T*(1-P);
TH01 = (T-P)*PI + P*(1-T);
TH00 = (1-P-T)*PI + P*T;
TERM1 = (P+T-1)**2*(P*T+(1-P)*(1-T)) + (P-T)**2*(T*(1-P)+P*(1-T));
DENO = (P+T-1)**2 + (P-T)**2;
VAROS = TERM1/(4*DENO**2)-(2*PI-1)**2/4;
EZ = A*((1-P)*(1-T))*TH00 + B*(P*(1-T))*TH10 +C* ((1-P)*T)*TH01 + D*(P*T)*TH11;
EZ2 = A**2*((1-P)*(1-T))**2*TH00 +B**2*(P*(1-T))**2*TH10 +C**2*((1-P)*T)**2*TH01 +
D**2*(P*T)**2*TH11;
```

```
VARZ = EZ2-EZ**2;
DENOE = A*((1-P)*(1-T))*(1-P-T) + B*(P*(1-T))*(P-T) + C*(1-P)*T*(T-P) + D*(P*T)*(P+T-1);
VARP = VARZ/DENOE**2;
RE = VAROS*100/VARP;
T1 = (P*T)**2*TH11*(1-TH11);
T2 = (P*(1-T))**2*TH10*(1-TH10);
T3 = ((1-P)*T)**2*TH01*(1-TH01);
T4 = ((1-P)*(1-T))**2*TH00*(1-TH00);
T5 = -2*P**2*T*(1-T)*TH11*TH10;
T6 = -2*P*T**2*(1-P)*TH11*TH01;
T7 = -2*P*T*(1-P)*(1-T)*TH11*TH00;
T8 = -2*P*(1-P)*T*(1-T)*TH10*TH01;
T9 = -2*P*(1-P)*(1-T)**2*TH10*TH00;
T10 = -2*(1-P)**2*T*(1-T)*TH01*TH00;
DENON = 1-2*P*(1-P)-2*T*(1-T);
VARNEW = (T1+T2+T3+T4+T5+T6+T7+T8+T9+T10)/DENON**2;
RE2 = VAROS*100/VARNEW;
LB = (P+T-1)**2/TH11+(P-T)**2/TH10 + (T-P)**2/TH01 + (1-P-T)**2/TH00;
LBOS = 1/LB;
RE3 = VAROS*100/LBOS;
G1 = PI*(1-PI);
G2 = (2*P-1)**2*(2*T-1)**2*(P*(1-P)+T*(1-T))*PI/DENOE**2;
G3 = P*T*(1-P)*(1-T)*(1-16*P*T*(1-P)*(1-T))/DENOE**2;
VCHECK = G1 + G2 + G3;
H1 = (P+T-1)*(P*T*TH11*(1-TH11)-P*(1-T)*TH10*TH11-(1-P)*T*TH01*TH11-(1-P)*(1-
T)*TH00*TH11);
H2 =(P-T)*(P*(1-T)*TH10*(1-TH10)-P*T*TH11*TH10-(1-P)*T*TH10*TH01-(1-P)*(1-
T)*TH10*TH00);
H3 =(T-P)*((1-P)*T*TH01*(1-TH01)-P*T*TH11*TH01-P*(1-T)*TH10*TH01-(1-P)*(1-
T)*TH01*TH00);
H4 = (1-P-T)*((1-P)*(1-T)*TH00*(1-TH00)-P*T*TH11*TH00-P*(1-T)*TH10*TH00-(1-
P)*T*TH01*TH00);
H5 = 2.0*(1-2*P*(1-P)-2*T*(1-T))*((P+T-1)**2+(P-T)**2);
COV1 =(H1+H2+H3+H4)/H5;
ALPHAO = (COV1-VARP)/(VAROS+VARP-2*COV1);
DENO1 = VAROS + VARP-2.0*COV1;
VARC = VARP - (COV1-VARP)**2/DENO1;
RE1 = VAROS*100/VARP;
RE2 = VAROS*100/VARC;
RE3 = LBOS*100/VARC;
KEEP PI P T RE1 RE2 RE3;
PROC PRINT DATA= DATA2;
VAR PI P T RE1 RE2 RE3;
DATA DATA3;
SET DATA2;
PROC EXPORT DATA=DATA3  OUTFILE='c:\SASDATAFILES\OUT11.xls' DBMS=EXCEL
REPLACE;
RUN;
```