RASHI 2 (2) : 07-18 (2017)

Privacy protection in estimating sensitive population proportion by hypergeometric randomized response model

Kajal Dihidar

Sampling and Official Statistics Unit, Indian Statistical Institute, Kolkata.

ABSTRACT

In a practical sample survey containing sensitsive questions such as the illegal use of drugs, illegal earning, or incidence of acts of domestic violence, etc., the respondents may prefer not to confide the correct answers to the interviewer. In such cases the individuals may elect not to reply at all or to reply with incorrect answers. The resulting evasive answer-bias is ordinarily dicult to assess. To overcome this difficulty, Warner (1965) has introduced the pioneering randomized response technique for estimating the proportion of individuals possessing those sensitive attributes which can potentially eliminate the bias. In this paper we consider the problem of estimating sensitive population proportion by hypergeometric randomized response model. While implementing the randomized response technique, an important aspect is to take care of the respondents' privacy regarding the sensitive attribute. Here we investigate the degree of privacy protection offered to the interviewees in case of using hypergeometric randomized response model. Based on the pioneering work of Leysieffer and Warner (1976), we derive the jeopardy measures for our proposed model. We present a numerical illustration on how to choose the device parameters ensuring the privacy protection within some desired limits as well as maintaining the efficiency in estimation.

INTRODUCTION

Collection of data in surveys on sensitive issues, such as, tax evasion, drug use, illegal abortion, etc. poses a very difficult task due to non-cooperation of the respondents, and even if they agree to participate, the truthful answers may not be obtained. To overcome this difficulty, Warner (1965) pioneered the Randomized Response (RR) technique for estimating the proportion of people bearing a stigmatizing attribute, say A in a community, based on a sample of respondents drawn by Simple Random Sampling With Replacement (SRSWR). In his method, each respondent is provided with a randomization device by which he chooses one of two questions 'Do you belong to A ?' or 'Do you belong to A^c ?' with respect to probabilities, say, $p : (1 - p)$, where $p \neq 1/2$. The selected respondent is asked to draw randomly one card from the box and is asked to report the 'match' or 'non-match' of his own characteristic with the question written on the card drawn by him. These RR's gathered from a sample of persons provide an unbiased estimator for the sensitive population proportion, say, θ_A. Based on these RRs the variance of this estimator and an unbiased estimator for that variance are also given by Warner (1965).

Later significant developments to Warner's model are made by many researchers. For example, to expect the greater participation rate of the respondents, Horvitz et al. (1967), Greenberg et al. (1969) developed the unrelated question model, where in place of both questions being about sensitive charateristic, one question is about sensitive, and the other is completely unrelated to the sensitive characteristic, e.g. 'Do you prefer football to cricket?' or 'Is red your favourite colour?'. Boruch (1971) introduced the forced response model where the randomization determines whether a respondent truthfully answers the sensitive question or simply replies with a forced answer, 'yes' or 'no'. The idea behind the forced response design is that a certain proportion of respondents are expected to respond 'yes' or 'no' regardless of their truthful response to the sensitive question, and the design protects the anonymity of respondents' answers. That is, interviewers and researchers can never tell whether observed responses are in reply to the sensitive question. Kuk (1990) proposed a method, where each person selected by simple random sampling with replacement (SRSWR) is given two boxes, say, Box-1 and Box-2. Each of the two boxes are filled with cards of two types, say, red and blue with their mixing proportions being $p_1 : (1 - p_1)$, $0 < p_1 < 1$ in one box and $p_2 : (1 - p_2)$, $0 < p_2 < 1$ in the other; $p_1 \neq p_2$ and $p_1 + p_2 \neq 1$. Every selected person is requested to draw cards for a fixed number of times, say, K times independently, either from the first box or from the second, according as whether this person bears characteristic A or not. The respondent is requested to report the number of red cards obtained out of K cards drawn. Based on these RRs an unbiased estimator for θ, variance and variance estimator are obtained.

Likewise, many contributors of this area have enriched the randomized response literature, for instance, Moors (1971), Raghavarao (1978), Eichhorn and Hayre (1983), Chaudhuri and Mukerjee (1987), Mangat and Singh (1990), Mangat (1994), Huang (2004), Kim and Warde (2004), Gjestvang and Singh (2009), Chaudhuri, Bose and Dihidar

E-mail: kajaldihidar@gmail.com

(2011a, 2011b), Dihidar and Chowdhury (2013), Singh and Grewal (2013), Singh and Sedory (2013), Chaudhuri and Dihidar (2014), Dihidar (2016) among others. We refer to Hedayat and Sinha (1991) as an example of an early text book on sampling which covers this area as a separate chapter (see Chapter 11). For a comprehensive review of the literature on these techniques, we refer to the books by Chaudhuri and Mukerjee (1988) and Chaudhuri (2011) and the various articles in Chaudhuri et al. (2016).

An important aspect of collecting data on sensitive variables is that the survey sampling practitioners need to take care of the respondent's privacy to reduce biases due to refusals to respond and intentionally misleading replies. Lanke (1976) studied the issue of respondent's privacy protection and the same issue was studied by Leysiefer and Warner (1976) for dichotomous populations, and by Loynes (1976) for polychotomus populations. Anderson (1977) studied the efficiency versus protection in a general randomized response model.

Later, Ljungqvist (1993) gave a unified approach to measures of privacy for dichotomous populations, and Nayak and Adeshiyan (2009), Chaudhuri, Christodes and Saha (2009) proposed measures of jeopardy. Among the researchers of this area, Giordano and Perri (2012) has compared the efficiencies of unrelated question model at same privacy protection degree while Dihidar and Basu (2017) has studied the privacy protection issue for a modied unrelated question model. For randomized response models suitable for discrete valued sensitive variables, Bose (2015) has investigated in detail the privacy protection and efficiency in estimation. For many other recent rich developments in this direction, we refer to Chaudhuri et al. (2016).

Motivated by these earlier researchers, in this paper, we make an attempt to investigate the matter namely, to what extent the respondents' privacy will be protected while using the hypergeometric randomized response model. We present some numerical illustrative design parameters ensuring the privacy protection at some desired level at the same time maintaining high efficiency in estimation. We organize our findings of this research work in the following sections.

2 Generating RR by Hypergeometric Distribution

Let $U = (1, ..., i, ..., N)$ denote a finite population of N persons labeled 1 through N. Let

$y_i = 1$ if i^{th} person bears the sensitive characteristic A

$= 0$, otherwise.

Our objective is to estimate the population proportion $\theta = \frac{1}{N}\sum_{i=1}^{N} y_i$ bearing the sensitive characteristic A, using randomized response technique (RRT).

For generating the hypergeometric randomized responses we proceed in the following way. We prepare two ramdomized response boxes, say, Box1 and Box2, where each of the two boxes are filled with cards of two types, say red and blue; suppose Box 1 contains total N_1 number of cards, of which r_1 cards are red and the rest $N_1 - r_1$ cards are blue; and Box 2 contains total N_2 number of cards, of which r_2 cards are red and the rest $N_2 - r_2$ cards are blue; and $r_1/N_1 \neq r_2/N_2$. We consider the simple random sampling with replacement (SRSWR) scheme for selection of respondents, this scheme being popularly used in most studies on randomized responses. Each respondent in sample s of units collected by SRSWR is given two boxes and requested to draw cards K times without replacement, either from the first box or from the second, according as whether this person bears the sensitive characteristic A or not, and is requested to give the randomized response as the number of red cards out of the K cards drawn. The collected randomized responses from n selected respondents will be used to estimate θ.

Let us denote E_p, V_p as the expectation and variance operators for sampling design p, being SRSWR here, and E_R, V_R as the conditional expectation and variance operators for randomized response collection stage given a sample unit is chosen. Then the overall expectation, variance operators denoted by E and V are given as $E = E_p E_R$ and $V = E_p V_R + V_p E_R$. So, if y_i denotes the y-value for a person chosen on the ith draw for (i = 1,..., n) and if f_i denotes the number of red cards happened to be obtained out of the K trials as reported by that person, then following the approach of Chaudhuri (2001) we can have

$$E_R(f_i) = K\left[y_i\frac{r_1}{N_1} + (1-y_i)\frac{r_2}{N_2}\right],$$

and

$$V_R(f_i) = K\left[y_i\frac{r_1}{N_1}\frac{N_1-r_1}{N_1}\frac{N_1-K}{N_1-1} + (1-y_i)\frac{r_2}{N_2}\frac{N_2-r_2}{N_2}\frac{N_2-K}{N_2-1}\right].$$

This leads to

$$E_R(f_i) = K\left[y_i\left(\frac{r_1}{N_1} - \frac{r_2}{N_2}\right) + \frac{r_2}{N_2}\right],$$

and

$$E_R\left[\frac{\frac{f_i}{K} - \frac{r_2}{N_2}}{\frac{r_1}{N_1} - \frac{r_2}{N_2}}\right] = y_i \quad \text{provided } \frac{r_1}{N_1} \neq \frac{r_2}{N_2}.$$

So if we call

$$z_i = \frac{\frac{f_i}{K} - \frac{r_2}{N_2}}{\frac{r_1}{N_1} - \frac{r_2}{N_2}}, \text{ then we have } E_R(z_i) = y_i.$$

And

$$V_R(z_i) = \frac{1}{\left(\frac{r_1}{N_1} - \frac{r_2}{N_2}\right)^2}\left[\frac{1}{K^2}V_R(f_i)\right]$$

$$= \frac{1}{\left(\frac{r_1}{N_1} - \frac{r_2}{N_2}\right)^2}\frac{1}{K^2}K\left[y_i\frac{r_1}{N_1}\frac{N_1-r_1}{N_1}\frac{N_1-K}{N_1-1} + (1-y_i)\frac{r_2}{N_2}\frac{N_2-r_2}{N_2}\frac{N_2-K}{N_2-1}\right]$$

$$= \frac{1}{K\left(\frac{r_1}{N_1} - \frac{r_2}{N_2}\right)^2}\left[y_i\left\{\frac{r_1}{N_1}\frac{N_1-r_1}{N_1}\frac{N_1-K}{N_1-1} - \frac{r_2}{N_2}\frac{N_2-r_2}{N_2}\frac{N_2-K}{N_2-1}\right\} + \frac{r_2}{N_2}\frac{N_2-r_2}{N_2}\frac{N_2-K}{N_2-1}\right]$$

$$= ay_i + b, \text{ say, where}$$

$$a = \frac{1}{K\left(\frac{r_1}{N_1} - \frac{r_2}{N_2}\right)^2}\left[\frac{r_1}{N_1}\frac{N_1-r_1}{N_1}\frac{N_1-K}{N_1-1} - \frac{r_2}{N_2}\frac{N_2-r_2}{N_2}\frac{N_2-K}{N_2-1}\right],$$

$$b = \frac{1}{K\left(\frac{r_1}{N_1} - \frac{r_2}{N_2}\right)^2}\left[\frac{r_2}{N_2}\frac{N_2-r_2}{N_2}\frac{N_2-K}{N_2-1}\right].$$

Clearly an unbiased estimator for $V_i = V_R(z_i)$ can be considered as

$$\widehat{V}_R(z_i) = v_i = az_i + b, \quad i \in s, \quad \text{as} \quad E_R(v_i) = V_R(z_i).$$

Now an unbiased estimator for $\theta = \frac{\sum_{i=1}^{N} y_i}{N} = \bar{Y}$ is given by

$$\hat{\theta} = \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad \text{as } E_p E_R(\bar{z}) = E_p(\bar{y}) = \bar{Y} = \theta.$$

And

$$V(\bar{z}) = V_p E_R(\bar{z}) + E_p V_R(\bar{z}) = V_p(\bar{y}) + E_p(\frac{1}{n^2}\sum_{i=1}^{n} V_R(z_i)) = \frac{1}{n}[\theta(1-\theta)] + \frac{1}{Nn}\sum_{i=1}^{N} V_i,$$

where $V_i = V_R(z_i)$. $V(\bar{z})$ can be unbiasedly estimated by

$$\hat{V}(\bar{z}) = v(\bar{z}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(z_i - \bar{z})^2,$$

because

$$\frac{1}{n(n-1)}\left[E_R(\sum_{i=1}^{n}(z_i-\bar{z})^2)\right] = \frac{1}{n(n-1)}\left[E_R(\sum_{i=1}^{n} z_i^2 -= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n} E_R(z_i^2) - nE_R(\bar{z})^2)\right]\right.$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}(V_R(z_i) + (E_R(z_i))^2) - n(V_R(\bar{z}) + (E_R(\bar{z}))^2)\right]$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}(V_i + y_i^2) - \frac{1}{n}\sum_{i=1}^{n} V_i - n\bar{y}^2\right]$$

$$= \frac{1}{n(n-1)}\left[\frac{n-1}{n}\sum_{i=1}^{n} V_i + (n-1)s^2\right],$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$;

and

$$E_p\left[\frac{1}{n(n-1)}\left(\frac{n-1}{n}\sum_{i=1}^{n} V_i + (n-1)s^2\right)\right] = \frac{1}{Nn}\sum_{i=1}^{N} V_i + \frac{1}{n}[\theta(1-\theta)].$$

We now note that $V(\hat{\theta})$ involves V_i and increases as V_i itself increases too. So, in order to increase the efficiency of the estimator $\hat{\theta}$, i.e. to decrease the $V(\hat{\theta})$, we need to control the V_i values. In this regard, we have seen earlier that V_i depends heavily on the parameters of the randomization device and hence it can be regarded as the technical aspects of the device. Uncontrolled use of device parameters may be harmful to the respondents in view of their privacy protection. Therefore in order to control the Vi values, it is appropriate to examine the behavior of V_i values in relation to some suitable measure of protection of privacy which is device dependent. This aspect is discussed in the following section.

3 Protection of privacy for hypergeometric RR model

Under SRSWR, $P(y = 1) = \theta = Y/N = P(A)$, is the probability that a person chosen from U at random bears the sensitive attribute A. Let R be a possible randomized response obtained from a model. On applying Bayes's theorem,

$$P(A|R) = \frac{P(A)P(R|A)}{P(A)P(R|A) + P(A^c)P(R|A^c)} = \frac{\theta P(R|A)}{\theta P(R|A) + (1-\theta)P(R|A^c)},$$

and

$$P(A^c|R) = \frac{P(A^c)P(R|A^c)}{P(A)P(R|A) + P(A^c)P(R|A^c)} = \frac{(1-\theta)P(R|A^c)}{\theta P(R|A) + (1-\theta)P(R|A^c)}$$

are respectively regarded as the 'revealing probabilities' about a person's actual characteristic A or A^c in reporting R. If $P(A|R) > \theta$, R is jeopardizing with respect to A and if $P(A^c | R) > (1 - \theta)$, then R is jeopardizing with respect to A^c. Following Chaudhuri et al. (2009), a measure of jeopardy is defined combining these two as

$$J(R) = \frac{P(A|R)/\theta}{P(A^c|R)/(1-\theta)},$$

and this is the 'response-specific' 'jeopardy measure' for the randomized response obtained as R from a respondent chosen by SRSWR. This measure depends on the specific response of the participant. However, since a measure of jeopardy quantifies the risk of revealing his/her status (i.e., whether he/she belongs to the stigmatizing group) which a person undertakes by agreeing to use the randomization device, it should be made known to the participants before they agree to participate in the survey, i.e, before any response is available. It is therefore justied to use a measure which is not response-specific but rather could be regarded as a technical characteristic of the device. Chaudhuri et al. (2009) advocated to combine the values of $J(R)$ into a single index which can be used to quantify the risk of revealing one's status at the same time that will depend only on the technical characteristics of the randomization device. For a measure of 'jeopardy', they have proposed to use J, the average of the $J(R)$ values over all the possible forms of randomized responses. The closer the J is to unity, the more the privacy is protected. However, in general, the better the privacy is protected the higher the variance of the estimator $\hat{\theta}$ turns out to be, for the choice of the randomized response device specific parameters.

We now examine the behaviour in between the efficiency and the privacy protection for the hypergeometric randomized response model. Prior to that, for the sake of notational simplicity, let us call the probability $P(A|R)$ as $L(R)$, i.e. if f denotes the number of red balls obtained as randomized response, $L(f)$ is the conditional probability of bearing the stigmatizing characteristic A given that the randomized response obtained is f.

We have

$$L(f) = \frac{\theta\,[P(f|y=1)]}{\theta\,[P(f|y=1)] + (1-\theta)\,[P(f|y=0)]}$$

$$= \frac{\theta\left[\frac{\binom{r_1}{f}\binom{N_1-r_1}{K-f}}{\binom{N_1}{K}}\right]}{\theta\left[\frac{\binom{r_1}{f}\binom{N_1-r_1}{K-f}}{\binom{N_1}{K}}\right] + (1-\theta)\left[\frac{\binom{r_2}{f}\binom{N_2-r_2}{K-f}}{\binom{N_2}{K}}\right]}$$

$$= \frac{\theta\left[\frac{\binom{r_1}{f}\binom{N_1-r_1}{K-f}}{\binom{N_1}{K}}\right]}{\theta\left[\frac{\binom{r_1}{f}\binom{N_1-r_1}{K-f}}{\binom{N_1}{K}} - \frac{\binom{r_2}{f}\binom{N_2-r_2}{K-f}}{\binom{N_2}{K}}\right] + \frac{\binom{r_2}{f}\binom{N_2-r_2}{K-f}}{\binom{N_2}{K}}}.$$

As $\frac{r_1}{N_1} \longrightarrow \frac{r_2}{N_2}$, i.e. the two boxes are as alike as possible in proportion property, it can be shown by numerical illustration that $L(f) \to \theta$ and this is the desirable property for privacy to be protected, but under such situation $V_i \to \infty$, and hence the total variance value $\to \infty$, thus destroys the efficiency in estimation. The jeopardy value corresponding to a particular randomized response as defined above is obtained for hypergeometric randomized response model as

$$J(f) = \frac{L(f)/\theta}{(1-L(f))/(1-\theta)} = \frac{\frac{\binom{r_1}{f}\binom{N_1-r_1}{K-f}}{\binom{N_1}{K}}}{\frac{\binom{r_2}{f}\binom{N_2-r_2}{K-f}}{\binom{N_2}{K}}}.$$

On noting that the randomized response f values can range from 0 to K, and taking into account these all possible randomized response values, the final jeopardy measure is obtained as

$$\bar{J} = \frac{1}{K+1} \sum_{f=0}^{K} J(f).$$

The closer $\bar{J}$ is to unity, the more the privacy is protected. As in general, the better the privacy is protected the higher the variance of the estimator of θ turns out to be for the choice of the device specific parameters, in order to study how to keep a balance between the efficiency and privacy protection for the hypergeometric randomized response model, we present some numerical performance based results in Section 4.

4. Numerical Illustration

In this section we present the numerical illustration considering a hypothetical population for which θ value is assumed to be equal to 0.3. For comparison purpose, we consider various types of two devices Box 1 and Box 2 in the following way. The N_1 and N_2 values are chosen as 40, 41, 42, ..., 50, and r_1 and r_2 values are chosen as 20, 21, 22, ..., 30. And the total number of draws i.e. K is taken as 12. As K = 12, we may note that the number of red balls observed in an attempt of K draws can happen to be any value of 0, 1, 2, ..., K = 12.

For each value of f = 0, 1, 2, ..., K = 12, we compute the $L(f)$ and $J(f)$ for all possible combination of above mentioned N_1, r_1, N_2, r_2 values. As the total number of all possible combinations of N_1, r_1, N_2, r_2 values is very large, being 11^4 = 14641, for easy inspection the values of $L(f)$ and $J(f)$ for all these combinations are presented graphically. Moreover, the $J(f)$ values obtained are of very large ranges, and so the log ($J(f)$) values are plotted for clear visualization.

In Figure 1 the values of $L(f)$ for f = 0, 1, 2, ..., K = 12 are plotted against $(r_1 / N_1) - (r_2 / N_2)$. It is clear from Figure 1 that for each value of f = 0, 1, 2,...., K = 12, as the difference of $\frac{r_1}{N_1}$ and $\frac{r_2}{N_2}$ approaches to zero, the $L(f)$ values approach to θ = 0.3, as is desirable for privacy to be protected. But from Section 2, we have seen that as $\frac{r_1}{N_1} \longrightarrow \frac{r_2}{N_2}$, $V_R(r_i)$ values approach to ∞, and hence the $V(\hat{\theta})$ approaches to ∞, meaning the efficiency in estimation approaching to zero.

In Figure 2 the values of log ($J(f)$) for f = 0, 1, 2, ... , K = 12 are plotted against $(r_1 = N_1) - (r_2 / N_2)$. It is clear from Figure 2 that for every f value, as the difference of $\frac{r_1}{N_1}$ and $\frac{r_2}{N_2}$ approaches to zero, the log ($J(f)$) values approach to 0, meaning the $J(f)$ values approach to 1, as is ideal for privacy to be protected.

Next, we examine the behaviour of $\bar{J}$ over all f = 0, 1, ..., K = 12 in relation with $V(\hat{\theta})$. Deleting the combinations of N_1, r_1, N_2, r_2 values causing the zero denominator for $V(\hat{\theta})$ values, the range of $\bar{J}$ and $V(\hat{\theta})$ happen to be respectively (0.8483, 799.7576), (0.007, 3874.335), meaning that both the ranges are too much wide to see in a graphics window. So, for clear visual inspection, we make an attempt to look into the quantile values. For this purpose, following Chaudhuri (1996) and Chaouch and Goga (2010), we investigate the bivariate geometric quantile values for the joint distribution of $\bar{J}$ and $V(\hat{\theta})$ as presented in the Table 1. According to Chaudhuri (1996), the geometric quantile corresponding to a fixed direction u and based on the d-dimensional data Y_1, ... Y_N of finite population U = {1,..., k,...,N}, where N is the size of the finite population and $d \geq 2$, is defined by

$$Q(u) = arg \ \ min_{\alpha \in R^d} \sum_{k=1}^{N} \phi(u, Y_k - \alpha) \quad \text{for } u \in B^d = \{z \in R^d : ||z|| < 1\},$$

where R^d is the d-dimensional real value space and the multivariate loss function $\phi : B^d \times R^d$ is given by

$$\phi(u,t) = ||t|| + < u, t >$$

with ||.|| as the usual Euclidean norm and < .,. > as the usual Euclidean inner product. The u-th geometric quantile $Q(u)$ is indexed by a directional 'outlyingness' parameter u. The spatial median is obtained for u = 0 and $Q(0)$ is called the center of the data cloud formed by the Y_k's. On the other hand, for $u \neq 0$, Chaudhuri (1996) interprets $\|u\|$ as an 'extent of deviation' of $Q(u)$ from the center of the data cloud. As per his definition, the geometric quantile is called 'central' for $\|u\|$ close to 0 and 'extreme' for $\|u\|$ close to 1. Following this denition, in Table 1, the central quantile as well as the medium extent and extreme quantiles in both positive and negative directions are obtained.

According to the measures obtained in Table 1, to examine the behaviour of $\bar{J}$ and $V(\hat{\theta})$, for clear visual inspection, in Figure 3(a) we plot the values of $\bar{J}$ and $V(\hat{\theta})$ versus $\frac{r_1}{N_1} - \frac{r_2}{N_2}$ upto extreme quantile values, and as the jeopardy values close to 1 means the privacy is protected, for much more clear visual inspection in Figure 3(b) we

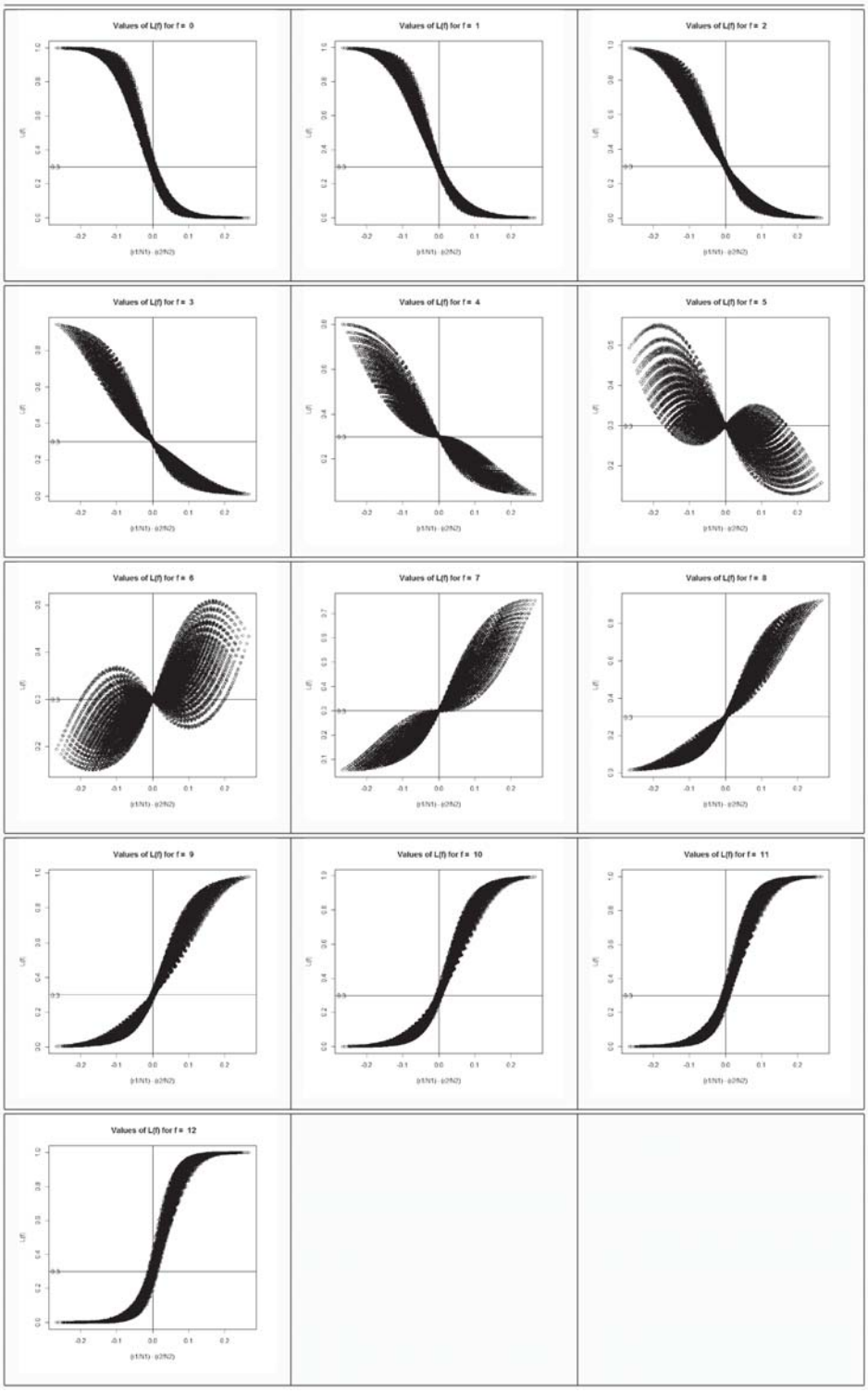

Figure 1: Values of L(f) for f = 0, 1, 2, .., 12 versus $(r_1=N_1) - (r_2/N_2)$ for = 0:3

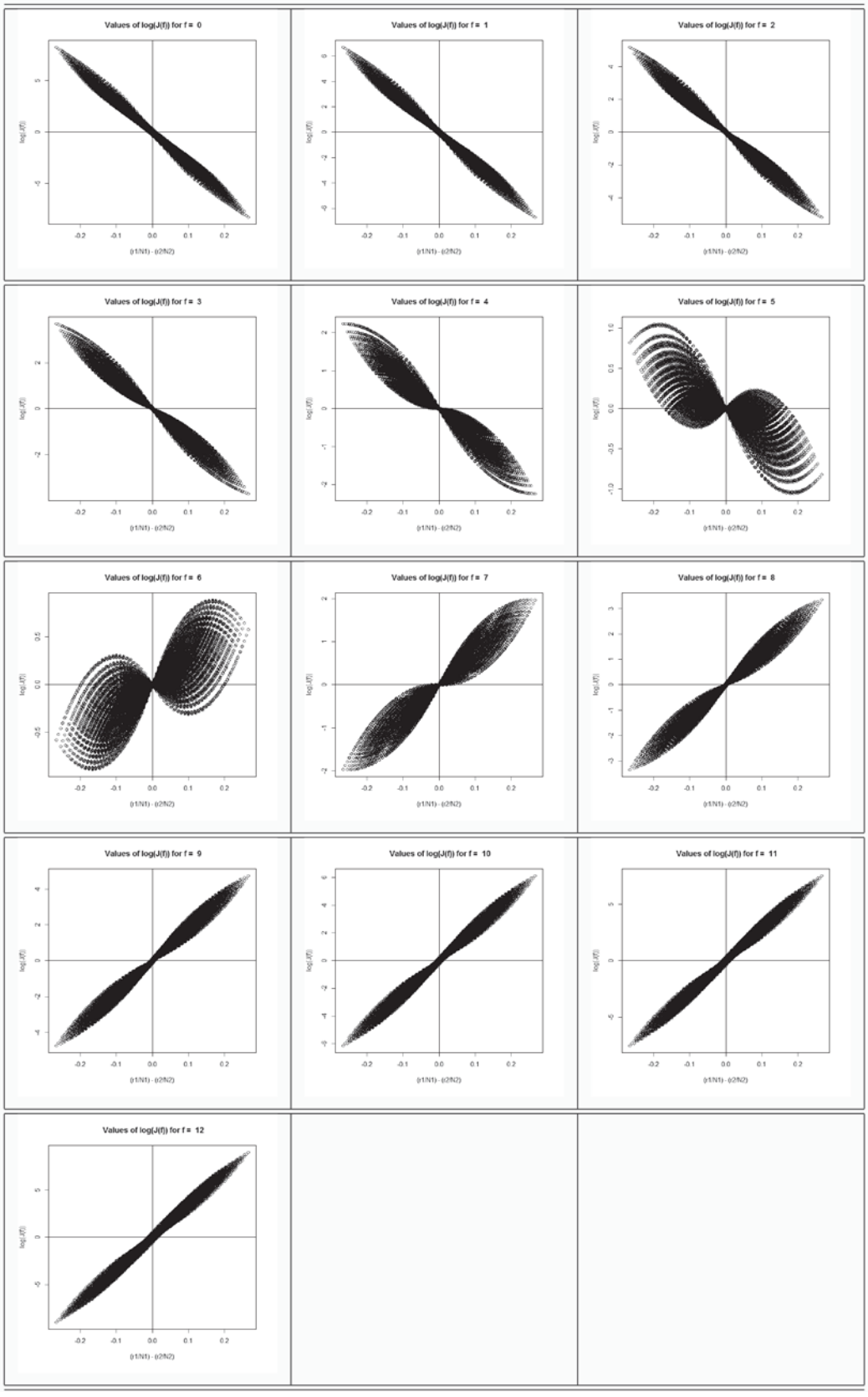

Figure 2: Values of log (J(f)) for f = 0, 1, 2, ..., 12 versus $(r_1/N_1) – (r_2/N_2)$ for = 0:3

Table 1: Bivariate geometric quantile values obtained for $\bar{J}$ and V ($\hat{\theta}$) for θ = 0:3 and n = 100.

	Extreme(-ve)	Medium(-ve)	Central(-ve)	Median	Central(+ve)	Medium(+ve)	Extreme(+ve)
	u = (–0:75, – 0:34) $\|\|u\|\|$= 0:8235	u = (–0:4, –0:3) $\|\|u\|\|$ = 0:5	u = (–0:2,– 0:2) $\|\|u\|\|$ = 0:2828	u = (0, 0) $\|\|u\|\|$ = 0:0	u = (0:2, 0:2) $\|\|u\|\|$= 0:2828	u = (0:4, 0:4) $\|\|u\|\|$= 0:5657	u = (0:65, 0:65) $\|\|u\|\|$= 0:9192
$\bar{J}$	0.9393	1.4791	1.8018	2.2865	3.2762	5.8047	37.6292
$V(\hat{\theta})$	0.0233	0.0937	0.1229	0.2999	0.8395	2.6087	31.8903

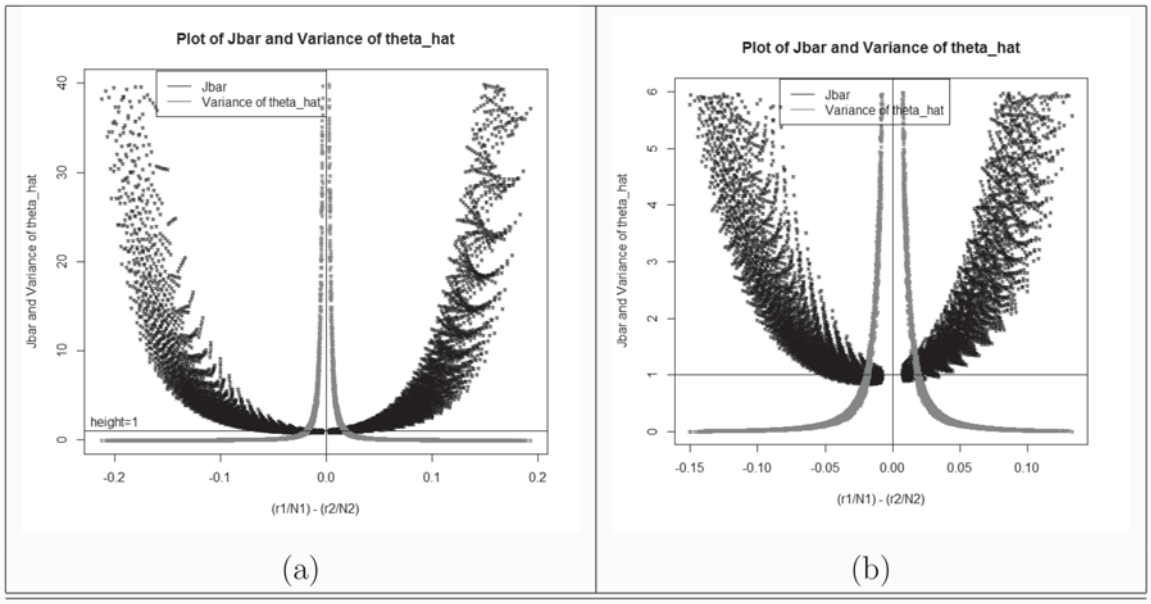

Figure 3: Jeopardy and variance values plotted against $(r_1/N_1) – (r_2/N_2)$ for θ = 0:3 and n = 100 : (a) Around Extreme(+ve), (b) Around Medium Extent(+ve) quantile

plot the values near about the medium extent quantile values. It is clear from both Figure 3(a) and Figure 3(b) that as $\frac{r_1}{N_1} \longrightarrow \frac{r_2}{N_2}$, the jeopardy values approach to 1 and $V(\hat{\theta})$ approach to high values, meaning the decrease in efficiency in estimation. Thus, the two aspects of ensuring high efficiency in estimation and guaranteeing a high degree of respondent privacy protection, are inherently coflicting. So, we have to choose our randomization device parameters in such a way that the efficiency of estimation can be maximized while maintaining a stipulated level of privacy protection. For example, examining the results as shown in Figure 3(a) and Figure 3(b), we may decide to make the randomized response devices so as to keep the jeopardy values within 0.8 and 1.2 as well as the variance of the estimator within 1.2. In such a stipulated decision, to have an idea about what may be the device parameters, we present in Table 2 some numerical observations. In this table, for some illustrative device parameters, along with the $\bar{J}$ and $V(\hat{\theta})$ values, we present the values of the efficiency of the estimator as dened by $Eff = (1/V(\hat{\theta})) * 100{:}0$ and the randomization effect as defined by $Reff = (V(\hat{\theta})/V_{Direct}(\hat{\theta}))$, where $V_{Direct}(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$ denotes the variance of the estimator of θ for direct response survey. Though the $Reff$ value more than 1 is evident in surveys with randomized response techniques, still any statistician's aim should be to keep the $V(\hat{\theta})$ as small as possible to obtain high efficiency, that means to keep the $Reff$ values as small as possible, but together with taking into account the respondents' privacy to be protected well. Hence to prepare the randomized response devices, we need to look into the three aspects, namely jeopardy, efficiency and randomization eect as illustrated below.

The illustrative results of Table 2 indicate that if we make the two devices with N_1 = 60, r_1 = 20, and N_2 = 59, r_2 = 23, we can make the respondents assured to have the almost sure privacy protection measure, being 1.047739, as well as we can keep the variance of our estimator as small as 0.134942, resulting the efficiency as 741.06% and the randomization effect as 64.26 times high as compared to direct response survey. Similary, if we make the two

devices with N_1 = 58, r_1 = 21, and N_2 = 57, r_2 = 24, we can make the respondents assured to keep the privacy protection at 1.190472, slightly departed from its ideal value 1, at the same time we can maintain the variance of our estimator as small as 0.117098, resulting the efficiency as 853.98% and the randomization effect as 55.76 times high as compared to direct response survey. However, the other device parameters yielding the smaller randomization effect can be chosen at the cost of departing the privacy protection, whatever small departure it may be, from its ideal value 1.

5 Concluding remarks

In this work, an attempt is made to examine how the hypergeometric randomized response model performs in terms of the both efficiency and protection of privacy measures in estimating the sensitive population proportion. We have developed the essential theories in this table.

Table 2 : Illustrative device parameters ensuring privacy protection and eciency for θ = 0:3 and n = 100, Eff = (1/$V(\hat{\theta})$) * 100.0 and $Reff$ = ($V(\hat{\theta})/V_{Direct}(\hat{\theta})$)

N_1	r_1	r_1/N_1	N_2	r_2	r_2/N_2	abs (r_1/N_1 - r_2/N_2)	$\bar{J}$	V ($\hat{\theta}$)	Eff	$Reff$
50	20	0.400000	57	26	0.456140	0.056140	1.191797	0.119896	834.06	57.093333
50	20	0.400000	59	27	0.457627	0.057627	1.188896	0.114331	874.65	54.443333
51	20	0.392157	58	26	0.448276	0.056119	1.165098	0.122207	818.28	58.193810
51	20	0.392157	60	27	0.450000	0.057843	1.167136	0.115539	865.51	55.018571
52	20	0.384615	59	26	0.440678	0.056063	1.139544	0.124612	802.49	59.339048
53	20	0.377358	60	26	0.433333	0.055975	1.115105	0.127103	786.76	60.525238
54	20	0.370370	54	23	0.425926	0.055556	1.170505	0.128496	778.23	61.188571
54	20	0.370370	56	24	0.428571	0.058201	1.181653	0.117648	849.99	56.022857
54	20	0.370370	58	25	0.431034	0.060664	1.193526	0.108777	919.32	51.798571
55	20	0.363636	57	24	0.421053	0.057416	1.143071	0.122849	814.01	58.499524
55	20	0.363636	59	25	0.423729	0.060092	1.158100	0.112617	887.97	53.627143
56	20	0.357143	53	22	0.415094	0.057951	1.191342	0.120488	829.96	57.375238
56	20	0.357143	58	24	0.413793	0.05665	1.108009	0.128153	780.32	61.025238
56	20	0.357143	60	25	0.416667	0.059524	1.125413	0.116522	858.21	55.486666
56	21	0.375000	58	25	0.431034	0.056034	1.163405	0.127216	786.07	60.579048
56	21	0.375000	60	26	0.433333	0.058333	1.173426	0.117866	848.42	56.126666
57	20	0.350877	54	22	0.407407	0.056530	1.141814	0.128627	777.44	61.250952
57	20	0.350877	56	23	0.410714	0.059837	1.161561	0.115326	867.11	54.917143
57	20	0.350877	58	24	0.413793	0.062916	1.182491	0.104759	954.57	49.885238
57	20	0.350877	59	24	0.40678	0.055902	1.076007	0.133559	748.73	63.599524
57	21	0.368421	59	25	0.423729	0.055308	1.128398	0.13263	753.98	63.157143
58	20	0.344828	50	20	0.400000	0.055172	1.171120	0.134693	742.43	64.139524
58	20	0.344828	52	21	0.403846	0.059019	1.191746	0.118354	844.92	56.359048
58	20	0.344828	55	22	0.400000	0.055172	1.098325	0.137083	729.49	65.277619
58	20	0.344828	57	23	0.403509	0.058681	1.119438	0.121685	821.79	57.945238
58	20	0.344828	59	24	0.406780	0.061952	1.142065	0.109602	912.39	52.191429
58	20	0.344828	60	24	0.400000	0.055172	1.046684	0.139068	719.07	66.222857
58	21	0.362069	55	23	0.418182	0.056113	1.175502	0.128809	776.34	61.337619
58	21	0.362069	57	24	0.421053	0.058984	1.190472	0.117098	853.98	55.760952
59	20	0.338983	53	21	0.396226	0.057243	1.137258	0.127713	783.00	60.815714
59	20	0.338983	55	22	0.400000	0.061017	1.161656	0.112927	885.53	53.774762
59	20	0.338983	57	23	0.403509	0.064526	1.187942	0.101418	986.02	48.294286
59	20	0.338983	58	23	0.396552	0.057569	1.081707	0.128223	779.89	61.058571
59	20	0.338983	60	24	0.400000	0.061017	1.105406	0.11455	872.98	54.547619
59	21	0.355932	58	24	0.413793	0.057861	1.147491	0.123503	809.70	58.810952
59	21	0.355932	60	25	0.416667	0.060734	1.165624	0.112523	888.71	53.582381
60	20	0.333333	51	20	0.392157	0.058824	1.174286	0.121215	824.98	57.721429
60	20	0.333333	54	21	0.388889	0.055556	1.090099	0.137561	726.95	65.505238
60	20	0.333333	56	22	0.392857	0.059524	1.114814	0.120338	830.99	57.303810
60	20	0.333333	58	23	0.396552	0.063218	1.141902	0.107108	933.63	51.003810
60	20	0.333333	59	23	0.389831	0.056497	1.047739	0.134942	741.06	64.258095
60	20	0.333333	60	24	0.400000	0.066667	1.170384	0.096676	1034.38	46.036190
60	21	0.350000	54	22	0.407407	0.057407	1.178667	0.125336	797.86	59.683810
60	21	0.350000	56	23	0.410714	0.060714	1.198867	0.112566	888.37	53.602857
60	21	0.350000	59	24	0.406780	0.056780	1.108990	0.130086	768.72	61.945714
60	22	0.366667	59	25	0.423729	0.057062	1.174379	0.125321	797.95	59.676667

regard. We have presented the situations when the privacy measure can be attained at its ideal level and how the efficiency behaves in that situation. Our numerical illustration based presenations also support the theoretical derivations. Finally, we have given some idea about how the devices can be made so as to ensure a stipulated level of privacy measures as well as to maintain the efficiency in estimation as high as possible. However, for θ value we shall have to depend on some previous guess value obtained from some reliable sources or from the findings of some pilot survey. The survey sampling practitioners aiming to estimate the sensitive population proportion with hypergeometric randomized response model in current situation, and if he can avail such approximate knowledge about the sensitive population proportion, he may use this idea to design their randomized response devices looking at the triple aspects of privacy protection, efficiency in estimation and randomization effects. Hence this is the justication of this research.

REFERENCES

Anderson, H. (1977). Efficiency versus protection in a general randomized response model. *Scandinavian Journal of Statistics*. **4**:11-19.

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *The American Sociologist*. **6**: 308-311.

Bose, M. (2015). Respondent privacy and estimation efficiency in randomized response surveys for discrete-valued sensitive variables. *Statistical Papers*. **56**: 1055-69.

Chaouch, M. and Goga, C. (2010). Design-based estimation for geometric quantiles with application to outlier detection. *Computational Statistics and Data Analysis*. **54** : 2214-29.

Chaudhuri, A. (2001). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *Journal of Statistical Planning and Inference*. **94**: 37-42.

Chaudhuri, A. (2011) *Randomized response and indirect questioning techniques in surveys*. CRC Press, Boca Raton, FL.

Chaudhuri, A., Bose, M. and Dihidar, K. (2011a). Estimation of a sensitive proportion byWarner's randomized response data through inverse sampling. *Statistical Papers*. **52**: 343-54.

Chaudhuri, A., Bose, M. and Dihidar, K. (2011b). Estimating sensitive proportions by Warner's randomized response technique using multiple randomized responses from distinct persons sampled. *Statistical Papers*. **52**: 111-24.

Chaudhuri, A., Christodes, T. C. and Rao, C.R. (Eds). (2016). Handbook of Statistics Vol. 34,

Data gathering, analysis and protection of privacy through randomized response techniques. Amsterdam : Elsevier.

Chaudhuri, A., Christodes, T. C. and Saha, A. (2009). Protection of privacy in efficient application of randomized response techniques. *Statistical Methods and Applications*. **18**: 389-418.

Chaudhuri, A. and Dihidar, K. (2014). Generating randomized response by inverse mechanism. *Model Assisted Statistics and Applications*. **9**: 343-51.

Chaudhuri, A. and Mukerjee, R. (1987) Randomized response techniques: a review. *Statistica Neerlandica*. **41**: 27-44.

Chaudhuri, A. and Mukerjee, R. (1988) Randomized responses: Theory and Techniques. Marcel Dekker, New York, NY.

Cahaudhuri, P. (1996). On a geometric notation of quantiles for multivariate data. *Journal of the American Statistical Association*. **91**: 862-72.

Dihidar, K. (2016). Estimating Sensitive Population Proportion by Generating Randomized Response Following Direct and Inverse Hypergeometric Distribution. In Chaudhuri, A., Christodes, T. C. and Rao, C.R. (Eds). Handbook of Statistics Vol. 34, pp - 427-441, Data gathering, analysis and protection of privacy through randomized response techniques. Amsterdam : Elsevier.

Dihidar, K. and Basu, L. (2017). Privacy Protection in Estimating Sensitive Population Proportion by a Modified Unrelated Question Model. *Statistics and Applications*. **15**(1 and 2): 19-25.

Dihidar, K. and Chowdhury, J. (2013). Enhancing a Randomized Response Model to Estimate Population Means to Sensitive Questions. Mathematical Population Studies: *An International Journal of Mathematical Demography*. **20(3)**: 123-36.

Eichhorn, B.H. and Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*. **7**: 307-16.

Giordano, S. and Perri, P. F. (2012). Efficiency comparison of unrelated question models based on same privacy protection degree. *Statistical Papers*. **53**: 987-99.

Gjestvang, C.R. and Singh, S. (2009). An improved randomized response model: Estimation of mean. *Journal of Applied Statistics*. **36(12)**: 1361-67.

Greenberg, B.G., Abul-Ela, Abdel-Latif, A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question RR model : theoretical framework. *Journal of American Statistical Association*. **64**: 520-39.

Hedayat, A. S. and Sinha, B. K. (1991). Design and Inference in Finite Population Sampling. New York: Wiley.

Horvitz, D.G., Shah, B.V. and Simmons, W.R. (1967). The unrelated question randomized response model. Social Statistics Section, *Proceedings of the American Statistical Association*. 65-72.

Huang, K.C. (2004). Survey technique for estimating the proportion and sensitivity in a dichotomous finite population. *Statistica Neerlandica*. **58**: 75-82.

Kim J. and Warde, W.D. (2004). A mixed randomized response model. *Journal of Statistical Planning and Inference*. **110**: 1-11.

Kuk, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*. **77**: 436-38.

Lanke, J. (1976). On the degree of protection in randomized interviews. *International Statistical Review*. **44**: 197-203.

Lee, C-S., Sedory, S.A. and Singh, S. (2013). Simulated minimum sample size requirements in various randomized response models. *Communications in Statistics : Simulation and Computation*. **42**: 771-89.

Leysieffer, R.W. and Warner, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of American Statistical Association.* **71**: 649-56.

Ljungqvist, L. (1993). A unified approach to measures of privacy protection in randomized response models: a utilitarian perspective. *Journal of American Statistical Association*. **88**: 97-103.

Loynes, R.M. (1976). Asymptotically optimal randomized response procedures. *Journal of American Statistical Association*. **71**: 924-28.

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

Mangat, N. S. and Singh, R. (1990). An alternative randomized response procedure. *Biometrika*. **77**: 439-42.

Moors, J. J. A. (1971). Optimization of the unrelated question randomised response model. *Journal of the American Statistical Association*. **66(335)**: 627-29.

Nayak, T. K. and Adeshiyan, S. A. (2009). A unified framework for analysis and comparison of randomized response surveys of binary characteristics. *Journal of Statistical Planning and Inference*. **139**: 2757-66.

Raghavarao, D. (1978). On an estimation problem in Warner's RR technique. *Biometrics*. **34**: 87-90.

Singh, S. and Grewal, I.S. (2013). Geometric distribution as a randomization device: Implemented to the Kuks model. *International Journal of Contemporary Mathematical Sciences.* **8(5)**: 243-48.

Singh, S. and Sedory, S.A. (2013). A new randomized response device for sensitive characteristics: An application of negative hypergeometric distribution. *Metron*. **71**: 3-8.

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