STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 193–210, DOI 10. 21307

ESTIMATING SENSITIVE POPULATION PROPORTION USING A COMBINATION OF BINOMIAL AND HYPERGEOMETRIC RANDOMIZED RESPONSES BY DIRECT AND INVERSE MECHANISM

Kajal Dihidar [1], **Manjima Bhattacharya** [2]

ABSTRACT

For various reasons individuals in a sample survey may prefer not to confide to the interviewer the correct answers to certain potentially sensitive questions such as the illegal use of drugs, illegal earning, or incidence of acts of domestic violence, etc. 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 difficult to assess. The use of a randomized response method for estimating the proportion of individuals possessing those sensitive attributes can potentially eliminate the bias. Following Chaudhuri and Dihidar (2014) and Dihidar (2016), here, as a possible variant, we have made an attempt to estimate the sensitive population proportion using a combination of binomial and hypergeometric randomized responses by direct and inverse mechanism. Along with the traditional simple random sampling, with and without replacement, we consider here sampling of respondents by unequal probabilities. Essential theoretical derivations for unbiased estimator, variance and variance estimators are presented for several sampling schemes. A numerical illustration is performed to make a comparative study of the relative efficiencies of the direct and inverse mechanism..

1. Introduction

Surveys for eliciting information on sensitive or stigmatizing attributes are plagued by the problem of untruthful responses or non-cooperation by respondents, both of which lead to biased estimates. To avoid this evasive answer bias and to preserve the privacy of the respondent, Warner (1965) introduced an innovative technique commonly referred to as randomized response (RR) technique. In his model, a respondent answers ‘Yes’ or ‘No’ to either the sensitive question of interest or the complementary question. For example, suppose that we are interested in whether

[1] Corresponding Author. Sampling and Official Statistics Unit, Indian Statistical Institute, Kolkata. E-mail: kajaldihidar@gmail.com

[2] Credit Swiss Company, Mumbai, Maharastra, India. E-mail: bhattacharya.manjima@gmail.com

a person belongs to the sensitive group A. The respondent uses a chance device to select Question 1:'Do you belong to Group A?,' or Question 2: 'Do you belong to Group A^c?,' where A^c is the complement of A, with probabilities, say, $p:(1-p)$, where $p \neq 0.5$. Thus, if the respondent says 'Yes', the interviewer does not know whether the 'Yes' refers to Question 1 or Question 2. These RRs gathered from a sample of persons chosen by SRSWR provide an unbiased estimator for the sensitive population proportion, say, θ. 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 a sensitive characteristic, 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), Haung (2004), Kim and Warde (2004), Gjestvang and Singh (2006), Chaudhuri, Bose and Dihidar (2011a, 2011b), Singh and Grewal (2013), Singh and Sedory (2013) 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).

In general, all the approaches of RR techniques assume that the respondents answer truthfully according to the randomized response device. However, the methods are often criticized as being susceptible to cheaters, that is, respondents who do not answer truthfully as directed by the randomizing device. Clark and Desharnais (1998) has shown that by splitting the sample into two groups and assigning each group a different randomization probability, it is possible to detect whether significant cheating is occurring and to estimate its extent while simultaneously protecting the identity of cheaters and those who may have engaged in sensitive activities. In Feth et al. (2015) different forms of cheating is described and it has been shown in detail how to obtain general solution for detecting the extent to which various forms of cheating occurs and extends these analyses with practical hints for the flexible use of these methods. However, although there may be some possibility of cheating occurrences, for the present research work we assume that the respondents are tried to be well convinced to answer truthfully according to the randomizing device and therefore, based on this assumption, below we make an attempt to develop alternative RR techniques for estimating the sensitive proportion.

As stated above, in Kuk's (1990) approach the cards are drawn from either of the two boxes with replacement. A natural question arises - what will happen if the cards are drawn without replacement? In this paper we look into this matter. We know that while drawing n cards with replacement from a box containing two types of cards, the number of the first type of cards obtained follows binomial distribution whereas the number of trials to obtain a fixed number of the first type of cards follows a negative binomial distribution, and drawing the cards without replacement instead of with replacement will result in the hypergeometric and negative hypergeometric distributions respectively for the same. Here, we consider estimating the sensitive population proportion by generating randomized responses using a combination of binomial and hypergeometric distributions in the direct approach as well by using a combination of negative binomial and negative hypergeometric distributions in the inverse approach. Also, keeping in mind that many large scale sample surveys consist of sampling of respondents by unequal probability sampling even without replacement, in this paper we develop unbiased estimators for sensitive population proportion by general sampling schemes instead of only simple random sampling with replacement scheme of respondents. We organize our findings of this research work in the following sections.

In Section 2 below, we present the necessary derivations for generating randomized responses using binomial distribution for Box-1 and hypergeometric distribution for Box-2. In Section 3, we present the same by negative binomial and negative hypergeometric distributions respectively. In Section 4, we present the unbiased estimators for θ, variance and variance estimators based on some sampling methods, namely simple random sampling (SRS) both with and without replacement (WR or WOR), and some unequal probability sampling methods, namely probability proportional to size with replacement (PPSWR), Rao, Hartley and Cochran's (1962) and Midzuno's (1952) sampling schemes. We present the numerical illustration in Section 5 for comparison purpose. Finally, we give some concluding remarks in Section 6.

2. Generating RR by direct approach

Let $U = (1,2,\ldots,N)$ denote a finite, identifiable population of N persons labeled 1 to N. Let

$$\begin{aligned} y_i &= 1, \text{ if } i^{th} \text{ person bears the sensitive character, say, } A \\ &= 0, \text{ otherwise.} \end{aligned}$$

We want to estimate the population proportion $\theta = \frac{1}{N}\sum_{i=1}^{N} y_i$, proportion of individuals bearing the sensitive character A.

In our proposed methodology, two randomized response boxes, say Box-1 and Box-2 are used, and each of the two boxes are filled with two types of cards, say 'Red' and 'Blue'; in proportion $p_1 : (1-p_1)$ in Box-1; and in proportion $p_2 : (1-p_2)$ in Box-2, $0 < p_1 \neq p_2 < 1$. Suppose Box-1 contains N_1 total number of balls out of which r_1 are red and the rest are blue, and Box-2 contains N_2 total number of balls out of which r_2 are red and the rest are blue. Hence, $p_1 = r_1/N_1$ and $p_2 = r_2/N_2$. Each respondent in sample s of units, collected with a given probability $p(s) > 0$ according to a given sampling design p, is given two boxes. Every selected person is instructed to use the first box if he bears A, otherwise to use the second box, unnoticed by the interviewer, thus protecting the privacy of the respondent. Additional instruction is also given to the selected respondent to draw cards at random independently for a specified number of times, say, K times, with replacement if he chooses the Box-1 and without replacement if he chooses the Box-2. Every selected person is requested to report finally how many times a 'Red' marked cards are actually drawn out of K trials. Let us denote f_i as the number out of K trials, a 'Red' card happened to be obtained as reported by the person labeled i. Additionally, let E_R, V_R, C_R denote the expectation, variance and covariance operators with respect to the randomized response generation. Then

$$E_R(f_i) = K[y_i p_1 + (1-y_i)p_2]$$

and

$$V_R(f_i) = K\left[y_i p_1(1-p_1) + (1-y_i)\frac{N_2-K}{N_2-1}p_2(1-p_2)\right]$$

leading to

$$E_R\left[\frac{f_i}{K}\right] = y_i(p_1-p_2)+p_2$$

$$\Rightarrow E_R\left[\frac{\frac{f_i}{K}-p_2}{p_1-p_2}\right] = y_i, \quad \text{on noting that} \quad p_1 \neq p_2.$$

Let $r_i = \dfrac{\frac{f_i}{K}-p_2}{p_1-p_2}$ with $E_R[r_i] = y_i$ and

$$V_R(r_i) = \frac{1}{(p_1-p_2)^2}V_R(\frac{f_i}{K})$$

$$= \frac{1}{(p_1-p_2)^2}\frac{1}{K^2}V_R(f_i)$$

$$= \frac{1}{K(p_1-p_2)^2}\left[y_i p_1(1-p_1) - y_i\left(\frac{N_2-K}{N_2-1}\right)p_2(1-p_2) + \left(\frac{N_2-K}{N_2-1}\right)p_2(1-p_2)\right]$$

$$= ay_i + b$$

$$= V_i, \quad \text{say,}$$

where

$$a \quad = \quad \frac{1}{K(p_1-p_2)^2}\left[p_1(1-p_1) - \left(\frac{N_2-K}{N_2-1}\right)p_2(1-p_2)\right]$$

and

$$b = \frac{p_2(1-p_2)(N_2-K)}{K(p_1-p_2)^2(N_2-1)}.$$

Then, an unbiased estimator of V_i is

$$\hat{V}_R(r_i) = v_i = ar_i + b, \;\; i \in s$$

because $E_R(v_i) = ay_i + b = V_i$.

3. Generating RR by inverse mechanism

Here, every selected respondent is given instruction to use the first box if he bears A, otherwise to use the second box. Additional instruction is given to each respondent that, if he uses Box-1, he should draw cards WR until he gets a specified number, say, t_1 'Red' cards, then he should report the number of required draws to obtain t_1 'Red' cards, say, $G = g$, where G is the random variable denoting the number of draws obtained from Box-1; similarly, if he uses Box-2, he should draw cards WOR until he gets a specified number, say, $t_2 (t_2 < r_2)$ 'Red' cards, then he should report the number of required draws to obtain t_2 'Red' cards, say, $H = h$, where H is the random variable denoting the number of draws obtained from Box-2.

Then G follows a negative binomial distribution with parameters t_1 and p_1 and its probability mass function is given by:

$$P(G = g|t_1, p_1) = \binom{g-1}{t_1-1} p_1^{t_1}(1-p_1)^{g-t_1}; g = t_1, t_1+1, \ldots$$

Similarly, the random variable H follows the negative hypergeometric distribution with parameters N_2, r_2, t_2 and its probability mass function is given by:

$$P(H = h|N_2, r_2, t_2) = \frac{\binom{r_2}{t_2-1}\binom{N_2-r_2}{h-t_2}}{\binom{N_2}{h-1}} \times \frac{r_2-t_2+1}{N_2-h+1}; h = t_2, t_2+1, \ldots, (N_2-r_2+t_2)$$

At this stage, we may note that it may be possible that the response of an individual with A could be $g < t_2$ or $g > N_2 - r_2 + t_2$, in which case it would be known that the individual has characteristic A, compromising the privacy of the respondent. So, in order to protect the privacy of the respondent, we consider $t_1 = t_2 = t$, say, and ask the respondent to stop drawing when he reaches at the number of draws at $N_2 - r_2 + t$, so that after getting the number of draws from respondent it will not be possible to find out from which box the draws are made. Hence, instead of usual negative binomial distribution, we consider the following truncated negative binomial distribution. We also note that as the number of successes is fixed at t, the number of failures is the random variable, and following Mir (2008) and Shonkwiler (2016) we utilize below the properties of the un-truncated and the truncated random variable. So, if X is the random variable denoting the number of failures preceding t successes, then the probability mass function of the usual negative binomial distribution is given by :

$$P(X = x|t, p_1) = \frac{t}{t+x}\binom{t+x}{x} p_1^t(1-p_1)^x; x = 0, 1, 2, \ldots,$$

for which the expectation and variance are

$$E(X) = t\frac{1-p_1}{p_1} \quad \text{and } V(X) = t\frac{1-p_1}{p_1^2}.$$

Following Shonkwiler (2016), we obtain the expectation and variance of the right truncated negative binomial distributed variable as

$$E(X|X \leq N_2 - r_2) = t\frac{1-p_1}{p_1} - \frac{\frac{t}{p_1}(N_2 - r_2 + 1)h(N_2 - r_2 + 1)}{tP(X \leq N_2 - r_2)} = \mu_0, \quad \text{say,}$$

where $h(N_2 - r_2 + 1)$ is the un-truncated negative binomial probability mass function $P(X = x|t, p_1)$ evaluated at $N_2 - r_2 + 1$, and

$$V(X|X \leq N_2 - r_2)$$
$$= \mu_0 + (N_2 - r_2)\left(\mu_0 - t\frac{1-p_1}{p_1}\right) + \mu_0 t\frac{1-p_1}{p_1}\left(1 + \frac{1}{t}\right) - \mu_0^2 = V_0, \quad \text{say.}$$

Hence,

$$E(G = t + X|G \leq N_2 - r_2 + t) = t + E(X|X \leq N_2 - r_2)$$
$$= t + t\frac{1-p_1}{p_1} - \frac{\frac{t}{p_1}(N_2 - r_2 + 1)h(N_2 - r_2 + 1)}{tP(X \leq N_2 - r_2)} = t + \mu_0 = \mu_1, \quad \text{say.}$$

And

$$V(G = t + X|G \leq N_2 - r_2 + t) = V(X|X \leq N_2 - r_2) = V_0.$$

So, if Z_i denotes the randomized response obtained from i^{th} chosen person, and if G_T denotes the above defined truncated negative binomial distribution, then

$$\begin{aligned} Z_i &= G_T \quad \text{if } i^{th} \text{ person bears } A \\ &= H \quad \text{if } i^{th} \text{ person bears } A^c \, . \end{aligned}$$

On noting the expectation and variance of G_T as derived above and that for the negative hypergeometric distribution $H(N_2, r_2, t)$ as

$$E(H(N_2, r_2, t)) = t\frac{N_2 + 1}{r_2 + 1}, \quad V(H(N_2, r_2, t)) = t\frac{(N_2 - r_2)(N_2 + 1)(r_2 + 1 - t)}{(r_2 + 1)^2(r_2 + 2)},$$

we have

$$\begin{aligned} E_R(Z_i) &= y_iE_R(G_T)+(1-y_i)E_R(H) \\ &= y_i\mu_1+(1-y_i)t\frac{N_2+1}{r_2+1} \\ &= y_i\left(\mu_1-\frac{t(N_2+1)}{r_2+1}\right)+t\frac{N_2+1}{r_2+1} \end{aligned}$$

This implies that if

$$\mu_1-\frac{t(N_2+1)}{(r_2+1)}\neq 0 \quad \text{and} \quad r'_i=\frac{Z_i-t\frac{N_2+1}{r_2+1}}{\mu_1-\frac{t(N_2+1)}{(r_2+1)}}$$

then

$$E_R(r'_i)=y_i.$$

We now note that

$$\begin{aligned} V'_i=V_R(r'_i) &= \frac{V_R(Z_i)}{\left[\mu_1-\frac{(N_2+1)t}{r_2+1}\right]^2} \\ &= cy_i+d, say \end{aligned}$$

where, on writing

$$\phi=\left[\mu_1-\frac{(N_2+1)t}{r_2+1}\right]^2,$$

$$c=\frac{V_0-\frac{t(N_2-r_2)(N_2+1)(r_2+1-t)}{(r_2+1)^2(r_2+2)}}{\phi},$$

$$d=\frac{t(N_2-r_2)(N_2+1)(r_2+1-t)}{(r_2+1)^2(r_2+2)\phi}.$$

An unbiased estimator for $V'_i=V_R(r'_i)$ is

$$\hat{V}_R(r'_i)=v'_i=cr'_i+d, \;\; i\in s,$$

because

$$E_R(v'_i)=E_R(cr'_i+d)=cE_R(r'_i)+d=cy_i+d=V'_i.$$

4. Comparative efficiencies of the inverse method versus direct one under different sampling schemes

We now present a study of the relative efficiencies of the direct versus inverse RRT as $e = 100\frac{V}{V'}$, where V is the variance of the usual estimator of θ for direct method and V' as the variance of the estimator of θ for the inverse method in different situations. We consider (1) Simple Random Sampling With Replacement (SRSWR) by n draws and (2) Simple Random Sampling Without Replacement (SRSWOR) in n draws, in these two cases the sample means of the transformed randomized responses are used to estimate θ. Also some unequal probability schemes, for example, (3) probability proportional to size with replacement (PPSWR), (4) Rao, Hartley and Cochran's (RHC,1962) sampling scheme and (5) Midzuno's (1952) scheme are used for the estimation of θ and variance of that estimator.

Let us denote E_p, V_p as the expectation and variance operators for design p, then the overall expectation, variance operators denoted by E and V are given as $E = E_pE_R$ and $V = E_pV_R + V_pE_R$. We present below the essential formulation for the estimator considering the direct method for generating RR (as described earlier) and variance and variance estimators for θ based on the various sampling schemes considered in this paper. For the inverse counterpart, r_i, V_i and v_i will of course change in the manners described already, replacing them by r_i', V_i' and v_i' respectively.

4.1. SRSWR in n draws

Let us denote y_k as the y-value for a person chosen on the k^{th} draw $(k = 1, \ldots, n)$ and r_k as the transformed RR generated by the direct method from that person. Then an unbiased estimator for $\theta = \frac{1}{N}\sum_{i=1}^{N} Y_i = \overline{Y}$ is given by $\overline{r} = \frac{1}{n}\sum_{k=1}^{n} r_k$ with $V(\overline{r}) = V_pE_R(\overline{r}) + E_pV_R(\overline{r}) = V_p(\overline{y}) + E_p(\frac{1}{n^2}\sum_{k=1}^{n} V_R(r_k)) = \frac{1}{n}[\theta(1-\theta)] + \frac{1}{Nn}\sum_{i=1}^{N} V_i$, where $V_i = V_R(r_i)$. $V(\overline{r})$ can be unbiasedly estimated by

$$\hat{V}(\overline{r}) = v(\overline{r}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}(r_k - \overline{r})^2.$$

4.2. SRSWOR in n draws

In this case also an unbiased estimator for θ is $\overline{r} = \frac{1}{n}\sum_{i\in s} r_i$ because $E(\overline{r}) = E_pE_R(\overline{r}) = E_p(\overline{y}) = \overline{Y} = \theta$ and $V(\overline{r}) = \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2 + \frac{1}{Nn}\sum_{i=1}^{N} V_i$. $V(\overline{r})$ is unbiasedly estimated by

$$\hat{V}(\overline{r}) = v(\overline{r}) = \frac{N-n}{Nn}\frac{1}{(n-1)}\sum_{i\in s}(r_i - \overline{r})^2 + \frac{1}{Nn}\sum_{i\in s} v_i.$$

4.3. PPSWR in n draws

Let us consider that for unequal probability sample drawing the normed size measures $p_i s$ are from an auxiliary variable z with known $z_i > 0$ for all i having $Z = \sum_{i=1}^{N} z_i$ such that $p_i = \frac{z_i}{Z}$, where $0 < p_i < 1, i = 1,2,\ldots,N$ and $\sum_{i=1}^{N} p_i = 1$. Let us denote p_k as the normed size measure, y_k as the y-value for a person chosen at the k^{th} draw $(k = 1,2,\ldots,n)$. And also let us denote r_k as the transformed RR generated by the direct method for generating randomized response for a person chosen at the k^{th} draw, for $k = 1,2,\ldots,n$. Then, following Hansen and Hurwitz (1943) an unbiased estimator for θ is given by $e_{PPSWR} = \frac{1}{Nn}\sum_{k=1}^{n}\frac{r_k}{p_k}$ with $V(e_{PPSWR}) = V_p E_R(e_{PPSWR}) + E_p V_R(e_{PPSWR}) = V_p(\frac{1}{Nn}\sum_{k=1}^{n}\frac{y_k}{p_k}) + E_p(\frac{1}{N^2n^2}\sum_{k=1}^{n}\frac{V_R(r_k)}{p_k^2}) = \frac{1}{N^2}\left[\frac{V}{n} + \frac{1}{n}\sum_{i=1}^{N}\frac{V_i}{p_i}\right]$, where

$$V = \sum_{i=1}^{N-1}\sum_{j>i}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$$

and $V_i = V_R(r_i)$. $V(e_{PPSWR})$ can be unbiasedly estimated by

$$\hat{V}(e_{PPWSR}) = v(e_{PPSWR}) = \frac{1}{N^2}\left[\frac{1}{2n^2(n-1)}\sum_{k=1}^{n}\sum_{k' \neq k, k=1}^{n}\left(\frac{r_k}{p_k} - \frac{r_{k'}}{p_{k'}}\right)^2\right].$$

4.4. Rao, Hartley and Cochran's sampling scheme of size n

Rao, Hartley and Cochran's (RHC, 1962) sampling of n persons from N population units consists of making n non-overlapping random groups of the population units of group sizes being N_i, $i = 1,\ldots,n$ such that $\sum_{i=1}^{n} N_i = N$. Let Q_i denote the sum of the normed size measures of the N_i units falling in the i^{th} group. Then, independently from every group only one unit is selected with probability proportional to the normed size measures, thus yielding a sample of required size n by RHC method. For simplicity in notation, we denote the value obtained from the unit selected from i^{th} group as y_i and its normed size measure as p_i. With this notation, the unbiased estimator for θ is

$$e_{RHC} = \frac{1}{N}\sum_{n} r_i \frac{Q_i}{p_i}.$$

Here $\sum_n$ means the sum over the n disjoint groups into which the population U is divided into random groups. Following Rao et al. (1962), the optimal choices of group sizes N_is are given by $N_i = [N/n]$ for $i = 1,2,\ldots,k$ and $N_i = [N/n] + 1$ for $i = k+1, k+2,\ldots,n$, k being determined by solving $\sum_{i=1}^{n} N_i = N$. Following

Chaudhuri and Dihidar (2014) we have

$$V(e_{RHC}) = \frac{1}{N^2}\left[C\sum_{i=1}^{N}\frac{V_i}{p_i} + (1-C)\sum_{i=1}^{N}V_i + C\left(\sum_{i=1}^{N}\frac{y_i^2}{p_i} - Y^2\right)\right], \quad \text{with } C = \frac{\sum_n N_i^2 - N}{N(N-1)}.$$

$V(e_{RHC})$ is unbiasedly estimated by

$$\hat{V}(e_{RHC}) = v(e_{RHC}) = \frac{1}{N^2}\left[D\sum_n\sum_{n'} Q_i Q_{i'}\left(\frac{r_i}{p_i} - \frac{r_{i'}}{p_i'}\right)^2 + \sum_n v_i\frac{Q_i}{p_i}\right],$$

where

$$D = \frac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2}.$$

Here $\sum_n \sum_{n'}$ denotes the sum over non-overlapping pairs of n groups.

4.5. Midzuno's (1952) sampling scheme of n persons

For our illustrative purpose we consider the fifth scheme as Midzuno's (1952) scheme of unequal probability sampling of n units. Sampling by this scheme is done first by drawing one unit by probability proportional to size measure of the auxiliary variable, say, z with $Z = \sum_{i=1}^{N} z_i$. Then, keeping the selected unit aside, the remaining $(n-1)$ units are chosen by simple random sampling without replacement (SRSWOR) out of the remaining $(N-1)$ population units. Under this scheme, the first and second order inclusion probabilities, π_i and $\pi_{ij}, i \neq j$ are as follows.

$$\pi_i = \frac{z_i}{Z} + \frac{Z - z_i}{Z}\frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} = \frac{z_i}{Z}\frac{N-n}{N-1} + \frac{n-1}{N-1} \quad \forall i = 1,2,\ldots,N, \tag{1}$$

and

$$\pi_{ij} = \frac{z_i}{Z}\frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} + \frac{z_j}{Z}\frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} + \frac{Z - z_i - z_j}{Z}\frac{\binom{N-3}{n-3}}{\binom{N-1}{n-1}}$$

$$= \frac{z_i + z_j}{Z}\frac{(N-n)(n-1)}{(N-1)(N-2)} + \frac{(n-1)(n-2)}{(N-1)(N-2)}, \quad \forall i \neq j \in U. \tag{2}$$

For this scheme, $\pi_i\pi_j > \pi_{ij}, \forall i \neq j \in U$. An unbiased estimator for the sensitive population proportion θ is given by Horvitz and Thompson(1952)'s estimator as

$$e_{HT} = \frac{1}{N}\sum_{i \in s}\frac{r_i}{\pi_i}.$$

Utilizing Yates and Grundy (1953)'s form of variance of the HT estimator the variance of e_{HT} is given by

$$V(e_{HT}) = \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1,j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{V_i}{\pi_i}\right].$$

It is unbiasedly estimated by

$$\hat{V}(e_{HT}) = v(e_{HT}) = \frac{1}{N^2}\left[\sum_{i\in s}\sum_{j\in s,j>i}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{v_i}{\pi_i}\right].$$

4.6. Comparison of the efficiencies

It is clear from the variance formulae of the unbiased estimators for θ as given in the above section, that in each case since $E_R(r_i) = y_i = E_R(r_i')$, the V_pE_R term will remain same for both the direct and inverse RRT and the difference will be only in the E_pV_R term. So, to compare the efficiencies of the two methods, we need to examine the relative magnitudes of $V_R(r_i')$ versus $V_R(r_i)$. So, the inverse method will be superior to the direct one if

$$V_R(r_i') \leq V_R(r_i), \quad \text{that is if } cy_i + d \leq ay_i + b \quad \text{or } y_i \leq \frac{b-d}{c-a}, \quad \text{provided } c - a > 0.$$

Maintaining the constraint $c - a > 0$, this condition may be equivalently stated by

$$0 \leq \theta \leq \frac{b-d}{c-a} \quad \text{or } 0 \leq \theta \leq \frac{b-d}{(b-d) + \left[\frac{V_0}{\phi} - \frac{r_1(N_1 - r_1)}{KN_1^2(\frac{r_1}{N_1} - \frac{r_2}{N_2})^2}\right]}.$$

Because of the complicated form of the above inequality, it seems excessively difficult to have any insightful idea about the superiority or otherwise of the estimators of θ realized by survey data obtained through the revised RRT approach rather than the procedure following direct method of randomized response generation. However, we present below a simulation-based numerical evaluation for efficiency comparison.

5. Numerical illustration

For numerical illustration, the values of y_i s and size measures z_i s for $i = 1, 2, \ldots, N = 117$ are taken from Chaudhuri and Dihidar (2014) and n is taken throughout as 24. For those data we have $\theta = 0.188$. For illustration of the simulation purpose, we have used the device parameters as $N_1 = 30, r_1 = 17, t = 9, N_2 = 33, r_2 = 13$ and $K = 12$. We have checked that all the conditions evolved in earlier sections are satisfied with these chosen device parameters. Below we present the relative efficiencies of the inverse mechanism for RR generation versus the direct one based on the various sampling schemes considered in this paper for comparative illustration purpose. We also show below few instances of our findings for the estimated standard error (se), which is the positive square root of $\hat{V}(\hat{\theta})$ and estimated coefficient of variation (cv), which is $cv = 100\frac{se}{\hat{\theta}}$ for various situations.

Table 1. Relative performances of the direct RRT versus the inverse RRT based on SRSWR

	Method 1:Direct RRT			Method 2:Indirect RRT		
serial number	est	se	cv	est	se	cv
1	0.227	0.192	84.582	0.183	0.118	64.481
2	0.289	0.187	64.706	0.372	0.180	48.387
3	0.229	0.127	55.459	0.201	0.105	52.239
4	0.209	0.132	63.158	0.212	0.115	54.245
5	0.278	0.150	53.957	0.292	0.144	49.315

Efficiency = 100(V(Method 1)/V(Method 2)) = 119.36.

Out of 100 cases estimated cv(Method 2) < estimated cv(Method 1) in 60 cases.

Table 2. Relative performances of the direct RRT versus the inverse RRT based on SRSWOR

	Method 1:Direct RRT			Method 2:Indirect RRT		
serial number	est	se	cv	est	se	cv
1	0.248	0.151	60.887	0.252	0.131	51.984
2	0.208	0.151	72.596	0.223	0.118	52.915
3	0.294	0.181	61.565	0.251	0.127	50.598
4	0.234	0.148	63.248	0.241	0.127	52.697
5	0.224	0.143	63.839	0.264	0.121	45.833

Efficiency = 100(V(Method 1)/V(Method 2)) = 120.52.

Out of 100 cases estimated cv(Method 2) < estimated cv(Method 1) in 65 cases.

We observe from Tables 1-5 that the randomized response model considered in this paper can be profitably modified by generating randomized responses by the inverse method having greater efficiencies in comparison to the direct one. Also, from the results obtained from the simulation exercise, it reveals that the inverse RRT has relatively lower values of the estimated coefficient of variations than the ones for

Table 3. Relative performances of the direct RRT versus the inverse RRT based on PPSWR

	Method 1:Direct RRT			Method 2:Indirect RRT		
serial number	est	se	cv	est	se	cv
1	0.214	0.193	90.187	0.174	0.146	83.908
2	0.149	0.128	85.906	0.162	0.131	80.864
3	0.139	0.108	77.698	0.192	0.108	56.250
4	0.289	0.202	69.896	0.164	0.102	62.195
5	0.214	0.131	61.215	0.194	0.106	54.639

Efficiency = 100(V(Method 1)/V(Method 2)) = 103.92.

Out of 100 cases estimated cv(Method 2) $<$ estimated cv(Method 1) in 54 cases.

Table 4. Relative performances of the direct RRT versus the inverse RRT based on Rao, Hartley and Cochran's sampling

	Method 1:Direct RRT			Method 2:Indirect RRT		
serial number	est	se	cv	est	se	cv
1	0.401	0.269	67.082	0.179	0.108	60.335
2	0.328	0.260	79.268	0.249	0.111	44.578
3	0.421	0.266	63.183	0.223	0.121	54.260
4	0.321	0.254	79.128	0.253	0.145	57.312
5	0.317	0.249	78.549	0.173	0.125	72.254

Efficiency = 100(V(Method 1)/V(Method 2)) = 104.89.

Out of 100 cases estimated cv(Method 2) $<$ estimated cv(Method 1) in 56 cases.

Table 5. Relative performances of the direct RRT versus the inverse RRT based on Midzuno's Scheme of sampling

	Method 1:Direct RRT			Method 2:Indirect RRT		
serial number	est	se	cv	est	se	cv
1	0.157	0.119	75.796	0.273	0.157	57.509
2	0.147	0.118	80.272	0.263	0.139	52.852
3	0.217	0.125	57.604	0.203	0.109	53.695
4	0.272	0.124	45.588	0.293	0.119	40.614
5	0.162	0.134	82.716	0.190	0.131	68.947

Efficiency = 100(V(Method 1)/V(Method 2)) = 120.20.

Out of 100 cases estimated cv(Method 2) $<$ estimated cv(Method 1) in 62 cases.

the direct RRT. Therefore, one may use the inverse method profitably in practical survey situation in place of the direct counterpart, in any general sampling scheme, some of which are considered here including the unequal probability sampling of respondents, and hence this is the justification of this research.

6. Concluding Remarks

The study in this paper is relevant for socio-economic surveys where the underlying variable is stigmatizing and qualitative and the objective is to estimate the population proportion of the variable. Here we propose a randomization device which generates the randomized response data from a combination of the binomial and hypergeometric distribution. We also present the alternative procedure of generating randomized responses by combining the inverse of these two distributions. While preparing the randomized response devices, we take care about the privacy of the respondents. In both cases we present the related estimation procedures considering the sample of respondents as chosen by simple random sampling and various unequal probability sampling schemes as well. At the same time, we concentrate on comparing these two approaches. Our numerical simulation-based comparison shows that the inverse approach may be used in practical survey situation in place of the direct approach not only in simple random sampling of respondents, but also profitably in general unequal probability sampling of respondents.

Acknowledgements

The authors are grateful to two anonymous reviewers for giving many insightful and constructive comments and suggestions which led to improvement of the earlier manuscript.

REFERENCES

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

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

CHAUDHURI, A., BOSE, M., DIHIDAR, K., (2011a). Estimation of a sensitive proportion by Warner's randomized response data through inverse sampling. Statistical Papers. 52, 343–354.

CHAUDHURI, A., BOSE, M., DIHIDAR, K., (2011b). Estimating sensitive proportions by Warner's randomized response technique using multiple randomized responses from distinct persons sampled, Statistical Papers, 52, 111–124.

CHAUDHURI, A., CHRISTOFIDES, T. C., RAO, C. R., (2016). Handbook of Statistics 34, Data gathering, analysis and protection of privacy through randomized response techniques. Elsevier, Amsterdam.

CHAUDHURI, A., DIHIDAR, K., (2014). Generating randomized response by inverse mechanism. Model Assisted Statistics and Applications, 9, 343–351.

CHAUDHURI, A., MUKERJEE, R., (1987). Randomized response techniques: a review. Statistica Neerlandica, 41, 27–44.

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

CLARK, S. J., DESHARNAIS, R. A., (1998). Honest answers to embarrassing: Detecting cheating in the randomized response model. Psychological Methods, 3 (2), 160–168.

DIHIDAR, K., (2016). Estimating sensitive population proportion by generating randomized response following direct and inverse hypergeometric distribution. Handbook of Statistics, 34 : Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits. Edited by Arijit Chaudhuri, Tasos C. Christofides and C.R. Rao. Elsevier, North Holland, Amsterdam, The Netherlands. 427–441.

EICHHORN, B. H., HAYRE, L. S., (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. Journal of Statistical Planning and Inference, 7, 307–316.

FETH, S., FRENGER, M., PITSCH, W., SCHMELZEISEN, P., (2015). Cheater Detection in the Randomized Response technique: derivation, analysis and application. Monsenstein & Vannerdat.
URL: http://universaar.uni-saarland.de/monographien/volltexte/2015/134/

GJESTVANG, C. R., SINGH, S., (2006). A new randomized response model. Journal of the Royal Statistical Society, Series B, 68, 523–530.

GREENBERG, B. G., ABUL-ELA, ABDEL-LATIF, A., SIMMONS, W. R., HORVITZ, D. G., (1969). The unrelated question RR model : theoretical framework. Journal of the American Statistical Association. 64, 520–539.

HANSEN, M. H., HURWITZ, W. N., (1943). On the theory of sampling from finite populations. Annals of Mathematical Statistics, 14 (4), 333–362.

HAUNG, K. (2004). A survey technique for estimating the proportion and sensitivity in a dichotomous finite population. Statistica Neerlandica, 58, 75-82.

HEDAYAT, A.S., SINHA, BIKAS. K., (1991). Design and Inference in Finite Population Sampling. Wiley.

HORVITZ, D. G., THOMPSON, D. J., (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663–685.

HORVITZ, D. G., SHAH, B. V., SIMMONS, W. R., (1967). The unrelated question randomized response model. Social Statistics Section, Proceedings of the American Statistical Association, 65–72.

KIM J., 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 (2), 436–438.

MANGAT, N.S., (1994). An improved randomized response strategy. Journal of the Royal Statistical Society, Series B, 56, 93–95.

MANGAT, N. S., SINGH, R., (1990). An alternative randomized response procedure. Biometrika, 77 (2), 439–442.

MIDZUNO, H., (1952). On the sampling system with probabilities proportionate to sum of sizes. Annals of the Institute of Statistical Mathematics, 3, 99–107.

MIR, K. A., (2008). Size-biased generalized negative binomial distribution. Journal of Modern Applied Statistical Methods, 7 (2), 446–453.

MOORS, J. J. A., (1971). Optimization of the Unrelated Question Randomized Response Model. Journal of the American Statistical Association, 66, 627–629.

RAGHAVARAO, D., (1978). On an estimation problem in Warner's randomized response technique. Biometrics, 34, 87–90.

RAO, J. N. K., HARTLEY, H. O., COCHRAN, W. G., (1962). On a simple procedure of unequal probability sampling without replacement. Journal of the Royal Statistical Society, Series B, 24, 482–491.

SHONKWILER, J. S., (2016). Variance of the truncated negative binomial distribution. Journal of Econometrics, 195, 209–210.

SINGH, S., GREWAL, I. S., (2013). Geometric distribution as a randomization device: Implemented to the Kuks model. International Journal of Contemporary Mathematical Sciences, 8 (5), 243–248.

SINGH, S., 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 the American Statistical Association, 60, 63–69.

YATES, F. & GRUNDY, P. M., (1953). Selection without replacement from within strata with probability proportional to size. Journal of the American Statistical Association, 75, 206–211.