Statistics and Applications {ISSN 2452-7395 (online)}
Volume 15 Nos. 1&2, 2017 (New Series), pp 79-92

Bootstrap technique for Randomized Response Surveys

Sanghamitra Pal
West Bengal State University, Barasat, West Bengal, India

Received: December 20, 2016; Revised: January 07, 2017; Accepted: January 24, 2017

Abstract

The objective of this paper is to suggest bootstrap technique for randomized response surveys dealing with the complex survey data. We mainly discuss modifications on Rao and Wu's (1988) rescaling bootstrap and Sitter's (1992) mirror-match bootstrap concentrating on estimation for correlation coefficients between two sensitive quantitative characteristics. The randomized responses (RR) are available for the two sensitive characters. Estimators for mean square error and confidence interval of the proposed estimator are also developed. A simulation study has been carried out to demonstrate our proposed method.

Keywords: Bootstrap, confidence interval, mean square error estimation, randomized response.

1. Introduction

For questions related to abortion, illegitimate birth, AIDs, illegal betting, shoplifting, drug-taking, tax evasion, annual income and students' cheating behavior etc.,some respondents may simply refuse to answer. Even worse, they may provide wrong answers to maintain privacy when these questions are being asked directly. Warner (1965) introduced the idea of Randomized Response (RR) technique for eliciting information on sensitive characters and later it has been studied by various authors. Horvitz et al. (1967) and Greenberg et al. (1969) proposed an unrelated question model introducing an character which is totally unrelated to the sensitive character. A rich growth of literature can be found in the monograph of Chaudhuri (2011). Fox and Tracy (1986), Hedayat and Sinha (1991, Chap. 11). Eichorn and Hayre (1983) suggested a multiplicative model to collect information on sensitive quantitative variables like income, tax evasion, amount of drug used etc.

Several techniques have been introduced for both of qualitative and quantitative variables. In many cases it is important to study on correlation coefficient estimators of two variables where they are of sensitive nature at the same time. This really prompted us to extend the bootstrap technique for randomized response survey data.

Rao and Wu (1988) proposed a "rescaling-bootstrap technique" to construct useful confidence intervals (CI) and mean square error estimation for non-linear functions of finite population totals of several variables like correlation coefficients, regression coefficients etc. without demanding normal distributions of relevant pivotals. In randomized response survey several extensions are needed to implement their techniques Sitter's (1992) mirror-match technique is also useful to deal with the above problem where samples are drawn by unequal

Corresponding author: Sanghamitra Pal
E-mail: mitrapal2013@gmail.com;

probability sampling eg. by Rao-Hartley-Cochran's (RHC, 1962) scheme. Their method seems reasonable to design a resampling scheme that parallels the original sampling scheme as closely as possible. For randomized response survey an extension of Sitter's (1992) mirror-match technique also has been developed where original samples are drawn by RHC (1962) scheme.

In Section 2 we deal with the estimation of non-linear statistics. Section 3 gives us the details of RR procedure. In Section 4.1, we extend our "rescaling-bootstrap technique" for RR survey data. Section 4.2 gives us the extension of Sitter's (1992) mirror-match bootstrap. A simulation study has been carried out in our Section 5. In Section 6, we conclude with a brief summary discussion.

2. Estimation of Non-linear Statistics

Given a finite population, namely U of N elements, let y denote a variable of interest with the population values y_i and x denote another variable of interest with the population values $x_i, i = 1,2,....,N$. The variables x and y are related to the sensitive characteristics A and B respectively. A and B may be related to illegal drugs, induced abortion etc. The collection of data through personal interview survey on such sensitive issue is a serious matter in social sciences. In this paper we are interested in the estimation of the finite population correlation coefficient and the regression coefficient between the variables x and y for complex survey design dealing with sensitive issues. Various resampling procedures for variance estimation and confidence intervals for non-linear statistics (such as correlation coefficient, regression coefficient etc.) in sample survey data have been proposed in the literature. Rao and Wu (1988), describe an application of the bootstrap under the design-based approach to sample survey inference. It is "Rescaling bootstrap" technique. Sitter (1992) has proposed his "mirror-match" method for complex survey design.

In this article we try to modify their approaches where x and y are related to the sensitive characters A and B respectively. The parameter of interest is assumed to be a non-linear function of several population totals. In this paper we are considering the regression coefficient and the correlation coefficient for our detailed estimation. The correlation coefficient between x and y, say R_{xy} and the regression coefficient of y on x, say B_{yx} may be written as

$$R_{xy} = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})}{[\sum_{i=1}^{N}(y_i - \bar{Y})^2 \sum_{i=1}^{N}(x_i - \bar{X})^2]^{1/2}} \text{ where } \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \text{ and } \bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$= \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{[N\sum_{i=1}^{N} y_i^{\,2} - (\sum_{i=1}^{N} y_i)^2]^{1/2}[\sum_{i=1}^{N} x_i^{\,2} - (\sum_{i=1}^{N} x_i)^2]^{1/2}}$$

$$= g(N, \sum_{i=1}^{N} y_i, \sum_{i=1}^{N} x_i, \sum_{i=1}^{N} y_i x_i, \sum_{i=1}^{N} y_i^{\,2}, \sum_{i=1}^{N} x_i^{\,2}). \qquad (1)$$

and $B_{yx} = \dfrac{\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})}{\sum_{i=1}^{N}(x_i - \bar{X})^2} = \dfrac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^{\,2} - (\sum_{i=1}^{N} x_i)^2} = f(N, \sum_{i=1}^{N} y_i, \sum_{i=1}^{N} x_i, \sum_{i=1}^{N} x_i y_i, \sum_{i=1}^{N} x_i^{\,2})$. (2)

So both R_{xy} and B_{yx} are non-linear functions (g and f) of the totals namely $N, \sum_{i=1}^{N} y_i, \sum_{i=1}^{N} x_i, \sum_{i=1}^{N} y_i x_i, \sum_{i=1}^{N} y_i^{\,2} and \sum_{i=1}^{N} x_i^{\,2}$.

A sample s is chosen according to a design P with a selection probability $P(s)$. By E_P and V_P we shall denote operators for expectation and variance with respect to the design P. An estimate of the parameters R_{xy} and B_{yx} are given by $\hat{R}_{xy} = g(\hat{N}, e_y, e_x, e_{xy}, e_{y^2}, e_{x^2})$ where $\hat{N}, e_y, e_x, e_{xy}, e_{y^2}, e_{x^2}$ are the unbiased estimates for the totals $N, \sum_{i=1}^{N} y_i, \sum_{i=1}^{N} x_i, \sum_{i=1}^{N} y_i x_i, \sum_{i=1}^{N} y_i^{\,2}$ and $\sum_{i=1}^{N} x_i^{\,2}$ respectively. Similarly, $\hat{B}_{yx} = f(\hat{N}, e_y, e_x, e_{xy}, e_{x^2})$. In section 3 we shall describe the estimation of above totals in detail.

Chaudhuri (2011) has given a general treatment of RR adopting general sampling schemes liberating it from the erstwhile compulsion of simple random sampling with replacement (SRSWR) alone. In this paper general schemes are adopted following Chaudhuri's (2011) approach.

We suppose that the values y_i's and x_i's, $i = 1,2,....,N$ are non-ascertainable for a person i in a sample. In Section 3 we describe the Randomized Response (RR) device to estimate the values y_i's and x_i's respectively.

3 Randomized Response Procedure and Overall Estimation of the Totals

3.1. RR estimation for y_i and x_i

The respondent labeled $i\,(i = 1,2,.....N)$ may be instructed to report only a "randomized response" (RR) following a specified device . The variable y is supposed to relate to a stigmatizing characteristic A. The characteristic A may be related to earning by gambling, bribery, number of days of drunken driving, amounts of tax evaded and the like.

Eichhorn and Hayre(1983) have considered a scrambled response technique involving a quantitative response variable and proposed an RR technique for it. According to them, ith respondent in the sample is requested to report the scrambled response $Z_i = Sy_i$ where y_i is the real value of the sensitive character A for ith person. and S is the scrambling variable whose distribution is assumed to be known. Later, Pollock and Bek (1976), Singh, Joarder and King (1996), Padmawar and Vijayan (2000), Gupta et al (2002), Bar-Lev et al. (2004) , Pal (2008) and others dealt with certain aspects of scrambled response technique.

Taking a cue from Eichhorn and Hayre(1983)'s scrambled response technique, here the respondents are requested to report the value

$$y_i \frac{u_i}{\theta} = r_i \text{ (say).}$$

Here u (non-stigmatizing scrambling variable) be a random variable with the values u_i $(i = 1,2,...,N)$ (with a known non-zero mean θ and a known positive standard deviation σ_u).

Denoting E_R and V_R as the expectation and variance operators with respect to RR device, we can write $E_R(r_i) = y_i \frac{\theta}{\theta} = y_i$.

and $V_R(r_i) = \frac{y_i^2}{\theta^2}(\theta^2 + \sigma_u^2) - y_i^2 = \frac{y_i^2 \sigma_u^2}{\theta^2}$.

Writing $\mathrm{v}_i = r_i^2 = \mathrm{v}_i(y_i)$, say (3)

it is easy to write $E_R(\mathrm{v}_i) = V_R(r_i)$.

(Since y_i^2 may be estimated by $\hat{y}_i^2 = \frac{\theta^2}{\theta^2 + \sigma_u^2} r_i^2$).

Similarly, the variable x is related to the characteristic B. Let u' be another random variable(scrambling variable) with a known non-zero mean θ' and a known positive standard deviation σ_u'.

The respondents are also requested to report scrambled response $x_i \frac{u_i'}{\theta'} = r_i'$ (say)

for the variable x .

Here $E_R(r_i') = x_i$; $V_R(r_i') = \frac{x_i^2 \sigma_u'^2}{\theta'^2}$ and $\mathrm{v}_i' = r_i'^2 = \mathrm{v}_i'(x_i)$ (4)

So, $E_R(\mathrm{v}_i') = V_R(r_i')$.

So adopting a suitable Randomized Response (RR) device, from an i in a sample, an RR may be procured as r_i and r_i' (for y_i and x_i respectively) such that

(i) $E_R(r_i) = y_i$, (ii) $V_R(r_i) = V_i (> 0)$, (iii) r_i's are independent over i in U, (iv) $\exists$ v_i such that $E_R(\mathrm{v}_i) = V_i, i \in U$ and

(i)' $E_R(r_i') = x_i$, (ii)' $V_R(r_i') = V_i' (> 0)$, (iii)' r_i''s are independent over i in U, (iv)' $\exists$ v_i' such that $E_R(\mathrm{v}_i') = V_i', i \in U$ where E_R and V_R denote operators for expectation and variance with respect to the RR devices. An interested reader may see Chaudhuri (2001) for the clarifications of (i) to (iii) and (i)' to (iii)'.

3.2. RR estimation for $x_i y_i$, y_i^2 and x_i^2

Suppose the variable $\log u \approx normal(\log \alpha, \sigma^2)$ then with mean $u \approx normal$ with mean $\theta = \alpha e^{\sigma^2/2}$ and variance $\sigma_u^2 = \alpha^2 e^{\sigma^2}(e^{\sigma^2} - 1)$.

Similarly, the variable $\log u' \approx normal(\log \alpha', \sigma'^2)$ with mean $\theta' = \alpha' e^{\sigma'^2/2}$ and variance $\sigma_u'^2 = \alpha'^2 e^{\sigma'^2}(e^{\sigma'^2} - 1)$.

From this we may write that $\log u^2 \approx normal(2\log\alpha, 4\sigma^2)$, so u^2 follows lognormal with mean $\alpha^2 e^{2\sigma^2}$ and variance $\alpha^4 e^{4\sigma^2}(e^{4\sigma^2}-1)$.

Also, $\log u'^2 \approx normal(2\log\alpha', 4\sigma'^2)$, so u'^2 follows lognormal with mean $\alpha'^2 e^{2\sigma'^2}$ and variance $\alpha'^4 e^{4\sigma'^2}(e^{4\sigma'^2}-1)$.

Also, $\log(uu') \approx normal(\log(\alpha\alpha'), (\sigma^2+\sigma'^2))$. Then uu' follows lognormal with mean $\alpha\alpha' e^{(\sigma^2+\sigma'^2)/2}$ and variance $(\alpha\alpha')^2 e^{(\sigma^2+\sigma'^2)}(e^{(\sigma^2+\sigma'^2)-1})$.

Here y_i^2 may be unbiasedly estimated by $\hat{y_i^2} = \frac{\theta^2}{\theta^2+\sigma_u^2}\mathrm{v}_i$. We may write

$$E_R(\hat{y_i^2}) = y_i^2;\ V_R(\hat{y_i^2}) = E_R(\hat{y_i^2})^2 - y_i^4 = \frac{\theta^4}{(\theta^2+\sigma_u^2)^2}E_R(r_i^4) - y_i^4. \text{ (since } \mathrm{v}_i = r_i^2)$$

The term y_i^4 may be unbiasedly estimated by $\hat{y_i^4} = \frac{\theta^4}{E_R(u_i^4)}r_i^4 = \frac{\theta^4 r_i^4}{\alpha^4 e^{8\sigma^2}}$. On simplification, an unbiased estimate of $V_R(\hat{y_i^2})$ is

$$\hat{V}_R(\hat{y_i^2}) = \alpha^2\frac{r_i^4}{e^{4\sigma^2}}(e^{4\sigma^2}-1) = \mathrm{v}_i(y_i^2). \text{ (say)} \tag{5}$$

Here $\sigma_u^2 = \alpha^2 e^{\sigma^2}(e^{\sigma^2}-1)$ and $E_R(u_i^4) = \alpha^4 e^{8\sigma^2}$.

Similarly $V_R(y_i'^2)$ may be estimated by $\mathrm{v}_i(x_i^2) = \alpha'^2\frac{r_i'^4}{e^{4\sigma'^2}}(e^{4\sigma'^2}-1)$ (6)

Now, $E_R(r_i r_i') = y_i x_i E_R(\frac{u_i u_i'}{\theta\theta'}) = y_i x_i \alpha\alpha'\frac{e^{\frac{1}{2}(\sigma^2+\sigma'^2)}}{\theta\theta'} = y_i x_i$. So the term $y_i x_i$ may be unbiasedly estimated by $r_i r_i'$.

Writing, $V_R(r_i r_i') = (\frac{y_i x_i}{\theta\theta'})^2 V_R(u_i u_i') = (\frac{y_i x_i}{\theta\theta'})^2(\alpha\alpha')^2 e^{(\sigma^2+\sigma'^2)}(e^{(\sigma^2+\sigma'^2)-1})$, the term $V_R(y_i x_i)$ may be estimated by

$$v_i(x_i y_i) = \frac{(\theta\theta')^2(r_i r_i')^2}{E_R(u_i u_i')^2} = \frac{(\theta\theta')^2(r_i r_i')^2}{(\alpha\alpha')^2 e^{2(\sigma^2+\sigma'^2)}} = \frac{(r_i r_i')^2}{e^{(\sigma^2+\sigma'^2)}}. \tag{7}$$

3.3. Overall estimation for the totals $\sum_{i=1}^N y_i$, $\sum_{i=1}^N x_i$, $\sum_{i=1}^N y_i^2$, $\sum_{i=1}^N x_i^2$ and $\sum_{i=1}^N x_i y_i$

Denoting the design based expectation and variance as E_P and V_P, the overall expectation and variance can be written as $E = E_P E_R = E_R E_P$ and $V = E_P V_R + V_P E_R = E_R V_P + V_R E_P$.

Writing $I_{si}=1$ if $i\in s$, 0, else, $I_{sij}=I_{si}I_{sj}$ and considering Horvitz-Thompson (HT, 1952) method of estimation, Chaudhuri's (2001) recommended unbiased estimators for y and x based on RR are

$$e_b=\sum_{i\in s}\frac{r_i}{\pi_i}=e_y, \text{say for which } E(e_b)=\sum_{i=1}^{N}y_i \tag{8}$$

$$\text{and } e_b'=\sum_{i\in s}\frac{r_i'}{\pi_i}=e_x, \text{ say for which } E(e_b')=\sum_{i=1}^{N}x_i \tag{9}$$

where π_i is the first order inclusion probability for the unit i, $\pi_i=\sum_{s\ni i}p(s)$.

$$V(e_b)=E_PV_R(e_b)+V_PE_R(e_b)=E_P(\sum_{i\in s}\frac{V_i}{\pi_i^2})+V_P(\sum_{i\in s}\frac{y_i}{\pi_i})$$

The variance of e_b is then,

$$=\sum_{i=1}^{N}\frac{V_i}{\pi_i}+\sum_{i<j=1}^{N}\sum(\pi_i\pi_j-\pi_{ij})(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j})^2$$

Here π_{ij} is the second order inclusion probability for the unit i and $j(i\neq j)$. So, $\pi_{ij}=\sum_{s\ni i,j}p(s)$.

Two unbiased estimators for $V(e_b)$ are

$$\mathrm{v}_1(e_b)=\sum_{i\in s}\frac{\mathrm{v_i}}{\pi_i^2}+\mathrm{v_r}-\sum_{i<j\in s}\sum\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}(\frac{\mathrm{v_i}}{\pi_i^2}+\frac{\mathrm{v_j}}{\pi_j^2}) \tag{10}$$

$$\text{and } \mathrm{v}_2(e_b)=\mathrm{v_r}+\sum_{i\in s}\frac{\mathrm{v_i}}{\pi_i^2} \tag{11}$$

(where $\mathrm{v_r}=\sum_{i<j\in s}\sum\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}(\frac{\mathrm{r_i}}{\pi_i}-\frac{\mathrm{r_j}}{\pi_j})^2$) for which

$E_PE_R(\mathrm{v}_1(e_b))=V(e_b)=E_RE_P((\mathrm{v}_2(e_b))$.

$V(e_b')$ is same as $V(e_b)$ only replacing y_i's by x_i's and V_i's by V_i''s. And their unbiased estimators are $\mathrm{v}_1(e_b')$ and $\mathrm{v}_2(e_b')$ are also same as $\mathrm{v}_1(e_b)$ and $\mathrm{v}_2(e_b)$ only replacing r_i by r_i''s and v_i by v_i''s. In this paper we shall always use $\mathrm{v}_2(e_b)$ as it always produces non-negative variance estimate.

Writing $e_y(=e_b$ here) for the Horvitz-Thompson estimator (HTE) for a total of a variable y, the population parameter R_{xy} may be estimated a

$$\hat{R}_{xy}=\frac{\hat{N}e_{xy}-e_xe_y}{[\hat{N}e_{x^2}-e_x^2]^{1/2}[\hat{N}e_{y^2}-e_y^2]^{1/2}} \tag{12}$$

where $e_y=e_b;\ e_x=e_b';\ e_{xy}=\sum_{i\in s}\frac{\hat{x_iy_i}}{\pi_i};e_{x^2}=\sum_{i\in s}\frac{\hat{x_i^2}}{\pi_i};\ e_{y^2}=\sum_{i\in s}\frac{\hat{y_i^2}}{\pi_i};\hat{N}=\sum_{i\in s}\frac{1}{\pi_i}$

and similarly $\hat{B}_{yx}=\frac{\hat{N}e_{xy}-e_xe_y}{\hat{N}e_{x^2}-e_x^2}$. (13)

The population size N has to be estimated to make sure of the fact $0 < \hat{R}_{xy} < 1$.

The terms $\hat{x_i y_i}$, $\hat{x_i^2}$ and $\hat{y_i^2}$ are the estimated RR values for the variables $x_i y_i, x_i^2$ and y_i^2. Here

$E_R(\hat{x_i y_i}) = x_i y_i,\ V_R(\hat{x_i y_i}) = V_i(x_i y_i)$;

$E_R(\hat{x_i^2}) = x_i^2,\quad V_R(\hat{x_i^2}) = V_i(x_i^2)$;

$E_R(\hat{y_i^2}) = y_i^2,\quad V_R(\hat{y_i^2}) = V_i(y_i^2)$

Proceeding as in (4), the unbiased estimators of $V(e_{xy}), V(e_{y^2}), V(e_{x^2})$ are

$$\mathrm{v}_2(e_{xy}) = \sum\sum \frac{\pi_i\pi_i - \pi_{ij}}{\pi_{ij}}\left(\frac{\hat{x_i y_i}}{\pi_i} - \frac{\hat{x_j y_j}}{\pi_j}\right)^2 + \sum_{i\in s}\frac{\mathrm{v_i}(x_i y_i)}{\pi_i},$$

$$\mathrm{v}_2(e_{y^2}) = \sum\sum \frac{\pi_i\pi_i - \pi_{ij}}{\pi_{ij}}\left(\frac{\hat{y_i^2}}{\pi_i} - \frac{\hat{y_j^2}}{\pi_j}\right)^2 + \sum_{i\in s}\frac{\mathrm{v_i}(y_i^2)}{\pi_i^2},$$

and $\mathrm{v}_2(e_{x^2}) = \sum\sum \frac{\pi_i\pi_i - \pi_{ij}}{\pi_{ij}}\left(\frac{\hat{x_i^2}}{\pi_i} - \frac{\hat{x_j^2}}{\pi_j}\right)^2 + \sum_{i\in s}\frac{\mathrm{v_i}(x_i^2)}{\pi_i^2}$ where $\mathrm{v_i}(x_i y_i), \mathrm{v_i}(y_i^2)$ and $\mathrm{v_i}(x_i^2)$ are unbiased RR related variance estimators of $V_i(x_i y_i), V_i(y_i^2)$ and $V_i(x_i^2)$ respectively (see 5,6,7 etc.). Using the above results we shall go ahead to modify on Rao and Wu's (1988) rescaling bootstrap procedure in the RR context.

4. Proposed Bootstrap Procedures

4.1 Modifications on Rescaling Bootstrap procedure for RR survey

Here it is assumed that the non-linear statistics are non-linear functions of several linear HT estimators. For mean square error estimation and to construct the confidence interval of non-linear statistics, Rao and Wu (1988) proposed a rescaling bootstrap method which rescales the resampled values appropriately so that the resulting bootstrap variance of the linear bootstrap estimator is the same as the usual unbiased variance estimate in the linear case. Denoting the expectation and variance operators for the bootstrap sampling as E^* and V^* and the bootstrap estimate of the population total Y as e_y^* and following Rao and Wu (1988), we may write

$E_*(e_y^*) = e_b = e_y$ and $V_*(e_y^*) = \mathrm{v}_2(e_y)$.

In finite population sampling the theory we can not verify that the bootstrap distribution needs to converge to the sampling distribution. One may see Rao and Wu (1988) and Sitter (1992).

The resulting bootstrap estimates for the correlation coefficient R_{xy} and the regression coefficient B_{yx} are

$$\hat{R}_{xy}^{*} = \frac{\hat{N}^{*} e_{xy}^{*} - e_{x}^{*} e_{y}^{*}}{[\hat{N}^{*} e_{x^2}^{*} - e_{x}^{*2}]^{1/2} [\hat{N}^{*} e_{y^2}^{*} - e_{y}^{*2}]^{1/2}} \tag{14}$$

$$\text{and } \hat{B}_{yx}^{*} = \frac{\hat{N}^{*} e_{xy}^{*} - e_{x}^{*} e_{y}^{*}}{\hat{N}^{*} e_{x^2}^{*} - e_{x}^{*2}} \tag{15}$$

where $\hat{N}^{*} = \sum_{i \in S^{*}} \frac{1}{\pi_i}$ denoting s^{*} as the bootstrap sample drawn from the original sample s. $N^{*}, e_{xy}^{*}, e_{x^2}^{*} e_{y^2}^{*}$ are the bootstrap estimates of the totals $N, \sum_{i=1}^{N} x_i y_i, \sum_{i=1}^{N} x_i^{2}, \sum_{i=1}^{N} y_i^{2}$ respectively. Now we describe the bootstrap methodology for the total $\sum_{i=1}^{N} y_i$.

The bootstrap methodology is as follows.

(1) Out of $n(n-1)$ ordered pairs of units $(i, j)(i \neq j)$ in s, let the first Bootstrap sample s_1^{*} of pairs (i^{*}, j^{*}) in m draws with replacement (WR) be chosen with probabilities $q_{i^{*}j^{*}} (i^{*} \neq j^{*}), q_{i^{*}j^{*}} \neq q_{j^{*}i^{*}}$. With numbers $k_{i^{*}j^{*}}$ to be presently specified, let

$$\mathrm{e}_1^{*} = \frac{1}{m} \sum_{i^{*}j^{*} \in S_1^{*}} \sum k_{i^{*}j^{*}} \left(\frac{r_{i^{*}}}{\pi_{i^{*}}} - \frac{r_{i^{*}}}{\pi_{i^{*}}}\right) + \sum_{i \in S} \frac{r_i}{\pi_i}$$

Then, $\mathrm{E}_{*}(\mathrm{e}_1^{*}) = \sum_{i \neq j \in s_1} \sum q_{ij} k_{ij} \left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j}\right) + e_y$

and $V_{*}(\mathrm{e}_1^{*}) = \frac{1}{m} \sum_{i \neq j \in s} \sum q_{ij} k_{ij}^{2} \left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j}\right)^2$.

$V_{*}(\mathrm{e}_1^{*})$ equals $\sum_{i \neq j} \sum \frac{\pi_i \pi_i - \pi_{ij}}{\pi_{ij}} \left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j}\right)^2$ on choosing $m = n(n-1), q_{ij} = \frac{1}{m}$

and $k_{ij} = m\left(\frac{\pi_i \pi_i - \pi_{ij}}{\pi_{ij}}\right)^{1/2}, i \neq j \in s$.

(2) Let us now draw from s, a second bootstrap sample s_2^{*}, 'independently' of the selection of s_1^{*} following the Poisson sampling scheme as described by Hájek (1958) with k_i as the probability of 'success' associated with i in s implementing a Bernoullian trial.

Let $\mathrm{e}_2^{*} = \sum_{i^{*} \in s_2^{*}} \frac{\sqrt{\frac{\mathrm{v}_{i^{*}}}{(\pi_{i^{*}})^2}}}{k_{i^{*}}} - \sqrt{\frac{\mathrm{v}_{\mathrm{i}}}{(\pi_{\mathrm{i}})^2}}$

Then, $\mathrm{E}_{*}(\mathrm{e}_2^{*}) = 0$

and $V_{*}(\mathrm{e}_2^{*}) = \sum_{i \in s} \left(\frac{1}{k_i} - 1\right) \frac{\mathrm{v}_{\mathrm{i}}}{\pi_{\mathrm{i}}^{2}}$.

The final bootstrap estimate for $\sum_{i=1}^{N} y_i$ is $e_y^{*} = e_1^{*} + e_2^{*}$.

Now, $E_*(e_y{}^*) = \sum_{i\in S} \frac{r_i}{\pi_i} = e_y$ and $V_*(e_y{}^*) = \mathrm{v}_2(e_y)$. Here our bootstrap sample $s*$ is composed of two samples $s_{1}*$ and $s_{2}*$.

Next we proceed similarly with the other totals namely N, $\sum_{i=1}^{N} x_i y_i$, $\sum_{i=1}^{N} x_i$, $\sum_{i=1}^{N} x_i{}^2$ and $\sum_{i=1}^{N} y_i{}^2$.

Next we calculate the final bootstrap estimates for correlation coefficient and the regression coefficients using (14) and (15).

Calling such a sample s^* as a bth bootstrap sample $s_b{}^*$, one has now to replicate the same, a large number of times, say, B=10,000 and calculate $\hat{R}_{xy}{}^*$ and $\hat{B}_{yx}{}^*$ for these B replicated bootstrap samples $s_b{}^*, b = 1,2,....,B$. Then,

$\overline{R}_{xy} = \frac{1}{B}\sum_{b=1}^{B} \hat{R}_{xy}{}^*(b)$ gives us the "bootstrap estimate of the population correlation coefficient" and we take

$$\mathrm{v}_g{}^R = \frac{1}{B-1}\sum_{b=1}^{B} (\hat{R}_{xy}{}^*(b) - \overline{R}_{xy})^2 \qquad (16)$$

as the bootstrap estimate of the Mean Square Error (MSE) of the original estimator $\hat{R}_{xy}$. Similarly, for the regression coefficient B_{yx}, the bootstrap estimate and the bootstrap estimate of the Mean Square Error (MSE) of the original estimator $\hat{B}_{yx}$ are

$$\overline{B}_{yx} = \frac{1}{B}\sum_{b=1}^{B} \hat{B}_{yx}{}^*(b) \text{ and } \mathrm{v}_g{}^B = \frac{1}{B-1}\sum_{b=1}^{B} (\hat{B}_{yx}{}^*(b) - \overline{B}_{yx})^2 \text{ respectively.} \qquad (17)$$

A 95% confidence interval (CI) for $\overline{R}_{xy}$ is then calculated as $(L_{2.5}(R), U_{97.5}(R))$ using the lower 2.5% tail point $L_{2.5}(R)$ and the upper 2.5% tail point $U_{97.5}(R)$ of the 'histogram' of the $\hat{R}_{xy}{}^*(b)$ values, $b = 1,.....,B$. This is by the well known 'Percentile method'. An alternative 'Double Bootstrap' CI may be calculated as follows. From $s_b{}^*$, B=10,000 more bootstrap samples are independently drawn using the same bootstrap sampling scheme and they are used to calculate $\mathrm{v}_g{}^R$ or $\mathrm{v}_g{}^B$ as above—to be denoted by $\mathrm{v}_g{}^R(b)$ or $\mathrm{v}_g{}^B(b)$. This is repeated for every initial $\mathrm{b} = 1,2,....., \mathrm{B} = 10{,}000$. For the histogram of

$\frac{\hat{R}_{xy}{}^*(b) - \hat{R}_{xy}}{\sqrt{\mathrm{v}_g{}^R(b)}}, b = 1,2,....., B$, the lower 2.5% point $l_{2.5}$ and the upper 2.5% point $u_{97.5}$

are then calculated and $\hat{R}_{xy} - u_{97.5}\sqrt{\mathrm{v}_g{}^R}, \hat{R}_{xy} - l_{2.5}\sqrt{\mathrm{v}_g{}^R}$ is taken as the 95% 'double bootstrap' CI for Y. Similarly the 95% double bootstrap CI may be calculated for $\hat{B}_{yx}$. For more clarification in bootstrap confidence intervals we may refer Rao and Wu (1988).

4.2. Extension on Sitter's Mirror-match bootstrap (1992) for Randomized response surveys

Sitter's (1992) mirror-match technique is useful to construct bootstrap samples from an initial sample chosen by Rao, Hartley and Cochran (RHC, 1962) scheme. The sampling method is as follows.

We suppose that from U, let a sample of n units be chosen by Rao-Hartley and Cochran (RHC, 1962) scheme. In order to select a sample of n units from a population of size N, the N units are divided at random into n groups of sizes N_i each of which is closest to $\frac{N}{n}$ subject to $\sum_n N_i = N$, denoting by $\sum_n$ the sum over the n groups. Writing Q_i as the sum of the p_i's of the N_i units in the ith group for the RHC scheme, the RHC unbiased estimator for the total $Y = \sum_{i=1}^{N} y_i$ is $e_b(RHC) = \sum_n \frac{Q_i}{p_i} r_i$.

Proceeding as in Section 4.1, an unbiased estimator of the variance of $e_b(RHC)$ is

$$\mathrm{v}_2(e_b(RHC)) = B[\sum_n \frac{Q_i}{{p_i}^2} {r_i}^2 - {e_b}^2] + \sum_n \frac{Q_i}{{p_i}^2} r_i \qquad (18)$$

$$= \text{(I) + (II)}$$

where $B = \frac{\sum_n {N_i}^2 - N}{N^2 - \sum_n {N_i}^2}$, The expression (I) is $B[\sum_n \frac{Q_i}{{p_i}^2} {r_i}^2 - {e_b}^2]$ and (II) is $\sum_n \frac{Q_i}{{p_i}^2} r_i$. For this $E_R E_P(\mathrm{v}_2(e_b(RHC))) = V(e_b(RHC))$.

In RR surveys two bootstrap samples are required as an extension of Sitter's Mirror match method. The method is as follows.

(I) For the first bootstrap sample let

(i) $a_i = \frac{Q_i}{p_i} r_i, i = 1,2,....,n$

ii) Choose $1 \le n^* < n,$ an integer and randomly partition the a_i's into n^* groups of sizes $n_g{}^* = \frac{n}{n^*}$

iii) Randomly select one a_i from each of the n^* groups with probability $\frac{p_i}{Q_i^*}$ where $Q_i^* = \sum_{i=1}^{n^*} p_i$ to get $a_1{}^*, a_2{}^*, ..., a_{n^*}{}^*$.

iv) Repeat steps (ii)-(iii) $k = \frac{(\sum_1^{n^*} n_g{}^* - n)}{n(n-1)} . \frac{N^2 - \sum_n {N_i}^2}{\sum_n {N_i}^2 - N}$ times independently, replacing the $a_i's$ drawn at each step, k is integer valued.

For the first bootstrap sample $s_1^*(RHC))$, define

$t_1^* = \frac{1}{k}\sum_{l=1}^{k}\sum_{i=1}^{n^*}\frac{Q_i^*}{p_i}a_i$.

Then, $E_*(t_1^*) = \frac{1}{k}\sum_{l=1}^{k}(\sum_n a_i) = \frac{1}{k}\sum_{l=1}^{k}(e_b(RHC)) = e_b(RHC)$.

The bootstrap variance is

$$V_*(t_1^*) = \frac{1}{k}\frac{(\sum_1^{n^*} n_g^* - n)}{n(n-1)}[\sum_n \frac{a_i^2}{R_i} - (\sum_n a_i)^2]$$

$$= \frac{1}{k}\frac{(\sum_1^{n^*} n_g^* - n)}{n(n-1)}[\sum_n \frac{Q_i^2 r_i^2}{Q_i p_i^2} - (e_b(RHC))^2]$$

$$= \frac{1}{k}\frac{(\sum_1^{n^*} n_g^* - n)}{n(n-1)}[\sum_n \frac{Q_i r_i^2}{p_i^2} - (e_b(RHC))^2] \quad (19)$$

Equating (I) of (18) and (19) we have,

$\frac{1}{k}\frac{(\sum_1^{n^*} n_g^* - n)}{n(n-1)} = \frac{N^2 - \sum_n N_i^2}{\sum_n N_i^2 - N}$ which gives $k = \frac{(\sum_1^{n^*} n_g^* - n)}{n(n-1)} \cdot \frac{N^2 - \sum_n N_i^2}{\sum_n N_i^2 - N}$

(II) From the original sample, a second bootstrap sample $s_2^*(RHC)$ (independently of $s_1^*(RHC))$ is to be drawn by poisson sampling scheme with success probability l_i.

Define, $t_2^* = \frac{1}{k'}[\sum_{S_2^*}\frac{1}{l_i^*}r_i^* - \sum_n r_i]$

Then, $E_*(t_2^*) = 0$; $V_*(t_2^*) = \frac{1}{k'}\sum_n(\frac{1}{l_i} - 1)r_i^2$ (20)

Equating (II) of equation (18) and (20) we have $\frac{1}{l_i} - 1 = \frac{Q_i}{p_i} \Rightarrow l_i = \frac{p_i}{Q_i + p_i}$

Now the combined bootstrap estimate (using Siiter's (1992) method) of $\sum_{i=1}^{N} y_i$ is $t^* = t_1^* + t_2^*$.

For this $E_*(t^*) = e_b(RHC)$ and $V_*(t^*) = V(e_b(RHC))$. This is the procedure for the total $\sum_{i=1}^{N} y_i$ only. A similar approach has been done for the other totals mentioned in Section 2. Proceeding as described in Section 4.1 we calculate here bootstrap estimates, bootstrap estimate of the Mean Square Error (MSE) the other related comparison criteria for the Sitter's (1992) scheme.

5. Simulation Study

A randomized response survey was undertaken among the students of a medical college at Calcutta. Here the characters A and B denote the amount of expenditure on alcohol consumption and the amount of expenditure on tobacco respectively. The variables x and y are related to the characters A and B respectively. We treat $N = 117$ medical students as our population. To estimate the population correlation coefficient (R_{xy}) between x and y and the regression coefficient (B_{yx}) of y on x, a sample (s) of size n is drawn by Hartley-Rao (HR, 1962) scheme which is required in applying Rescaling bootstrap procedure. The family income of a particular student is the size measure variable in drawing a sample. In the HR scheme the units of U are permuted at random and then n units are chosen systematically with probabilities proportional to sizes.

The sampled persons are requested to report their randomized responses (RR) as mentioned in our Section 3.1.

We considered a practicable choice of parameters as:

$\theta = 3.4,\ \sigma^2 = 20.3, \theta' = 29.6,\ \sigma^{2'} = 34.2$ and $\alpha = 4.3, \alpha' = 13.2$. Here $N = 117,\ n = 13$.

Another sample of size n is also drawn by Rao-Hartley-Cochran (RHC, 1962) scheme to apply Sitter's Mirror Match (1992) Procedure.

Now we examine possible efficacies of our approach. Usually, a very large number of identically independently drawn bootstrap samples are required to be taken, labeled, say, as $b = 1,2,....., B$. Here B is taken as 10,000. In order to judge the efficiency of the proposed method, the population is assumed to be known. We draw $T = 1000$ samples from the population to calculate ACP, ACV and the AL values. For each sample we perform bootstrap methods to calculate bootstrap estimates for population correlation and regression coefficients using B bootstrap samples.

The point estimator will be judged good if the coefficient of variation, namely

$CV = 100 \dfrac{\sqrt{\text{Estimated } MSE(estimate)}}{estimate}$ has a small magnitude, preferably smaller than 10% or at most 30% . *MSE(estimate)* is the mean square error of the required estimate. A CI will be judged good if on drawing a large number of simulated samples, say, T in number taken as 1000, from a population at hand, the (i) CI's happen to cover the known value of the parameter, say, R_{xy} or B_{yx} a percentage of times close to 95% – this percentage is called the ACP, the Average Coverage Percentage, and (ii) if the average value of the length, AL, say, of a CI is small enough. Between two CI's the one with a smaller value of AL will be preferred unless its ACP is too far from 95% compared to that for the other.

Needless to say, our method extends to cover several non-linear functional other than correlation and regression coefficients with no extra efforts.

Based on the $T = 1000$ such simulated samples, if the average of the values of *CV*'s be calculated, called the ACV, the average coefficient of variation, then the point bootstrap estimate is judged good if this ACV is good, preferably smaller than 10% or at most 30% in value.

In Table 1 we illustrate results based on HR (1962) scheme. *Estimates* and their *Standard Errors* (*SE*) of 500th simulated sample are given. ACP, ACV and AL values are based on 1000 simulated samples. Table 2 gives us the same for the samples drawn by RHC (1962) scheme.

Table 1: Bootstrap Estimates and their MSE estimates for correlation coefficient and regression coefficient for 500th (t=500) simulated sample

(The Population Correlation Coefficient to be estimated is 0.76 *and the Population Regression Coefficient to be estimated is* 0.42*)*

Bootstrap estimate	**Sampling scheme**			
	HR		**RHC**	
	Estimate (t)	$\sqrt{MSE}$	***t***	$\sqrt{MSE}$
For correlation coefficient	$\overline{R}_{xy}$: 0.61	$\sqrt{v_g^{R}}$: 0.49	$\overline{R}_{xy}$: 0.83	$\sqrt{v_g^{R}}$ 0.61
For regression coefficient	$\overline{B}_{yx}$: 0.17	$\sqrt{v_g^{B}}$: 0.46	$\overline{B}_{yx}$: .34	$\sqrt{v_g^{B}}$: 0.21

Table 2: Performances of correlation coefficient estimator and regression coefficient estimator with the criteria ACP, ACV and AL

(The Population Correlation Coefficient to be estimated is 0.76 and the Population Regression Coefficient to be estimated is 0.42)

Method		**Sampling scheme**					
		HR			**RHC**		
		ACP	**ACV**	**AL**	**ACP**	**ACV**	**AL**
Bootstrap for correlation coefficient	**Percentile method**	92.2	15.08	0.51	93.2	17.2	0.64
	Double bootstrap method	90.1	12.8	0.46	89.5	16.6	0.49
Bootstrap method for regression coefficient	**Percentile method**	84.3	15.2	0.39	84.3	18.8	0.47
	Double bootstrap method	87.2	12.9	0.26	83.9	17.9	0.40

6. Concluding Remarks

To deal with non-linear statistics (*eg.*, correlation coefficient and regression coefficient) we often employ Bootstrap technique. Concentrating on Rao and Wu's rescaling bootstrap and Sitter's Mirror match bootstrap for unequal probability sampling, we need new developments for some practical situations dealing with sensitive issues. Randomized response aims to reduce false responses on sensitive issues. There is a rich growth of literature on Randomized Response (RR) procedure. In this paper we try to modify Rescaling bootstrap technique and Mirror match technique for the situations where RR surveys are needed.

From our simulation study we may conclude that Double bootstrap method gives better ACV's and the lengths of the CI's and a desirable magnitude of the ACP values. Our result

shows that any of the schemes mentioned above may be used for our bootstrap method. For both of the schemes the desired requirements are fulfilled.

References

Bar-Lev, S.K., Bobovitch, E. and Boukai, B.(2004). A note on randomized response models for quantitative data. *Metrika*, **60**, 255-260.

Chaudhuri A (2001).Using a 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*. Chapman and Hall, CRC Press, Taylor & Francis Group, Boca Raton, FL.

Eichhorn B.H. and Hayre, L.S. (1983). Scrambled RR method for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, **7**, 307-316.

Fox, J.A. and Tracy, P.E. (1986). *Randomized Response: A Method for Sensitive Surveys*. SAGE Publications.

Greenberg B.G., Abul-Ela A.A., Simmons,W.R., Horvitz. D.G. (1969) The unrelated question randomized response model: theoretical framework. *Journal of American Statistical Association,* **64**, 520–539 352

Gupta, S., Gupta, B. and Singh, S.(2002). Estimation of sensitivity level of personal interview survey question. *Journal of Statistical Planning and Inference,* **100**, 239-247.

Hajek, J. (1958). Some contributions to the theory of probability sampling. *Bulletin of International Statistical Institute,* **36(3),**127-134.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics,* **33**, 350-374

Hedayat, A.S. and Sinha, B.K. (1991) *Design and Inference in Finite Population Sampling*. Wiley, New York 356

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

Horvitz., D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association***, 47**, 663-685.

Padmawar, V.R. and Vijayan, K. (2000). Randomized response revisited. *Journal of Statistical Planning and Inference,* **90,** 293–304.

Pal, S. (2008). Unbiasedly estimating the total of a stigmatizing variable from a complex survey on permitting options for direct or randomized responses, *Statistical Papers,* **49(2),** 157-164.

Pollock, K. H. and Bek, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of American Statistical Association***,71 (356**), 884-886.

Rao, J.N.K. and Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of Royal Statistical Society,***B24,** 482-491.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of American Statistical Association*, **83,** 231-241.

Singh, S., Joarder, A,H. and King, M.L. (1996). Regression analysis using scrambled responses, *Australian Journal Statistics,* **38(2**), 201-211.

Sitter, R. (1992). A re-sampling procedure for complex survey data. *Journal of American Statistical Association*, **87,** 755-765.

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