J. Ind. Soc. Agril. Statist.
58(2), 2004 : 190-211

Utilizing Covariates by Logistic Regression Modelling in Improved Estimation of Population Proportions Bearing Stigmatizing Features through Randomized Responses in Complex Surveys*

Arijit Chaudhuri and Amitava Saha[1]
Indian Statistical Institute, Kolkata
(Received : February, 2003)

SUMMARY

The problem addressed here is one of estimating the proportion of people in a community possibly bearing a stigmatizing characteristic, which is one of the items in a multi-purpose large-scale survey in which as usual the sample is chosen with unequal selection-probabilities for the units of a finite population. Generating truthful data on a sensitive item is difficult. Warner (1965) with his pioneering Randomized Response (RR) technique (RRT) provided a well-known solution, protecting a respondent's privacy. But his selection procedure was restricted exclusively to simple random sampling with replacement (SRSWR) – a scheme rarely executed in practice. Numerous alternative RRT's confined to SRSWR emerged with higher accuracy and better protection of privacy. Maddala (1983), Sheers and Dayton (1988), Kerkvliet (1994), Heijden and Gils (1996) among others achieved improved efficiency on utilizing covariates employing appropriate logistic regression modelling, again restricting to SRSWR. Chaudhuri (2001a, b, 2002) extended a few well-known RRT's to cover 'unequal probability sampling'. Here is illustrated how the logistic regression modelling, in case of 'unequal probability sampling' does not apply, for example, to Warner's RRT, but may be profitably employed with a couple of other well-known RRT's. Nayak's (1994) treatment of 'protection of privacy' with RRT's confined to SRSWR is extended here to complex survey sampling schemes. Additional algebraic complexity is inevitable in extension beyond SRSWR, as is needless to mention.

Key words : Covariates, Logistic regression, Protection of privacy, Randomized response, Sensitive issues, Varying probability sampling.

---

1 DGMS, Dhanbad, Jharkhand

* This research is partially supported by CSIR Grant No. 21(0539)/02/EMR-II. The opinions expressed are of the authors, not of the organizations they work for.

1. Introduction

Suppose we consider a community of a given number of people and our interest is to survey a sample to obtain an estimate of the proportion out of them bearing a specific socially disapproved characteristic. Examples of such characteristics are habits of drunken driving, alcoholism, tax evasion, experiences of sexual abuses, induced abortion, testing H.I.V. positive, exercising economic fraud, drug abuse among others. In practice, it is hard to gather honest information by direct interrogation on such personal and sensitive issues. But a pioneering device of generating 'Randomized Responses' (RR) rather than 'Direct Responses' (DR) was introduced by Warner (1965) as a means of protecting the privacy of a respondent and simultaneously deriving a serviceable estimator for the parameter intended. Numerous modifications on it have emerged by way of effecting improvements as reviewed in the text by Chaudhuri and Mukerjee (1988). Most of these RR devices are applied with the restriction that the sample is selected by simple random sampling with replacement (SRSWR) method alone. Chaudhuri (2001a, b, 2002) showed how to extend the application covering unequal probability sampling. Maddala (1983) described how a logistic regression model connecting the parameter to be estimated and the values available on certain covariates possibly governing the incidence of the abominable feature in a respondent may be fitted for an improved estimation. Sheers and Dayton (1988), Kerkvliet (1994), Heijden and Gils (1996) among others pursued with this approach as well. In this paper we illustrate how to extend this covering unequal probability sampling. A specimen of a simulation exercise is presented to notice how one may acquire additional efficacy in estimation. This extension is necessary in practice as samples are usually chosen without replacement with unequal probabilities.

Selection probability of course cannot affect an individual's privacy. This is because, our approach is to attach an indicator function to each, estimate the population total of this, allowing 'a value 1' if the respondent 'bears' the sensitive feature or '0' if not, retaining the option of assigning these two values to the contrary case of 'not bearing it' as well. So, a respondent cannot reasonably suspect he/she is selected because the investigator guesses his/her value is '1', rather than '0', nor that the total number of persons bearing the sensitive item in the community is 'high' rather than 'low'. In estimating this total and the corresponding proportion, the most important point intended to be emphasized here is that the theory of tackling the RRT's accomplished with SRSWR does not readily carry over when the respondents are selected with unequal probabilities, as they usually are, as a rule, in large-scale surveys. Executing an SRSWR survey to estimate a few sensitive proportions alone can at most be a mere luxury and so, may not be sponsored in practice.

In Section 2 we narrate four RR devices and their applicability in (I) SRSWR and with (II) varying probability sampling. Also we add a note on the 'theme of protection of privacy' in the two situations demanding separate

treatments. Of course, numerous other RRT's exist yielding competitive results. But here we modestly illustrate only these four, which are simple and adequately interesting. More sophisticated ones are avoided to keep our illustrations simple. In Sections 3 and 4 respectively we present details about fitting logistic regression models in SRSWR and general sampling schemes. Section 5 gives numerical evidences illustrating possibilities for improved estimation. An Appendix finally gives estimates of accuracies of estimates of regression parameters.

2. Four RR Devices and Estimation

Let $U = (1, \cdots, i, \cdots, N)$ denote a population of N people and y_i, the value of a variable defined for i in U such that $y_i = 1$ if the person labeled i bears a stigmatizing characteristic A and = 0 if i bears the complement of A which is A^C. Writing $Y = \sum y_i$, $\pi_A = \frac{Y}{N}$ is the proportion of people in the community bearing A, by $\sum$ we mean $\sum_{i=1}^{N}$.

The problem is to suitably estimate π_A on surveying a sample or to estimate Y treating N as known. If a person is chosen at random from U, then π_A may be taken as the "probability" that a person sampled bears A and this remains so if the selection at random is continued 'with replacement'. So, assuming SRSWR the following four RR devices apply in unbiasedly estimating this probability π_A.

(i) Warner's (1965) RR Device

Every sampled person is given a device to observe one 'outcome' with a probability p or its 'complement' with the probability $(1-p)$ and is to (a) say "yes" or "no" according as the outcome 'matches' or 'mis-matches' his/her characteristic A or A^C (b) without divulging to the interviewer the outcome observed.

(ii) A 'Forced Response' Method

A sampled person is given a device which may yield three possible outcomes O_1, O_2 or O_3, say, with respective probabilities p_1, p_2 and $(1 - p_1 - p_2)$ and corresponding instructions for him/her is to respond "yes", "no", and "the honest reporting" of yes or no about bearing A or A^C, in each case of course not giving out the outcomes to the interviewer.

(iii) Kuk's (1990) Method

A sampled person is given two boxes marked respectively A and A^C containing respectively 'Red' and 'Black' cards in proportions "p_1: $(1 - p_1)$" and

"p_2: $(1 - p_2)$" with an instruction to draw from the box, marked matching his/her true "A or A^C – trait", k cards and report the number of 'Red' cards obtained.

(iv) Unrelated Question Model

Horvitz *et al.* (1967), Greenberg *et al.* (1969) and Greenberg *et al.* (1977) have given essentially the following device. In one SRSWR in n_1 draws every person is to use a randomizing device to truthfully say "yes" or "no" with probability p_1 about bearing A or A^C and with probability $(1 - p_1)$ about bearing an unrelated innocuous characteristic B or its complement B^C. Every person in another 'independent' SRSWR in n_2 draws is to act similarly with the difference in the device that the probabilities change to p_2 and $(1 - p_2)$, $p_1 \neq p_2$.

The RR device parameters p, k, p_1, p_2 are all pre-specified. From the literature it is known how to unbiasedly estimate π_A and obtain the variance of the estimate and unbiasedly estimate the latter in each of the above four cases (i) – (iv).

Chaudhuri (2001a, b, 2002) permitted unbiased estimation of π_A when a sample, say, s may be chosen with certain selection probabilities p(s), say, admitting certain general features but permitting selection of persons with or without replacement with equal or unequal probabilities. For the Warner's RR device (i) let for a person i no matter how selected.

$I_i = 1$ if the 'outcome type' matches ith person's attribute A or A^C

$= 0$, otherwise

Then, Prob $[I_i = y_i] = p$, Prob$[I_i = 1 - y_i] = 1 - p$ for $y_i = 1$ or 0 as earlier specified. We shall throughout use the generic notation E_R, V_R to denote expectation and variance operators with respect to any RR device.

Then

$$E_R(I_i) = py_i + (1-p)(1-y_i) = (1-p) + (2p-1)y_i$$

Taking $p \neq \frac{1}{2}$, $r_i = \frac{I_i - (1-p)}{(2p-1)}$ is an unbiased estimator for y_i noting that $E_R(r_i) = y_i, \forall i \in U$.

Also

$$V_R(I_i) = E_R(I_i)(1 - E_R(I_i)) = p(1-p)$$

leading to $V_R(r_i) = \frac{V_R(I_i)}{(2p-1)^2} = \frac{p(1-p)}{(2p-1)^2} = V_i$, say, $i \in U$

For the Forced Response Scheme in (ii) for any i in U let

$I_i = 1$ if i responds 'Yes'

$= 0$ if i responds 'No'

Then

Prob $[I_i = 1, y_i = 1] = p_1 + (1 - p_1 - p_2) = 1 - p_2$

Prob $[I_i = 1, y_i = 0] = p_1$, Prob $[I_i = 0, y_i = 1] = p_2$

Prob $[I_i = 0, y_i = 0] = p_2 + (1 - p_1 - p_2) = 1 - p_1$

Hence

$$\text{Prob}[I_i = y_i] = p_1 y_i + (1 - p_1 - p_2)y_i + p_2(1 - y_i) + (1 - p_1 - p_2)(1 - y_i)$$
$$= 1 - p_1 + (p_1 - p_2)y_i$$

$$\text{Prob}[I_i = 1 - y_i] = p_1(1 - y_i) + p_2 y_i$$

$E_R(I_i) = p_1 + (1 - p_1 - p_2)y_i$, noting that $y_i^2 = y_i$ and

$$V_R(I_i) = E_R(I_i)(1 - E_R(I_i)) = p_1(1 - p_1) + (p_1 - p_2)(1 - p_1 - p_2)y_i$$
$$= p_1(1 - p_1) \quad \text{if } y_i = 0$$
$$= p_2(1 - p_2) \quad \text{if } y_i = 1$$

Then, $r_i = \dfrac{I_i - p_1}{1 - p_1 - p_2}$ is unbiased for y_i with a variance

$V_i = V_R(I_i)/(1 - p_1 - p_2)^2$ with $p_1 + p_2 \neq 1$

For Kuk's (1990) model (iii) let f_i be number of red cards reported to have been drawn by the person labeled i in U. Then

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

$$V_R(f_i) = k[p_1(1 - p_1)y_i + p_2(1 - p_2)(1 - y_i)]$$

noting that f_i follows binomial distribution with parameters k and "p_1 if i bears A" or "p_2 if i bears A^{C}". Then, $r_i = \left(\dfrac{f_i}{k} - p_2\right)\Big/(p_1 - p_2)$, taking $p_1 \neq p_2$, is an unbiased estimator of y_i, $i \in U$. Also, $V_R(r_i) = a_i y_i + b_i = V_i$, say, with $a_i = \dfrac{1 - p_1 - p_2}{k(p_1 - p_2)^2}$, $b_i = \dfrac{p_2(1 - p_2)}{k(p_1 - p_2)^2}$ leading to $V_i = \dfrac{p_1(1 - p_1)}{k(p_1 - p_2)^2}$ if $y_i = 1$ and $V_i = \dfrac{p_2(1 - p_2)}{k(p_1 - p_2)^2}$ if $y_i = 0$. For both (ii) and (iii) an unbiased estimator v_i for V_i is obtained on replacing y_i in it by r_i, $i \in U$. It may be carefully noted that it is desirable to take "k as large as possible subject to a respondent's willingness to repeat the drawings from the box" in order to reduce $V_i, i \in U$.

With the approach of Chaudhuri (2001a, b, 2002), to apply the (iv) unrelated question model to a general sampling scheme one may do only with a single sample. But each sampled person i is required to make four independent draws as follows leading to the generation of the following 'indicator' random variables.

$I_i = 1$ if i draws from a box a card marked A or B in proportions $p_1 : (1-p_1)$ and the card type 'matches' his/her true characteristic and $I_i = 0$ if there is no match

I_i' is the variable generated independently in a manner identically as $I_i, i \in U$.

J_i is a variable generated as I_i with the difference that the cards marked A or B are in proportions $p_2 : (1-p_2)$ instead of $p_1 : (1-p_1)$; moreover J_i' is another variable generated independently and identically as $J_i, i \in U$.

Letting $t_i = 1$ if i bears B

$= 0$ if i bears B^C, one gets

$$E_R(I_i) = p_1 y_i + (1-p_1)t_i = E_R(I_i'), \qquad E_R(J_i) = p_2 y_i + (1-p_2)t_i = E_R(J_i')$$

Taking $p_1 \neq p_2$, let

$$r_i' = \frac{(1-p_2)I_i - (1-p_1)J_i}{(p_1 - p_2)}, \quad r_i'' = \frac{(1-p_2)I_i' - (1-p_1)J_i'}{(p_1 - p_2)}$$

Then, $E_R(r_i') = y_i = E_R(r_i''), i \in U$. Also, $r_i = \frac{1}{2}(r_i' + r_i'')$ is an unbiased estimator of y_i and $v_i = \frac{1}{4}(r_i' - r_i'')^2$ is an unbiased estimator of $V_i = V_R(r_i), i \in U$.

The advantages of estimating y_i by r_i and deriving the variance V_i as illustrated above relating to (i) – (iv), and may be in general also will be revealed briefly in what follows.

Let s be a sample chosen from U with a probability p(s) according to a design P and E_p, V_p be operators for expectation, variance with respect to P. Let the over-all expectation, variance operators be E and V respectively such that

$$E = E_p E_R = E_R E_p \text{ and } V = E_p V_R + V_p E_R = E_R V_p + V_R E_p$$

Let $t = \sum y_i b_{si} I_{si}$, $I_{si} = 1$ if $i \in s$, $I_{si} = 0$ if $i \notin s$, b_{si} be constants free of $Y = (y_1, \cdots\cdots, y_i, \cdots\cdots y_N)$ such that $E_p(b_{si} I_{si}) = 1, \forall i \in U$. The literature on survey sampling abounds with numerous examples of P and b_{si}. One example of b_{si} is $\frac{1}{\pi_i}$ with $\pi_i = \sum_{s \ni i} p(s) = \sum_s p(s) I_{si}$ assuming $\pi_i > 0$ for every $i \in U$ - a condition known to be necessary for the existence of an unbiased estimator of Y. For the above t we have $E_p(t) = Y$.

$$\text{Also, } V_p(t) = \sum y_i^2 c_i + \sum_{i \neq j} y_i y_j c_{ij}$$

where $c_i = E_p(b_{si}^2 I_{si}) - 1$, $c_{ij} = E_p(b_{si} b_{sj} I_{sij}) - 1$ writing $I_{sij} = I_{si} I_{sj}$. Let there exist constants d_{si}, d_{sij} free of Y such that $E_p(d_{si} I_{si}) = c_i$, $E_p(d_{sij} I_{sij}) = c_{ij}$ leading to the existence of $v_p(t) = \sum y_i^2 d_{si} I_{si} + \sum_{i \neq j} y_i y_j d_{sij} I_{sij} = v_p$, say, satisfying $E_p(v_p) = V_p(t)$.

The literature on survey sampling is full of examples of such d_{si}, d_{sij}'s. For example, if $b_{si} = \frac{1}{\pi_i}$, then $c_i = \frac{1-\pi_i}{\pi_i}$ suggesting $d_{si} = \frac{1}{\pi_i}\left(\frac{1-\pi_i}{\pi_i}\right)$, $c_{ij} = \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$, writing $\pi_{ij} = \sum_s p(s) I_{sij}$, $d_{sij} = \frac{c_{ij}}{\pi_{ij}}$ if $\pi_{ij} > 0$.

Turning to the main topic here with y_i's generally unavailable it is a simple step to employ

$$e = \sum r_i b_{si} I_{si}$$

as an unbiased estimator for Y because $E_R(e) = t$, $E_p(e) = \sum r_i$ $E(e) = E_p E_R(e) = E_p(t) = Y$ and also $E(e) = E_R E_p(e) = E_R(\sum r_i) = \sum y_i = Y$

$$\text{Now } V(e) = E_p V_R(e) + V_p E_R(e) = E_p\left[\sum V_i b_{si}^2 I_{si}\right] + V_p(t)$$

$$= E_p\left[\sum V_i b_{si}^2 I_{si}\right] + \sum y_i^2 c_i + \sum_{i \neq j} y_i y_j c_{ij}$$

$$\text{Writing } v_p(e) = v_p(t)\Big|_{Y-R}, \text{ where } R = (r_1, \cdots\cdots, r_N)$$

$$= \sum r_i^2 d_{si} I_{si} + \sum_{i \neq j} r_i r_j d_{sij} I_{sij}$$

it follows that $E_R v_p(e) = v_p(t) + \sum V_i d_{si} I_{si}$

$$\text{So, one may take } v_1(e) = v_p(e) + \sum v_i (b_{si}^2 - d_{si}) I_{si} \qquad (2.1)$$

as an unbiased estimator of V(e) because one may check that

$$E_R v_1(e) = v_p(t) + \sum V_i d_{si} I_{si}$$

and hence that $E v_1(e) = V(e)$

Also, $V(e) = E_R V_p(e) + V_R E_p(e) = E_R E_p v_p(e) + \sum V_i$ and hence

$$v_2(e) = v_p(e) + \sum v_i b_{si} I_{si} \tag{2.2}$$

is another possible unbiased estimator of V(e) because obviously

$$E\ v_2(e) = E_R E_p v_2(e) = E_R E_p v_p(e) + E_R \sum v_i = V(e)$$

Incidentally, if V_i for an RR device like Warner's (1965), be known then v_i in (2.1) - (2.2) should be replaced by $V_i, i \in s$.

In order to demonstrate how an RR as 'yes' or 'no' reveals the respondent's true characteristic A or A^C, writing

Prob (Yes| A) = a and Prob (No| A^C) = b

for any SRSWR-based RRT, Nayak (1994) observes the revealing posterior probabilities as

$$\text{Prob}(A \mid \text{Yes}) = \frac{a\theta_A}{a\theta_A + (1-b)(1-\theta_A)} \text{ and}$$

$$\text{Prob}(A \mid \text{No}) = \frac{(1-a)\theta_A}{(1-a)\theta_A + b(1-\theta_A)}$$

Then, he discusses how the parameters 'a and b' should be chosen in the lights of these to keep the 'revealing probabilities' close to θ_A and variance of an estimator for θ_A down. These two equations cease to be valid with unequal probability sampling.

Turning to the question of protection of privacy in case of unequal probability sampling with our formulation for tackling RRT's we need to revise Nayak's (1994) Bayesian approach in the following way.

Pretending that a postulated 'prior' probability $L_i (0 \le L_i \le 1)$ which is Prob(y_i = 1), is assigned to i, we note the conditional probabilities $L_i(C)$, namely the 'posterior' probability Prob(y_i = 1| I_i = 1) for the four illustrated RRT's as follows.

(I) For Warner's RRT

$$L_i(C) = \frac{L_i \text{Prob}(I_i = 1 \mid y_i = 1)}{L_i \text{Prob}(I_i = 1 \mid y_i = 1) + (1 - L_i)\text{Prob}(I_i = 1 \mid y_i = 0)_i}$$

$$= \frac{pL_i}{(1-p) + (2p-1)L_i} \to L_i \text{ as } p \to \frac{1}{2}$$

Also, $V_i \to \infty$ as $p \to \frac{1}{2}$

(II) For Kuk's RRT

$$L_i(C) = \text{Prob}(y_i = 1 \mid RR = f_i)$$

$$= \frac{L_i \,\text{Prob}(RR = f_i \mid y_i = 1)}{L_i \,\text{Prob}(RR = f_i \mid y_i = 1) + (1 - L_i)\,\text{Prob}(RR = f_i \mid y_i = 0)}$$

$$= \frac{L_i\left[\theta_1^{f_i}(1-\theta_1)^{k-f_i}\right]}{\theta_2^{f_i}(1-\theta_2)^{k-f_i} + L_i\left[\theta_1^{f_i}(1-\theta_1)^{k-f_i} - \theta_2^{f_i}(1-\theta_2)^{k-f_i}\right]} \to L_i$$

as $\theta_1 \to \theta_2$

Also, $V_i \to \infty$ as $\theta_1 \to \theta_2$.

(III) For Forced Response RRT

$$L_i(C) = \text{Prob}(y_i = 1 \mid I_i = 1)$$

$$= \frac{L_i \,\text{Prob}(I_i = 1 \mid y_i = 1)}{L_i \,\text{Prob}(I_i = 1 \mid y_i = 1) + (1 - L_i)\,\text{Prob}(I_i = 1 \mid y_i = 0)}$$

$$= \frac{(1-p_2)L_i}{p_1 + (1-p_1-p_2)L_i} \to L_i \text{ as } p_1 \to (1-p_2)$$

and also, $V_i \to \infty$ as $p_1 \to p_2$

In all the cases (I) – (III), we may write, formally

$$L_i(C) = \frac{L_i\phi_i}{\psi_i + L_i(\phi_i - \psi_i)} \to L_i \text{ as } \phi_i \to \psi_i \text{ so that 'protection of privacy'}$$

and 'preservation of accuracy' move in opposite directions. Also

(IV) For Unrelated Question Model

$$L_i(C) = \text{Prob}(y_i = 1 \mid I_i = 1, J_i = 1)$$

$$= \frac{L_i \,\text{Prob}(I_i = 1 \mid y_i = 1) \times \text{Prob}(J_i = 1 \mid y_i = 1)}{L_i \,\text{Prob}(I_i = 1 \mid y_i = 1) \times \text{Prob}(J_i = 1 \mid y_i = 1) + (1 - L_i)\,\text{Prob}(I_i = 1 \mid y_i = 0) \times \text{Prob}(J_i = 1 \mid y_i = 0)}$$

$$= \frac{L_i\phi_i}{\psi_i + L_i(\phi_i - \psi_i)} \to L_i \text{ as } p_1 \to p_2$$

and also, $V_i \to \infty$ as $p_1 \to p_2$ as in (I – III) too.

3. Use of Covariates in Improved Estimation and Logistic Regression Modelling

If the propensity to bearing a socially blameworthy characteristic may understandably be supposed to vary among the people in the community of

interest in a given investigation according to their other special features and the people may be classified by characteristics in discernible ways then the above estimators for π_A may be modified hopefully with higher accuracies. The characteristics may be distinguished in terms of real and/or dummy variables like, age, income, housing and martial conditions, educational levels and proficiency in the local languages etc. Following is the pioneering work of Maddala (1983) in the present context of dealing with sensitive issues, followed by subsequent contributors including Sheers and Dayton (1988), Kerkvliet (1994), Heijden and Gils (1996), each concerned with estimation of π_A by RR-based observations from SRSWR's.

Suppose $x_1, \cdots\cdots, x_k$ are $k(\geq 1)$ covariates for which observations $x_{ji}, (j = 1, \cdots\cdots, k)$ are available on the individuals i in a sample s which is selected by SRSWR. Then, we may write $\pi_A(x)$ as the probability that a person chosen at random from the given community bearing specific combination of values of $x = (x_1, \cdots\cdots, x_k)$ also bears the sensitive attribute A.

Then, since $0 < \pi_A(x) < 1$, in situations of interest we may consider the logit transform of $\pi_A(x)$, namely

$$L(x) = \log_e\left(\frac{\pi_A(x)}{1-\pi_A(x)}\right)$$

A next step is to postulate a linear model for $L(x)$ as, say

$$L(x) = \alpha(x) + \beta(x)$$

Then applying the principle of least squares, estimates $\hat{\alpha}(x)$, for $\alpha(x)$ and $\hat{\beta}(x)$ for $\beta(x)$ may easily be derived leading to an estimate of $\pi_A(x)$ in terms of $\hat{\alpha}(x)$ and $\hat{\beta}(x)$ by an appropriate transformation. Estimation of the variance or MSE of the resulting estimate, say, $\hat{\pi}_A(x)$ need not be easy or trivial.

Of course, we admit that we are not contributing here any new concepts adding to Maddala's (1983). Our sole purpose is to show how his logistic regression approach, confined to SRSWR, when extended to complex survey sampling, fails with Warner's RRT but applies profitably with a couple of others, demanding a sizeable manipulation in algebra, of course, with our formulation.

In Section 4 we pursue with this issue when RR's are derived by the procedures (i) - (iv) from persons selected by general sampling schemes considered in Section 2.

4. Logistic Regression Modelling in Improved Estimation of Proportions from RR in Complex Surveys

When RR's are gathered through complex surveys, i.e. in surveys with unequal probabilities of selection, the approach of Section 3 in estimation of $\pi_A(x)$ is not applicable but certain modifications narrated below may be successfully tried. For simplicity we shall restrict below to the case when only one covariate x is available with known values x_i for i in U.

Let us start with Kuk's (1990) RR procedure (iii). For this we have already defined r_i for i in U. Let us try to develop a suitable transform of r_i and relate that transform to x_i, $i \in U$ in a suitable way.

Let $u_i = r_i(p_1 - p_2) + p_2$ for i in U and r_i as defined for (iii). Then, $0 \le u_i \le 1, \forall i \in U$. So we take $k > 1$ to ensure that $0 < u_i < 1$ at least for certain i in U. Then, let us define

$$z_i = \log_e\left(\frac{u_i}{1-u_i}\right) \text{ for } 0 < u_i < 1, \text{ for every i in a subset, say, } \overline{U} \text{ of U}$$

Let us postulate the logistic regression model connecting u_i with x_i and write

$$z_i = \alpha + \beta x_i + \varepsilon_i, i \in \overline{U}$$

with α, β as unknown constants and ε_i's as random errors. In what follows wherever z_i is used it will be understood that z_i is to be calculated only for i in $\overline{U}$.

$$\text{Let } S = \sum (z_i - \alpha - \beta x_i)^2 b_{si} I_{si} \text{ with } E_p(b_{si} I_{si}) = 1, \ \forall i \in \overline{U}$$

$$\text{Solving } 0 = \frac{\partial S}{\partial \alpha} \text{ and } 0 = \frac{\partial S}{\partial \beta} \text{ for } \alpha \text{ and } \beta \text{ we get}$$

$$\hat{\alpha} = \frac{\sum (z_i - \hat{\beta} x_i) b_{si} I_{si}}{\sum b_{si} I_{si}}, \quad \hat{\beta} = \frac{\sum \left(z_i - \frac{\sum z_i b_{si} I_{si}}{\sum b_{si} I_{si}} \right) \left(x_i - \frac{\sum x_i b_{si} I_{si}}{\sum b_{si} I_{si}} \right)}{\sum \left(x_i - \frac{\sum x_i b_{si} I_{si}}{\sum b_{si} I_{si}} \right)^2}$$

as the least squares estimates of α and β. In order to estimate π_A using sample-based RR's and x_i-values we may proceed as follows.

Let $\hat{z}_i = \hat{\alpha} + \hat{\beta}x_i$, $\hat{u}_i = \frac{1}{1+e^{-\hat{z}_i}}$, $\hat{r}_i = (\hat{u}_i - p_2)/(p_1 - p_2)$ for $i \in s$, using r_i, u_i for $i \in s \cap \bar{U}$. Then our proposed revised estimator for π_A based on Kuk's procedure using the RR's and the covariate values is

$$\hat{e} = \sum \hat{r}_i b_{si} I_{si}$$

Though $E_R(r_i) = y_i$, $E_R(\hat{r}_i) \neq y_i$ and so $\hat{e}$ is a biased estimator for Y. This problem persists about the estimators of $\pi_A(x)$ for the approaches of Maddala (1983) and his followers as well.

In estimating the MSE of $\hat{e}$ about Y our recommended procedure is the following. We may note that

$$V_R(u_i) = (p_1 - p_2)^2 V_R(r_i) = (p_1 - p_2)^2 V_i, i \in U \text{ ; for } z_i = \log_e\left(\frac{u_i}{1-u_i}\right)$$

$$V_R(z_i) \cong \left(\frac{\partial z_i}{\partial u_i}\right)^2\Bigg|_{u_i = E_R(u_i)} V_R(u_i) = \frac{(p_1 - p_2)^2 V_i}{[E_R(u_i)(1 - E_R(u_i))]^2}$$

$$= \frac{(p_1 - p_2)^2 V_i}{[p_2(1-p_2)]^2 + y_i(p_1 - p_2)(1 - p_1 - p_2)[p_1(1-p_1) + p_2(1-p_2)]} = \sigma_i^2, \quad \text{say}$$

$= \frac{1}{kp_2(1-p_2)}$ if $y_i = 0$ though unknown in general since y_i though only 0 or 1 is really unknown. But σ_i^2 may be estimated by $\hat{\sigma}_i^2$ obtained on substituting r_i for y_i in the former. Now, writing

$$\hat{z}_i = \sum z_i b_{si} I_{si} - \sum b_{si} I_{si} + \left(x_i - \sum x_i b_{si} I_{si}\right) \frac{\sum \left(z_i - \sum z_i b_{si} I_{si}\right)\left(x_i - \sum x_i b_{si} I_{si}\right)}{\sum \left(x_i - \sum x_i b_{si} I_{si}\right)^2 b_{si} I_{si}}$$

and remembering that z_i's are independent across i, we may write down the formula for $V_R(\hat{z}_i) = W_i$ as a function of σ_i^2 with known coefficients. So, replacing σ_i^2 by $\hat{\sigma}_i^2$ in W_i the latter may be estimated by $\hat{W}_i$, say. So, an estimator $\hat{V}_R(\hat{u}_i)$ may be worked out for $V_R(\hat{u}_i)$ from

$$\hat{\sigma}_i^2 = \frac{1}{\hat{u}_i(1-\hat{u}_i)} \hat{V}_R(\hat{u}_i)$$

Hence $\hat{V}_R(\hat{r}_i)$ as an estimator for $V_R(\hat{r}_i)$ may be worked out from

$$\hat{V}_R(\hat{u}_i) = (p_1 - p_2)^2 \hat{V}_R(\hat{r}_i)$$

Writing $\hat{v}_i$ for $\hat{V}_R(\hat{r}_i)$ two estimators for the MSE of $\hat{e}$ may be proposed as the estimators $\hat{v}_j(\hat{e})$ obtained from $v_j(e)$, $j = 1, 2$ on replacing v_i in the latter by $\hat{v}_i$ and r_i by $\hat{r}_i$, $i \in s$.

In order to apply the logistic regression modelling, again with a single covariate, to the 'unrelated question model' (iv), let us define

$$u_i = \frac{(p_1 - p_2)}{2 - (p_1 + p_2)}\left(r_i + \frac{1 - p_i}{p_1 - p_2}\right), p_1 \neq p_2, i \in U$$

Since for (iv), $r_i = \frac{1}{2}\left[(1 - p_2)(I_i + I_i') - (1 - p_1)(J_i + J_i')\right]/(p_1 - p_2)$

it follows that

$$\min\ r_i = -\frac{1 - p_1}{p_1 - p_2} \text{ when } I_i = 0 = I_i', J_i = 1 = J_i' \text{ and}$$

$$\max\ r_i = \frac{1 - p_1}{p_1 - p_2} \text{ when } I_i = 1 = I_i', J_i = 0 = J_i', i \in U$$

Letting $u_i = \frac{(p_1 - p_2)}{2 - (p_1 + p_2)}\left(r_i + \frac{1 - p_1}{p_1 - p_2}\right)$ it follows that $0 \leq u_i \leq 1, \forall i \in U$.

It may be noted that $u_i = 0$ when r_i equals min r_i, $u_i = 1$ when r_i equals max r_i but $0 < u_i < 1$ for i in the remaining set $\overline{U}$, say, which is a subset of U.

For $i \in \overline{U}$, then as in the case (iii) we may take $z_i = \log_e\left(\frac{u_i}{1 - u_i}\right)$ and postulate the logistic regression model $z_i = \alpha + \beta x_i + \varepsilon_i$ as in (iii) and proceed accordingly.

Unfortunately, however, for Warner's (1965) model (i) and the Forced response model (ii) the above approach of fitting a logistic regression model is not applicable when RR's are obtained from a sample chosen with unequal probabilities of selection. To see this let us observe as follows

For Warner's case (i), $r_i = \frac{I_i - (1 - p)}{(2p - 1)}, p \neq \frac{1}{2}$

Let $u_i = (2p - 1)r_i + (1 - p), i \in U$. Then, only two possible values of r_i are $\min r_i = -\frac{1 - p}{2p - 1}$ and $\max r_i = \frac{p}{2p - 1}$ when $I_i = 0$ and $I_i = 1$ respectively so that

only two possible values of u_i are 0 and 1. So, a logit transform $z_i = \log_e\left(\frac{u_i}{1-u_i}\right)$ is not available and we can not proceed further.

For the Forced Response RR device (ii), we have $r_i = \frac{I_i - p_1}{1-p_1-p_2}$ which equals $-\frac{p_1}{1-p_1-p_2}$ when $I_i = 0$ and $\frac{p_1}{1-p_1-p_2}$ when $I_i = 1$ so that if we define $u_i = \left(r_i + \frac{p_1}{1-p_1-p_2}\right)(1-p_1-p_2)$, $u_i = 0$ if $I_i = 0$ and $u_i = 1$ when $I_i = 1$ and u_i can take no other values. So, a logit transform of u_i is not available and we cannot proceed further to apply a logistic regression model to utilize a covariate.

In Section 5 we present some results of our numerical exercises to illustrate for Kuk's RR device (iii) and the unrelated question model (iv) to show some comparative performances of estimators of π_A based on samples chosen with varying probabilities without replacement (1) without using a covariate (2) as well as with the use of one covariate. In an Appendix we present methods of estimating the MSE's of $\hat{\alpha}$ and $\hat{\beta}$ supposing this issue to be relatively less interesting to a reader.

5. Numerical Evidences of Accuracies in Estimation

We consider a fictitious community of N = 113 people with the known respective values y_i, t_i, a_i and x_i for $i = 1, \cdots\cdots, N$. Here $y_i = 1$ if i^{th} person has clandestine sources of income and/or incurs expenses on items he/she keeps secret from everybody else and $y_i = 0$ else

$t_i = 1$ if i prefers cricket to football and $t_i = 0$ else

a_i = the number of people in the household to which i belongs – this is the size-measure used in choosing a sample of households of which one particular member i is asked to 'respond' queries about y_i, t_i, and x_i for i in a sample s chosen from all the N = 113 households

x_i = the per capita expenditure in Indian Rupees incurred in the household to which i belongs – this expenditure is taken as the single covariate on which the sensitive variable y (with its values as y_i above) is regressed

The problem addressed is to estimate $Y = \sum_{i=1}^{N} y_i$

Table 1. Giving values relevant to the population of N = 113 people

Unit (i) serial numbers	y_i	t_i	a_i	x_i
(1)	(2)	(3)	(4)	(5)
1	1	1	9	2891.31
2	1	1	10	4261.13
3	1	0	1	2262.45
4	1	1	7	2530.20
5	1	1	8	2430.49
6	1	1	2	4226.83
7	1	0	11	3270.41
8	1	1	6	1179.95
9	1	1	6	1902.73
10	0	0	4	1482.09
11	1	1	3	1480.36
12	0	1	5	250.90
13	1	0	9	2255.33
14	0	1	4	2525.85
15	1	1	7	1241.19
16	1	0	9	1256.66
17	1	1	6	2194.89
18	1	1	3	3187.48
19	0	1	5	193.65
20	1	0	8	1669.54
21	1	1	7	3074.11
22	1	1	1	4187.81
23	1	0	1	1264.92
24	1	1	2	3196.59
25	1	1	1	3354.57
26	1	1	7	2717.12
27	1	1	3	2927.63
28	1	0	3	4147.14
29	1	1	10	3355.06
30	1	1	9	2644.63
31	1	0	3	2495.64
32	1	1	9	4400.64
33	1	1	3	3284.96
34	1	1	1	1334.98
35	1	0	3	1408.34
36	1	1	2	241.83
37	1	1	3	4649.75
38	1	1	2	2243.53
39	1	0	3	1120.97
40	1	0	3	1296.67
41	1	1	3	2878.00
42	1	1	7	1268.51
43	1	0	6	1258.95
44	1	1	2	2990.47

Unit (i) serial numbers	y_i	t_i	a_i	x_i
(1)	(2)	(3)	(4)	(5)
45	1	0	2	1299.93
46	0	1	4	205.55
47	1	1	6	1245.97
48	1	1	8	1241.24
49	0	1	4	195.59
50	1	0	1	2260.59
51	0	1	4	242.99
52	0	1	5	195.08
53	1	1	3	3194.31
54	0	0	5	307.38
55	1	1	10	4524.01
56	1	1	6	2904.35
57	1	1	3	3154.77
58	1	1	2	2191.78
59	1	1	3	2241.53
60	1	1	3	1241.82
61	0	1	4	2636.53
62	1	0	2	1344.76
63	1	1	3	1544.81
64	1	1	2	1255.77
65	1	0	3	1328.88
66	1	1	8	3258.28
67	1	1	1	2740.52
68	1	0	3	4298.50
69	1	1	2	2185.70
70	1	0	3	251.27
71	1	1	3	3065.67
72	0	1	5	1194.89
73	0	1	5	179.98
74	1	1	8	3845.06
75	1	0	3	1188.66
76	0	1	4	189.36
77	0	0	5	1247.30
78	0	1	4	5004.93
79	1	0	3	1505.03
80	1	1	2	3240.26
81	1	1	8	3254.33
82	0	0	4	334.97
83	1	1	3	242.27
84	1	1	8	4181.90
85	1	1	3	187.78
86	1	1	6	3242.91
87	1	1	7	4334.62
88	1	1	3	1575.97
89	1	1	9	2608.09
90	1	1	8	4703.93

Unit (i) serial numbers	y_i	t_i	a_i	x_i
(1)	(2)	(3)	(4)	(5)
91	1	1	7	1940.05
92	1	1	2	2724.16
93	1	1	3	3199.71
94	1	1	7	1241.56
95	0	1	4	1173.01
96	1	0	1	1435.06
97	0	0	4	251.42
98	1	1	1	3236.45
99	1	0	2	1309.49
100	1	1	1	3247.36
101	1	0	3	1271.32
102	1	1	2	208.24
103	1	1	3	246.96
104	0	0	5	1474.40
105	1	1	2	2430.23
106	1	1	1	1148.49
107	1	1	3	640.08
108	1	1	9	3942.96
109	1	1	8	2202.25
110	0	1	4	241.63
111	1	1	1	4191.92
112	1	0	1	4269.03
113	1	1	3	2742.73

In order to estimate Y we employ the scheme of sample selection given by Rao *et al.* (RHC, 1962) to choose a sample of n = 33 households out of N = 113 households using the household-size a_i as size-measures. The selection consists in (1) dividing the population at random into n = 33 disjoint groups, (2) choosing from each group exactly one household with a probability proportional to the sizes of the households within the groups and (3) repeating the selection-process independently across the groups. Writing for simplicity y_i, a_i, and A_i as the y-value, for the respondent of the household chosen from the ith group, the size of this household and the sum of the sizes of the households falling in the ith group, the unbiased estimator for Y given by RHC is

$$t_R = \sum_n \frac{A_i}{a_i} y_i$$

writing $\sum_n$ for the sum over the n groups. Writing N_i for the number of households falling in the ith group to be chosen as positive integers closest to $\frac{N}{n}$ and subject to $\sum_n N_i = N$ and $\sum_n \sum_n$ to denote summing over the non-duplicated pairs of the n groups formed, it follows that

$$V(t_R)=\frac{\sum_n N_i^2-N}{N(N-1)}\sum p_i\left(\frac{y_i}{p_i}-Y\right)^2$$

where $p_i=\frac{a_i}{A}, A=\sum a_i$

An unbiased estimator for $V(t_R)$ is

$$v(t_R)=\frac{\sum_n N_i^2-N}{N^2-\sum_n N_i^2}\sum_n A_i\left(\frac{y_i}{p_i}-t_R\right)^2$$

$$=\frac{\sum_n N_i^2-N}{N^2-\sum_n N_i^2}\sum_n\sum_n A_iA_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$$

when y_i is unavailable and is unbiasedly estimated by r_i having $V_R(r_i)=V_i$ admitting an unbiased estimator v_i we shall use the unbiased estimator

$$e_R=\sum_n\frac{A_i}{a_i}r_i \text{ for } Y$$

and its unbiased variance estimator

$$v(e_R)=\frac{\sum_n N_i^2-N}{N^2-\sum_n N_i^2}\sum_n\sum_n A_iA_j\left(\frac{r_i}{p_i}-\frac{r_j}{p_j}\right)^2+\sum_n v_i\frac{A_i}{p_i}$$

The corresponding modifications on e_R and $v(e_R)$ will be employed in manners discussed in Section 4 when logistic regression modelling is applicable to utilize values x_i for the single covariate indicated earlier.

In order to assess the performance of any estimator e for Y with a variance (MSE) estimator v we shall consider the two criteria, namely, (1) ACV, which is the average, of the values of $100\times\frac{\sqrt{v}}{e}$ over 1000 replicates of the samples drawn by the RHC scheme, called the 'Average Coefficient of Variation', - the smaller its value the better and (2) ACP, the 'Actual Coverage Percentage' which is the percent of the above 1000 replicated samples for which the 95% confidence intervals (CI), namely $\left(e-1.96\sqrt{v},e+1.96\sqrt{v}\right)$ may cover the true value Y. Here we treat the pivot $\frac{e-Y}{\sqrt{v}}$ as a standard normal deviate, an assumption justifiable for large samples. The closer this ACP to 95 the better the performance of the pair (e,v). For the Table 1 we may mention that $\pi_A=\frac{Y}{N}=0.8230$ and $\frac{\sum t_i}{N}=0.7345$.

Table 2 below summarizes some of the examples of the performances in estimation based on (iii) Kuk's model and (iv) unrelated question RR model. The estimator using a covariate is referred to as a regression estimator, the other one as 'original'.

Table 2. Illustrating relative performances of estimators of a proportion using Kuk's and Unrelated Question devices for RR from unequal probability samples by RHC scheme with and without utilizing covariate information

Serial number of cases	Kuk's RR Model					Unrelated Question Model			
	RR-device parameters			ACV/ACP of original estimator	ACV/ACP of revised estimator	RR-device parameters		ACV/ACP of original estimator	ACV/ACP of revised estimator
	k	p_1	p_2			p_1	p_2		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	4	.21	.02	26.3/90.0	19.4/87.3	.72	.96	15.3/92.2	11.9/84.0
2	4	.40	.14	25.4/91.8	13.7/84.7	.91	.81	16.5/93.2	12.3/83.9
3	4	.76	.96	26.5/90.7	17.8/89.1	.95	.68	15.3/93.1	11.9/83.7
4	4	.65	.97	21.4/90.3	15.2/82.8	.81	.88	18.5/94.3	13.1/83.4
5	5	.50	.98	17.1/95.0	13.7/84.7	.72	.89	16.1/92.2	12.2/83.4
6	5	.57	.98	17.9/93.4	13.8/84.3	.72	.92	15.8/93.1	12.2/81.4
7	5	.32	.03	20.5/92.7	14.3/84.7	.89	.78	16.7/93.7	12.4/82.4
8	5	.26	.01	21.1/92.9	16.3/86.3	.94	.71	15.3/93.2	11.9/82.3
9	5	.75	.96	23.7/90.0	15.9/86.6	.90	.60	15.6/93.2	12.2/82.0
10	5	.27	.05	23.4/92.0	15.3/87.7	.04	.35	22.9/94.2	14.8/81.3
11	5	.64	.87	24.6/92.4	12.6/84.2	.84	.67	17.5/93.0	12.9/81.1
12	5	.73	.95	23.7/90.8	15.1/87.2	.31	.16	57.2/97.4	24.1/80.5
13	5	.69	.93	22.9/94.1	13.7/87.8	.38	.57	27.9/95.3	16.8/80.5
14	5	.25	.08	29.2/90.7	15.0/90.4	.62	.55	86.9/97.9	38.1/80.4
15	5	.84	.97	31.5/89.6	19.6/90.3	.31	.13	37.5/96.8	19.8/80.2
16	5	.21	.05	29.2/89.9	16.9/89.9	.21	.50	22.5/95.2	14.9/80.1
17	5	.73	.95	23.7/90.8	15.1/87.2	.78	.60	19.1/93.9	13.4/80.1
18	5	.77	.92	15.1/90.5	15.1/90.5	.50	.20	22.4/94.7	14.7/80.0
19	5	.39	.08	21.0/91.2	13.4/85.0	.59	.72	24.7/95.9	15.7/79.5
20	6	.50	.92	17.8/92.9	13.1/85.1	.43	.77	17.0/94.2	12.8/79.4
21	6	.19	.02	23.7/89.6	16.8/86.2	.32	.01	22.1/91.8	14.4/79.4
22	6	.52	.17	19.2/92.9	12.3/85.7	.50	.37	55.1/95.9	24.6/79.4
23	6	.16	.94	15.1/96.0	12.9/89.3	.36	.49	57.0/97.4	28.5/79.0
24	6	.39	.84	17.0/92.7	11.4/88.9	.42	.95	15.2/92.4	12.3/79.0
25	6	.23	.96	15.3/95.0	11.8/98.1	.41	.77	17.0/93.1	12.9/79.0
26	6	.13	.95	15.0/95.9	11.7/98.5	.30	.11	33.1/95.2	19.0/78.9

A couple of comments may be made from the results in Table 2. Use of a covariate cuts down the average coefficient of variation but undesirably reduces the coverage percentage as well. This applies to both the RR devices illustrated. Between the two devices the unrelated question model seems to be the better.

Appendix

First let us consider the finite population analogue of $\hat{\beta}$ as

$$B = \frac{\sum\left(z_i - \frac{\sum z_i}{N}\right)\left(x_i - \frac{\sum x_i}{N}\right)}{\sum\left(x_i - \frac{\sum x_i}{N}\right)^2} = \frac{\sum(z_i - \bar{Z})(x_i - \bar{X})}{\sum(x_i - \bar{X})^2} = \frac{\sum T_i}{\sum U_i}, \text{ say}$$

writing $\bar{Z} = \frac{Z}{N}$, $\bar{X} = \frac{X}{N}$, $T_i = (z_i - \bar{Z})(x_i - \bar{X})$, $U_i = (x_i - \bar{X})^2$

This B, if the entire finite population U could be surveyed, instead of only a sample s chosen from it, could be taken as the 'unweighted' or simple 'Least Squares Estimate' (LSE) of β.

Writing $\hat{T}_i$ for T_i on replacing $\bar{Z}$ by $\frac{\hat{Z}}{\hat{N}} = \frac{\sum z_i b_{si} I_{si}}{\sum b_{si} I_{si}}$ and $\bar{X}$ by $\frac{\hat{X}}{\hat{N}} = \frac{\sum x_i b_{si} I_{si}}{\sum b_{si} I_{si}}$ in T_i and similarly writing $\hat{U}_i$ for U_i on replacing $\bar{X}$ by $\frac{\hat{X}}{\hat{N}}$ in U_i we may note that we may write

$$\hat{\beta} = \frac{\sum \hat{T}_i b_{si} I_{si}}{\sum \hat{U}_i b_{si} I_{si}}$$

Thus, $\hat{\beta}$ is a 'ratio estimator' $\hat{R}$, say, for the ratio parameter $R = \frac{\sum T_i}{\sum U_i}$

In Section 3 we have given a formula for the estimator of variance of an unbiased estimator $\hat{Y} = \sum y_i b_{si} I_{si}$ of Y. So, we may likewise write down the formula for an estimator of the MSE of $\hat{T}$ about $T = \sum T_i$ as $m_p(\hat{T})$. So, following recommendations of Rao (1988) in pursuance of the approach of Woodruff's (1971) Taylor-series expansion-based approximation of the MSE of an estimator, we may take the estimator of MSE ($\hat{\beta}$) in the form

$$m_p(\hat{\beta}) = \frac{1}{\left(\sum \hat{U}_i b_{si} I_{si}\right)^2} m_p(\hat{T})\Big|_{\hat{T}_i = \hat{T}_i - \hat{\beta}\hat{U}_i}$$

Similarly, writing $w_i = z_i - \hat{\beta}x_i, \hat{\alpha} = \frac{\sum w_i b_{si} I_{si}}{\sum b_{si} I_{si}}$, the MSE $(\hat{\alpha})$ may be estimated by $m_p(\hat{\alpha}) = \frac{1}{\left(\sum b_{si} I_{si}\right)^2} m_p\left(\sum w_i b_{si} I_{si}\right)\Big|_{\hat{w}_i = \hat{w}_i - \hat{\alpha} I_{si}}$. If $m_p(\hat{\alpha}), m_p(\hat{\beta})$ be small so that $\hat{\alpha}$ and $\hat{\beta}$ may provide accurate estimates of α and β respectively then it is worthwhile to employ

$$\hat{z}_i = \hat{\alpha} + \hat{\beta}x_i$$

as a good estimator for $z_i = \log_e\left(\frac{u_i}{1-u_i}\right)$. Our ultimate interest however is in the accuracy in the estimation of Y which uses r_i and $\hat{z}_i, i \in s$.

ACKNOWLEDGEMENT

The authors wish to acknowledge the referee's helpful comments leading to this revision on an earlier draft.

REFERENCES

Chaudhuri, A. (2001a). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *J. Statist. Plann. Inf.*, **94**, 37-42.

Chaudhuri, A. (2001 b). Estimating sensitive proportions from unequal probability samples using randomized responses. *Pak. Jour. Stat.*, **17(3)**, 259-270.

Chaudhuri, A. (2002). Estimating sensitive proportions from randomized responses in unequal probability sampling. *Cal. Stat. Assoc. Bull.*, **52**, 315-322.

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Responses: Theory and Techniques.* Marcel Dekker, Inc. N.Y.

Greenberg, B.G., Abul-Ela, Abdel-Latif, A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question RR model: Theoretical frame-work. *J. Amer. Statist. Assoc.*, **64**, 520-539.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R. and Horvitz, D.G. (1977). Respondent hazards in the unrelated question randomized response model. *J. Statist. Plann. Inf.*, **1**, 53-60.

Heijden, P.G.M. and Gils, G. Van. (1996). Some logistic regression models for randomized response data. Statistical Modelling. *Proc. 11th Int. Workshop.* Orvieto, Italy, 341-348.

Horvitz, D.G., Shah, B. V. and Simmons, W.R (1967). The unrelated question RR model. *Proc. ASA Soc. Stat. Sec.*, 65-72.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from finite universes. *J. Amer. Statist. Soc.*, **47**, 663-685.

Kerkvliet, J. (1994) Estimating logit model with randomized data: The case of cocaine use. *Austr. J. Statist.*, **36(1)**, 9-20.

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

Maddala, G. (1983). *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge Univ. Press. New York.

Nayak, Tapan K. (1994). On randomized response survey for estimating a proportion. *Comm. Stat. –Theory Methods*, **23(11)**, 3303-3321.

Rao, J.N .K. (1988). *Variance Estimation in Sample Surveys.* In : Handbook of Stat., 6 Ed., C.R Rao and P.R. Krishnaih, North-Holland, Amsterdam.

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

Sheers, N.J. and Dayton, C.M. (1988). Covariate randomized response models. *J. Amer. Statist. Assoc.*, **83**, 969-974.

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

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *J. Amer. Statist. Assoc.*, **66**, 411-414.