# Christofides' randomized response technique in complex sample surveys

**Arijit Chaudhuri**[1]

Applied Statistics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata-700 108, India
(E-mail: achau@isical.ac.in)

**Abstract.** Christofides (2003) has given an improved modification of Warner's (1965) pioneering randomized response (RR) technique in estimating an unknown proportion of people bearing a sensitive characteristic in a given community. As both these RR devices are shown to yield unbiased estimators based only on simple random sampling (SRS) with replacement (WR) but in practice samples are mostly taken with unequal selection probabilities without replacement (WOR), here we present methods of estimation when Christofides' RR data are available from unequal probability samples. Warner's (1965) RR device was earlier shown by Chaudhuri (2001) to be applicable in complex surveys. For completeness we present estimators for the variance of our estimator and also describe what to do if some people opt to divulge truths.

AMS Subject classification: 62 D05

**Key words:** Estimation of proportion, Randomized response, Sensitive issues, Unequal probability sampling

## 1 Introduction

In social researches it is often necessary to make appraisals of proportions of people in a community that bear stigmatizing characteristics like habitual tax evasion, bribe-taking, drunken-driving, gambling, drug abuse etc. for example. Also sample surveys are often undertaken in gathering information relating to various items of interest concerning human societies. Some of these items may include the sensitive ones as illustrated above. Warner (1965) introduced the

pioneering randomized response (RR) devices of gathering information on a sensitive issue in a manner purported to protect privacy of a respondent. But his as well as many of his followers' RR devices require the respondents to be chosen by the SRSWR method. Usually, however, social surveys employ unequal probability sampling without replacement. Moreover, it is difficult to find sponsors funding surveys exclusively covering sensitive issues alone. So, it is considered important to apply RR techniques when people are actually sampled with unequal probabilities rather than by SRSWR. Chaudhuri's (2001) has shown how Warner's (1965) and some other RR techniques are applicable in estimation of proportions of people bearing sensitive features when instead of SRSWR a general WOR sampling is employed.

Christofides (2003), though referred to Chaudhuri's (2001) above publication, has given a new RR technique by way of improvement upon Warner's (1965) confining however only to SRSWR-based survey data. We present here details about how to apply Christofides' (2003) device when a varying probability sample is drawn WOR.

In section 2 we present our details about the estimation procedures. In section 3 we modify them to accommodate an option for direct responses (DR) from those respondents who do not care for privacy concerning the specific item of interest presumed to be sensitive by the investigator, while for the other sampled persons RR's are gathered following the method of Christofides (2003).

## 2 Estimation from complex surveys using RR device by Christofides'

Let $U = (1, \cdots, i, \cdots, N)$ denote a finite survey population of N people and y be an indicator variable such that

$y_i = 1$ if i bears a stigmatizing attribute A and $= 0$ if I bears $A^c$, the complement of A.

Then, writing $\sum$ as sum over i in U, the total number of members of U bearing A is

$Y = \sum y_i$ and $\theta = \frac{Y}{N}$ is the corresponding proportion. Supposing N to be known estimating Y is enough in estimating $\theta$. Christofides (2003) has given a method to estimate $\theta$ from an SRSWR from U taken in n draws if each time a person is selected an RR is realized from him/her on applying the following device in independent manners.

We shall instead suppose that a sample $s$ is selected from U with a probability p(s) according to a general design p and from each selected person an RR is realized on applying Christofides' device.

According to this device, from a given box containing $L(\geq 2)$ cards of a common shape, size, weight and thickness but separately marked as $1, \cdots, j, \cdots, L$ in known proportions

$$p_1, \cdots, p_j, \cdots, p_L (0 < p_j < 1, \sum_{j=1}^{L} p_j = 1),$$

a selected person randomly draws one, unnoticed by the interviewer, and reports

(i) the number, say, k actually drawn if he/she bears $A^c$ or (ii) the number (L + 1-k) if he/she actually draws k but bears A. Then, for the respondent

labelled i the RR randomly and independently of any one else's random realization of an RR, is

$$z_i = (L+1-k)y_i + k(1-y_i), i\epsilon U. \tag{1}$$

Writing $E_R, V_R$ as expectation, variance operators with respect to this RR device we get

$$\begin{aligned} E_R(z_i) &= \sum_{k=1}^{L} k_{p_k} + y_i\left[(L+1) - 2\sum_{1}^{L} k_{p_k}\right] \\ &= \mu_i + y_i(L+1-2_{\mu_i}), \text{ writing } \sum^{L} kp_k = \mu_i, \text{ say, and} \\ V_R(z_i) &= E_R(z_i^2) - \mu_i^2 \\ &= \sum_{1}^{L} p_k[(L+1-k)^2 y_i + k^2(1-y_i)] - \mu_i^2 \\ &= \sum k^2 p_k - \mu_i^2 \text{ on simplification on noting } y_i^2 = y_i, \\ &= V_R(k). \end{aligned}$$

It follows that if we write, assuming $L+1-2\mu_i \neq 0$,

$$r_i = \frac{z_i - \mu_i}{L+1-2\mu_i}, \text{ then } E_R(r_i) = y_i, i\epsilon U, \tag{2}$$

and

$$V_R(r_i) = \frac{V_R(z_i)}{(L+1-2_{\mu_i})^2} = \frac{V_R(k)}{(L+1-2\mu_i)^2} = V_i, \text{ say, } i\epsilon U. \tag{3}$$

If we take L = 2, then Christofides' (2003) scheme reduces to Warner's (1965) scheme, for which (2), (3) simplify respectively to

$$r_i' = \frac{z_i - (2-p_1)}{(2p_1 - 1)} \tag{4}$$

$$\text{and } V_i' = \frac{p_1(1-p_1)}{(2p_1-1)^2} \tag{5}$$

on noting that $p_1 + p_2 = 1$.

Let us proceed as follows to unbiasedly estimate Y using $(s, r_i | i\epsilon s)$, introducing $E_p, V_p$ as the expectation, variance operators with respect to the sampling design p and the constants $b_{si}$ free of $\underline{Y} = (y_1, \cdots, y_i, \cdots, y_N)$ and $\underline{R} = (r_1, \cdots, r_i, \cdots, r_N)$ and writing $R = \sum R_i, I_{si}$ equal to 1 if $i\epsilon s$ and to 0 if $i \notin s$. We shall also write the overall expectation, variance operators, E,V such that

$$E = E_p E_R = E_R E_p, V = E_p V_R + V_p E_R = E_R V_p + V_R E_p.$$

Let us follow Godambe (1955) to introduce $t = \sum y_i b_{si} I_{si}$ subject to $E_p(b_{si} I_{si}) = 1 \forall_i$ so as to get $e = \sum r_i b_{si} I_{si}$ as an unbiased estimator for Y because

$$E(e) = E_p(t) = Y \text{ and also } E(e) = E_R(R) = Y. \tag{6}$$

$$\text{Now, } V(e) = E_p \sum V_i b_{si}^2 I_{si} + V_p(t) \tag{7}$$

and alternatively,

$$V(e) = E_R V_p(e) + V_R(R) = E_R V_p(e) + \sum V_i. \tag{8}$$

We shall next introduce a few notations, namely $I_{sij} = I_{si}I_{sj}, w_i(\neq 0)$ as certain known constants,

$$d_{ij} = E_p(b_{si}I_{si} - 1)(b_{sj}I_{sj} - 1), d_{sij}, c_{si}$$

certain constants free of $\underline{Y}, \underline{R}, C_i = E_p(b_{si}^2 I_{si} - 1)$, such that

$$E_p(c_{si}I_{si}) = C_i \text{ and } E_p(d_{sij}I_{sij}) = d_{ij}$$

Examples of these entities and of p and $b_{si}$ abound in the literature on survey sampling and we may especially cite Rao (1979), and Chaudhuri and Stenger (1992).

Chaudhuri and Pal (2002) have noted that

$$V_p(t) = -\sum_{i=1}^{N}\sum_{\substack{j=2 \\ i<j}}^{N} d_{ij}w_i w_j \left(\frac{y_i}{w_i} - \frac{y_i}{w_j}\right)^2 + \sum \frac{y_i^2}{w_i}\alpha_i,$$

$$\alpha_i = \sum_{j=1}^{N} d_{ij}w_i$$

$$\text{and that } v_p(t) = -\sum\sum_{i<j} d_{sij}I_{sij}w_i w_j \left(\frac{y_i}{w_i} - \frac{y_j}{w_j}\right)^2 + \sum \frac{y_i^2}{w_i}\alpha_i b_{si}I_{si} \tag{9}$$

satisfies $E_p v_p(t) = V_p(t)$. Also we shall write

$$V_p(e) = V_p(t)|_{\underline{Y}=\underline{R}} \text{ and } v_p(e) = v_p(t)|_{\underline{Y}=\underline{R}}.$$

On checking that

$$E_R v_p(e) = v_p(t) - \sum\sum_{i<j} d_{sij}I_{sij}w_i w_j \left(\frac{V_i}{w_i^2} + \frac{V_j}{w_j^2}\right) + \sum \frac{V_i}{w_i}\alpha_i b_{si}I_{si} \tag{10}$$

and noting that $V_i$ is a known quantity we get the

**Theorem 1.**

$$\hat{V}_1 = v_p(e) + \sum_{i<j}\sum d_{sij}I_{sij}w_i w_j \left(\frac{V_i}{w_i^2} + \frac{V_j}{w_j^2}\right) + \sum V_i\left(b_{si}^2 - \frac{\alpha_i}{w_i}b_{si}\right)I_{si}$$

is an unbiased estimator for V(e).

*Proof:* Using (6),(7),(9),(10) it easily follows that $E(\hat{V}_1) = V(e)$. Also we have the

**Theorem 2.**

$\hat{V}_2 = v_p(e) + \sum V_i b_{si}I_{si}$ is an unbiased estimator for V(e).

*Proof:* On noting $E_p v_p(e) = V_p(e)$ and (8) we have $E(\hat{V}_2) = V(e)$. If following Raj (1968) or Rao (1975) we use the form

$$V_p(t) = \sum y_i^2 C_i + \sum_{i \neq j} \sum y_i y_j d_{ij},$$

then writing

$$w_p(e) = \sum r_i^2 c_{si} I_{si} + \sum_{i \neq j} \sum r_i r_j d_{sij} I_{sij} \text{ we get}$$

**Theorem 3.**

$$\hat{V}_3 = w_p(e) + \sum V_i (b_{si}^2 - c_{si}) I_{si}$$

is an unbiased estimator for V(e).

*Proof:* Using (7) and the above specifications we get $E(\hat{V}_3) = V(e)$.

## 3. Optional rather than compulsory RR's

Let the people in a sub-sample $s_1$ of $s$ opt to give out their true values $y_i$ but those in the complementary sub-sample $s_2$ give out the values $r_i$ generated by Christofides' RR technique, the former ones not believing the attribute sensitive enough. Since on knowing the values $y_i, i \epsilon s_1$, two estimators for Y seem to be available, namely the earlier

$$e \text{ and } e^* = \sum_{i \epsilon s_1} y_i b_i I_{si} + \sum_{i \epsilon s_2} r_i b_{si} I_{si},$$

writing $E_{DR}$ as the operator for the conditional expectation with respect to the RR device applicable only to those who divulge their true $y_i$ values we get now,

$$E_{DR}(e) = e^*.$$

$$\text{Now }, E_R(e^*) = \sum y_i b_{si} I_{si} = E_R(e) = t. \tag{11}$$

$$\text{Also }, E_R(e - e^*)^2 = E_R[(e - t) - (e^* - t)]^2 = V_R(e) - V_R(e^*) \tag{12}$$

$$\text{because } E_R(e^* - t)(e - t) = E_R(e^* - t)[E_{DR}(e - t)]$$

$$= E_R(e^* - t)^2 \tag{13}$$

So, if we write $\hat{V}(e)$ for any unbiased estimator for V(e) like $\hat{V}_j (j = 1, 2, 3)$ above then we have

**Theorem 4.**

$$\hat{V}(e^*) = \hat{V}(e) - (e - e^*)^2$$

is an unbiased estimator for $V(e^*)$.

*Proof:* Follows straightforwardly from (11) and (12) because

$$V(e) = E_p E_R (e - t)^2 + E_p (t - Y)^2$$

and

$$V(e^*) = E_p E_R (e^* - t)^2 + E_p (t - Y)^2.$$

## References

Chaudhuri A (2001) Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. Jour. Stat. Plan. Inf. 94:37–42

Chaudhuri A, Pal S (2002) On certain alternative mean square error estimators in complex survey sampling. Jour. Stat. Plan. Inf. 104(2):363–375

Chaudhuri A, Stenger H (1992) Survey Sampling: Theory and Methods. Marœl Dekker, Inc. N.Y.

Christofides TC (2002) A generalized randomized response technique. Metrika 57:195–200

Godambe VP (1955) A unified theory of sampling from finite populations. Jour. Roy. Stat. Soc. B 17:269–278

Raj Des (1968) Sampling theory, Mc Graw-Hill N.Y.

Rao JNK (1975) Unbiased variance estimation for multi-stage designs. Sankhya C 37:133–139

Rao JNK (1979) On deriving mean square errors and other non-negative unbiased estimators in finite population finite population sampling. Jour. Ind. Stat. Assoc. 17:125–136

Warner SL (1965) Randomized response: A survey technique for eliminating evasive answer bias. Jour. Amer. Stat. Assoc. 60:63–69