# Estimating Proportions from Unequal Probability Samples Using Randomized Responses by Warner's and Other Devices

Arijit Chaudhuri and Sanghamitra Pal
*Indian Statistical Institute, Calcutta-700 035*
(Received : December, 2000)

SUMMARY

Sarjinder Singh and Anwar H. Joarder [5] have given an improved method over Warner's [6] in unbiasedly estimating the proportion of people with a sensitive characteristic using randomized response (RR) data. Both use simple random samples (SRS) chosen with replacement (WR). Here we present theoretical and numerical results relating to both when a sample may be selected with unequal probabilities without replacement (WOR), as is the practice in social surveys.

*Key words:* Randomized response, Sensitive proportions, Unequal probability sampling.

## *1. Introduction*

In estimating the proportion $\theta$ of people bearing a stigmatizing characteristic A like habitual tax evasion, drunken driving, gambling etc. it is well-known that Warner [6] considered it useful to avoid seeking direct responses (DR) from respondents in a social survey. Instead he gave us a randomized response (RR) technique by way of protecting the respondent's privacy. According to this a sampled respondent is to implement a randomizing device by which with a pre-assigned probability $p \ (0 < p < 1)$ a truthful response is to be 'Yes' or 'No' about bearing A and with probability (1 – p) about bearing the complementary characteristic $\overline{A}$ without divulging to the interviewer whether the response relates to A or $\overline{A}$.

Based on such RR's procured from an SRSWR chosen in n draws an unbiased estimator for $\theta$ and an unbiased estimator for its variance are given by Warner [6]. Singh and Joarder [5] recommended a modification of Warner's RR procedure enjoining a (I) respondent bearing $\overline{A}$ to respond as in Warner's case but a (II) respondent bearing A to postpone the response to a second performance of Warner's randomizing device unless the first one induces a 'Yes' response.

With such responses from an SRSWR in n draws they prescribe a better unbiased estimator for $\theta$ along with an unbiased variance estimator.

Though the fact is not made explicit by these authors $\theta$ here is a 'finite survey population mean' of an 'Indicator' variable which is valued 1 for a population unit bearing A and 0 for one with $\overline{A}$. But in practice a finite population survey is implemented according to complex designs involving selection in multi-stages and through stratification with sampling in the early stages with unequal selection-probabilities. A sample survey in practice covers numerous, say, fifty items of enquiry of which only a few, say, five may relate to sensitive issues. From such a single survey one must derive good estimators based on 'direct responses' (DR) related to innocuous characteristics and those based on RR's related to the sensitive ones. So, we consider it important to present a theory how $\theta$ above may be estimated admitting variance estimates when RR's are obtained by Warner's [6] and Singh and Joarder's [5] techniques but the respondents are sampled by general sampling schemes with varying probabilities and without replacement.

After presenting revised methods of estimation we supplement Singh and Joarder's numerical findings with ours for the sake of comparison.

## 2. *Unbiased Estimators and Variance Estimators*

According to Warner's RR device the probability for a 'Yes' response about the possession of the characteristic A or its complement $\overline{A}$ is

$$Y_W = p\theta + (1-p)(1-\theta) = (2p-1)\theta + (1-p) \tag{2.1}$$

The corresponding probability for Singh *et al.*'s scheme is

$$\begin{aligned} Y_{SJ} &= p\theta + (1-p)\theta + (1-p)(1-\theta) \\ &= [(2p-1) + p(1-p)]\ \theta + (1-p) \\ &= Y_W + p(1-p)\theta \end{aligned} \tag{2.2}$$

Writing n as the number of draws in SRSWR and m as the number of 'Yes' responses in either case we have

Warner's well-known unbiased estimator for $\theta$ is

$$\hat{\theta}_W = \frac{\left(\frac{m}{n} - 1 + p\right)}{(2p-1)}, \text{ taking } p \neq \frac{1}{2} \tag{2.3}$$

Its variance and an unbiased estimator of the variance are

$$V(\hat{\theta}_W) = \frac{Y_W(1-Y_W)}{n(2p-1)^2} = \frac{\theta(1-\theta)}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{2.4}$$

and

$$v_W = \frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{(n-1)(2p-1)^2}$$

$$= \frac{\hat{\theta}_w\left(1-\hat{\theta}_w\right)}{(n-1)} + \frac{1}{4(n-1)}\left[\frac{1}{4(p-0.5)^2}-1\right] \tag{2.5}$$

Singh *et al.*'s unbiased estimator for $\theta$ is

$$\hat{\theta}_{SJ} = \frac{\left[\frac{m}{n}-(1-p)\right]}{[(2p-1)+p(1-p)]} \tag{2.6}$$

choosing its denominator non-zero.

Its variance and unbiased variance estimator are

$$V\left(\hat{\theta}_{SJ}\right) = \frac{Y_{SJ}(1-Y_{SJ})}{n[(2p-1)+p(1-p)]^2}$$

$$= \frac{\theta(1-\theta)}{n} + \frac{p(1-p)}{n[(2p-1)+p(1-p)]^2}$$

$$- \frac{\theta p(1-\theta)}{n[(2p-1)+p(1-p)]} \tag{2.7}$$

$$v_{SJ} = \frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{(n-1)[(2p-1)+p(1-p)]^2} \tag{2.8}$$

Singh *et al.*'s main theoretical result is

$$V\left(\hat{\theta}_W\right) \geq V\left(\hat{\theta}_{SJ}\right) \text{ for every } p > 0.5 \tag{2.9}$$

Following Chaudhuri ([1], [2]) we present below unbiased estimators for $\theta$ along with unbiased variance estimators based on RR's obtained by Warner's and Singh *et al.*'s devices when the respondents are sampled with unequal selection-probabilities.

Chaudhuri's ([1], [2]) approach is the following. Let $U = (1, \ldots, i, \ldots, N)$ denote a finite survey population of a known number of N people labeled $i = 1, \ldots, N$. Let y be an indicator variable with its value $y_i$ for i as

$$y_i = 1 \quad \text{if i bears A}$$
$$= 0 \quad \text{otherwise}$$

Then, $\theta = \frac{1}{N}\sum y_i$, writing $\sum$ as sum over $i \in U$

Let s be a sample from U chosen according to a design P with a selection-probability $p(s)$. By $E_p, V_p$ we shall denote operators for expectation and variance with respect to P.

We suppose that $y_i$ is not ascertainable for a person i in a sample but adopting a suitable RR device, from an i in a sample, an RR may be procured as $r_i$ 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 and (iv) there exist $v_i$ ascertainable from RR's such that $E_R(v_i) = V_i, i \in U$.

Here $E_R, V_R$ denote operators for expectation, variance with respect to RR devices. The over-all expectation and variance operators will be denoted by

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

Writing $I_{si} = 1$ if $i \in s$, 0 if $i \notin s$, $I_{sij} = I_{si} I_{sj}$ let it be possible to choose $b_{si}, d_{si}$, $I_{sij} = I_{si} I_{sj}$ as constants free of $\mathbf{Y} = (y_1, \ldots, y_i, \ldots, y_N)$ and $\mathbf{R} = (r_1, \ldots, r_i, \ldots r_N)$ such that

$$t_b = \frac{1}{N}\sum y_i b_{si} I_{si} \text{ subject to } E_p(b_{si} I_{si}) = 1 \ \forall\, i$$

Then, $V_p(t_b) = \frac{1}{N^2}\left[\sum y_i^2 C_i + \sum\sum_{i \neq j} y_i y_j C_{ij}\right]$

where
$$C_i = E_p(b_{si}^2 I_{si}) - 1$$
$$C_{ij} = E_p(b_{si} b_{sj} I_{sij}) - 1$$

Then $v_p(t_b) = \frac{1}{N^2}\left[\sum y_i^2 d_{si} I_{si} + \sum\sum_{i \neq j} y_i\ y_j\ d_{sij}\ I_{sij}\right]$

satisfies $E_p v_p(t_b) = V_p(t_b)$

provided $d_{si}, d_{sij}$'s are chosen subject to

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

The literature on 'Sample surveys' is full of numerous such possibilities of choices for P, $b_{si}, d_{si}, d_{sij}$'s. Since $y_i$'s are not ascertainable, $t_b$ is not available as an estimator for θ. So, Chaudhuri's ([1], [2]) recommended unbiased estimator for θ based on RR is

$$e_b = \frac{1}{N}\sum r_i b_{si} I_{si} \text{ for which } E(e_b) = \theta$$

Here $e_b$ is just $t_b$ with $y_i$'s replaced by $r_i$'s, $i \in s$.

Similarly we should write $V_p(e_b)$ as $V_p(t_b)$ with $y_i$ replaced by $r_i$ for i in U and $v_p(e_b)$ as $v_p(t_b)$ with $y_i$ replaced by $r_i$, $i \in s$.

Two unbiased estimators for the variance $V(e_b)$, of $e_b$ which is,

$$V(e_b) = E_p V_R(e_b) + V_p E_R(e_b)$$

$$= \frac{1}{N^2}\left(E_p\left[\sum V_i b_{si}^2 I_{si}\right]\right) + V_p(t_b) \quad (2.10)$$

$$= E_R V_p(e_b) + V_R E_p(e_b)$$

$$= E_R V_p(e_b) + \frac{1}{N^2}\left(V_R\left(\sum r_i\right)\right) \quad (2.11)$$

are

$$v(1) = v_p(e_b) + \frac{1}{N^2}\left(\sum v_i b_{si} I_{si}\right) \quad (2.12)$$

and

$$v(2) = v_p(e_b) + \frac{1}{N^2}\left[\sum v_i\left(b_{si}^2 - d_{si}\right)I_{si}\right] \quad (2.13)$$

It is easy to check that

$$Ev(1) = V(e_b) = Ev(2) \quad (2.14)$$

In order to develop formulae corresponding to $e_b$, $v(1)$, $v(2)$ for the specific RR devices by Warner [6] and Singh *et al.* [5] based on a sample of $r_i$'s for $i \in s$ let us use the following notations.

Let $I_i = 1$ if i responds "Yes"
$= 0$ otherwise

Then, for Warner's [6] scheme $r_i$ should be taken as

$$r_i = \frac{I_i - (1-p)}{(2p-1)} = r_i(W), \text{ say for which } E_R(r_i(W)) = y_i \quad (2.15)$$

with a variance, say, $V_i(W)$ as

$$V_i(W) = V_R(r_i(W))$$

$$= \frac{1}{(2p-1)^2}\left[y_i(2p-1) + (1-p) - \left(y_i(2p-1) + (1-p)\right)^2\right]$$

$$= \frac{p(1-p)}{(2p-1)^2}, \text{ noting } y_i = y_i^2 \quad (2.16)$$

Since $V_i(W)$ does not involve any unknown parameters we need not seek any estimator $v_i(W)$, say, for it and use this $V_i(W)$ straightaway for $v_i$ in (2.12)-(2.13).

For Singh *et al.*'s [5] scheme, $r_i$ should be taken as

$$r_i = \frac{I_i - (1-p)}{(2p-1) + p(1-p)} = r_i\ (SJ)\ \text{(say)}$$

Then $E_R(r_i(SJ)) = y_i$ (2.17)

Writing for simplicity, $\alpha = (2p-1) + p(1-p)$

we may work out the variance of $r_i(SJ)$ as, say

$$\begin{aligned} V_i(SJ) &= V_R(r_i(SJ)) = \frac{1}{\alpha^2}\left[E_R(I_i)(1 - E_R(I_i))\right] \\ &= \frac{1}{\alpha^2}\left[\alpha y_i + (1-p) - (\alpha y_i + (1-p))^2\right] \\ &= \frac{1}{\alpha^2}\left[\beta y_i + p(1-p)\right] \text{ writing } \beta = \alpha(1-\alpha) - 2\alpha(1-p) \end{aligned}$$

Since $\beta$ is thus known, this $V_i(SJ)$ may be estimated unbiasedly by

$$v_i(SJ) = \frac{1}{\alpha^2}\left[\beta r_i + p(1-p)\right]$$

which may be used to replace $v_i$ in (2.11), (2.12) in using $v(j), j = 1, 2$

On simplification we may check that

$$\begin{aligned} V_i(SJ) &= \frac{p(1-p)}{\alpha^2} && \text{if } y_i = 0 \\ &= \frac{p(1-p)^2(2-p)}{\alpha^2} && \text{if } y_i = 1 \end{aligned}$$

Writing $e_b(W), e_b(SJ)$ for $e_b$ based respectively on Warner's [6] and Singh *et al.*'s [5] schemes and $V(e_b(W)), V(e_b(SJ))$ as their respective variances we have

*Lemma 1.* $V(e_b(W)) \geq V(e_b(SJ))$ if $V_i(W) \geq V_i(SJ)\ \forall\ i$

*Proof.* Follows immediately from (2.10). Next we have

*Lemma 2.* $V_i(W) \geq V_i(SJ)\ \forall\ i$ if $p \geq .4384$

*Proof.* $V_i(W) - V_i(SJ) = \dfrac{p(1-p)}{(2p-1)^2} - \dfrac{\beta y_i + p\ (1-p)}{\alpha^2}$

$$= p(1-p)\left[\frac{1}{(2p-1)^2} - \frac{1}{((2p-1)+p(1-p))^2}\right] \text{if } y_i = 0$$

$> 0$ if $p > 0.4384$ as verifiable using Matlab (i)

$$= p(1-p)\left[\frac{1}{(2p-1)^2} - \frac{(1-p)(2-p)}{((2p-1)+p(1-p))^2}\right] \text{if } y_i = 1$$

$> 0$ if $p > 0.4366$ as verifiable using Matlab (ii)

Hence follows Lemma 2. Hence we have

*Theorem.* $V(e_h(W)) \geq V(e_h(SJ)) \geq 0$ if $p > 0.4384$

The next section presents a numerical study as a follow-up of Singh *et al.*'s exercise.

### 3. A Comparative Study with Numerical Illustrations

In order to maintain parity with Singh *et al.*'s [5] numerical illustration let us make separately 9 alternative choices of $y_i$'s in $\mathbf{Y} = (y_1, \ldots, y_i, \ldots, y_N)$ so as to get 9 alternative values for $\theta = \frac{1}{N}\sum y_i$ as 0.1 (0.1) 0.9 treating $y_i = 1$ for the $i^{th}$ person having a minimum monthly income $C_j$, say, with 9 choices of $j = 1, \ldots, 9$ with $y_i = 0$, else. Further, we associate with $\mathbf{Y}$ a vector $\mathbf{Z} = (Z_1, \ldots, Z_N)$ of positive numbers as size-measures to be used in drawing a sample with suitable unequal selection-probabilities. For illustration we take $N = 20$, $n = 7$ which in the case of (I) SRSWR is the number of draws and is the number of distinct units to be selected in employing two other sampling schemes, namely (II) Rao, Hartley and Cochran's (RHC [4]) scheme and (III) Hartley and Rao's (HR [3]) scheme. We take $\mathbf{Z} =$ (21.9, 20.1, 18.9, 18.3, 17.3, 17.2, 16.5, 16.4, 15.7, 11.6, 9.5, 9.3, 9.2, 9.2, 8.4, 8.4, 7.6, 7.5, 7.2, 5.8).

Writing, $Z = \sum z_i$, $p_i = \frac{z_i}{Z}$, which are the normed size-measures we may briefly describe the schemes (II), (III) as follows

In the RHC scheme the population is 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 $i^{th}$ group for the RHC scheme II, we have

$$t_h = \frac{1}{N}\sum_n y_i \frac{Q_i}{p_i}$$

$$V_p(t_b)=\frac{1}{N^2}\left[\frac{\sum_n N_i^2-N}{N(N-1)}\sum p_i\left(\frac{y_i}{p_i}-Y\right)^2\right], Y=\sum y_i$$

$$v_p(t_b)=\frac{1}{N^2}\left(\frac{\sum_n N_i^2-N}{N^2-\sum_n N_i^2}\right)\sum_n Q_i\left(\frac{y_i}{p_i}-t_b\right)^2; b_{si}=\frac{Q_i}{p_i}$$

$$d_{si}=\left(\frac{\sum_n N_i^2-N}{N^2-\sum_n N_i^2}\right)\left(\frac{Q_i}{p_i^2}+\left(\sum_n Q_i\right)\frac{Q_i^2}{p_i^2}-2\frac{Q_i^2}{p_i^2}\right)$$

$$N^2V(e_b)=B\sum\frac{V_i}{p_i}+(1-B)\sum V_i+B\left(\sum\frac{y_i}{p_i}-Y\right)^2$$

writing $B=\frac{\sum_n N_i^2-N}{N(N-1)}$

For the SRSWR scheme I, we have

$b_{si}=\frac{Nf_{si}}{n}$, writing $f_{si}$ = number of times i occurs in s

$$d_{si}=\frac{N^2}{n(n-1)}\left(f_{si}-\frac{f_{si}^2}{n}\right), \quad V(e_b)=\frac{\theta(1-\theta)}{n}+\frac{N+n-1}{nN^2}\sum V_i$$

In the HR scheme III the units of U are permuted at random and then *n* units are chosen circular systematically with probabilities proportional to sizes. Further, for this

$$t_b=\frac{1}{N}\sum\frac{y_i}{\pi_i}I_{si}, \pi_i=np_i=\sum_{s\ni i}p(s), \pi_{ij}=\sum_{s\ni i,j}p(s), b_{si}=\frac{1}{\pi_i}$$

$$v_p(t_b)=\sum y_i^2\frac{1-\pi_i}{\pi_i}\frac{I_{si}}{\pi_i}+\sum\sum_{i\neq j}y_iy_j\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\frac{I_{sij}}{\pi_{ij}}$$

$$d_{si}=\frac{1-\pi_i}{\pi_i^2}, b_{si}^2-d_{si}=\frac{1}{\pi_i}=b_{si} \text{ implying } v(1)=v(2)$$

$$V(e_b)=\frac{1}{N^2}\left[\sum y_i^2\frac{1-\pi_i}{\pi_i}+\sum\sum_{i\neq j}y_iy_j\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}+\sum\frac{V_i}{\pi_i}\right]$$

Table. Showing the values of PRE for 3 schemes (I, II, III) given from top to bottom

| θ \ p | 0.45 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|
| 0.1 | 219.85 | 477.60 | 226.23 | 154.48 | 119.18 |
| | 219.48 | 487.18 | 233.54 | 160.65 | 123.55 |
| | 219.23 | 487.62 | 234.06 | 161.21 | 124.06 |
| 0.2 | 222.40 | 477.31 | 224.63 | 153.06 | 118.64 |
| | 221.42 | 494.80 | 237.83 | 163.72 | 125.54 |
| | 220.83 | 494.75 | 238.29 | 164.39 | 126.31 |
| 0.3 | 225.21 | 482.88 | 226.87 | 154.40 | 119.55 |
| | 223.87 | 504.95 | 243.53 | 167.64 | 127.88 |
| | 223.06 | 504.02 | 234.55 | 168.07 | 128.51 |
| 0.4 | 228.33 | 494.65 | 232.88 | 158.14 | 121.54 |
| | 226.93 | 517.54 | 250.57 | 172.00 | 130.03 |
| | 226.09 | 516.31 | 250.26 | 172.39 | 130.63 |
| 0.5 | 231.76 | 513.58 | 243.29 | 164.66 | 124.77 |
| | 229.78 | 537.43 | 263.06 | 181.35 | 135.56 |
| | 228.70 | 535.38 | 262.73 | 181.91 | 136.61 |
| 0.6 | 235.54 | 541.48 | 259.69 | 175.18 | 129.92 |
| | 232.97 | 563.19 | 280.40 | 194.63 | 143.55 |
| | 231.71 | 560.47 | 280.04 | 195.75 | 145.60 |
| 0.7 | 239.73 | 581.54 | 285.44 | 192.69 | 138.68 |
| | 236.32 | 597.98 | 306.40 | 216.72 | 158.34 |
| | 234.84 | 594.84 | 306.62 | 219.88 | 163.90 |
| 0.8 | 244.37 | 639.46 | 327.98 | 225.04 | 155.99 |
| | 240.14 | 642.84 | 343.45 | 252.07 | 185.06 |
| | 238.56 | 641.44 | 347.12 | 262.89 | 204.73 |
| 0.9 | 249.50 | 726.07 | 407.08 | 301.04 | 204.80 |
| | 245.23 | 694.49 | 386.29 | 291.64 | 210.43 |
| | 244.08 | 702.17 | 401.99 | 323.65 | 262.03 |

Following Singh *et al.* we consider the criteria for comparison, namely

$$\mathrm{PRE} = 100\frac{V(\hat{\theta}_w)}{V(\hat{\theta}_{SJ})} \tag{3.1}$$

the higher its magnitude the better is $\hat{\theta}_{SJ}$ relative to $\hat{\theta}_W$ and present these values based on each of the three schemes of sampling we employ as above.

*Remark.* The entries in the first rows of the above table corresponding to $p = 0.6\,(0.1)\,0.9$ for each θ "equal to 0.1 (0.1) 0.9" match the PRE values given by Singh *et al.* (with a few slight discrepancies possibly because of misprints in Singh *et al.*) calculated by them using the formula

$$\mathrm{PRE} = 100\frac{V(\hat{\theta}_w)}{V(\hat{\theta}_{SJ})}$$

as they obviously should.

ACKNOWLEDGEMENT

The authors thankfully acknowledge the referee's helpful comments leading to this revised version of an earlier draft.

REFERENCES

[1] Chaudhuri, A. (1999). Towards a unified theory of randomized response surveys for dichotomous finite populations. *Tech. Rep. ASD/99/36*, ISI, Calcutta.

[2] Chaudhuri, A. (2000). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. To appear in *J. Stat. Plan. Inf.*

[3] Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Ann. Math. Stat.*, **33**, 350-374.

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

[5] Singh Sarjinder and Joarder Anwar, H. (1997). Unknown repeated trials in randomized response sampling. *Jour. Ind. Soc. Agril. Stat.*, **50**, 70-74.

[6] Warner, S.L. (1965). RR: A survey technique for eliminating evasive answer bias. *Jour. Amer. Stat. Assoc.*, **60**, 63-69.