Title: <u>Unequal Probability Sampling</u>

Analyzing Optional Randomized Response on Qualitative & Quantitative Variables Bearing Social Stigma

Author: **Arijit Chaudhuri**

Affiliation: **Indian Statistical Institute, Kolkata, India.**

Abstract:

We illustrate a few popular Randomized Response Techniques to elicit responses to sensitive items. Both qualitative and quantitative characteristics are covered. General Sampling Schemes even without replacement are permitted. Also allowed is an undisclosed option to respond directly instead. Certain relevant procedures are critically examined.

Introduction:

We consider a finite population U = (1, .., i, .. , N) of a known number N of persons identified by respective labels 1 through N. On U is defined a vector $\underline{Y} = (y_1,.., y_i,.., y_N)$ of values y_i of a variable y for the *ith* person, $i \in U$. By $Y = \sum_1^N y_i$ we denote the population total which we require to unbiasedly estimate from a sample s chosen from U according to a design P with a probability $p(s)$. For P the inclusion-probability $\pi_i = \sum_{s \ni i} p(s)$ for every i and also the probability $\pi_{ij} = \sum_{s \ni i, j} p(s)$ for every pair (i, j), $i \neq j$ are supposed to be positive. We shall restrict to an estimator for Y of the form

$$t_b = \sum_1^N y_i b_{si} I_{si} \quad \text{...(1)}$$

Here $I_{si} = 1/0$ if $i \in s / (i \notin s)$ and b_{si} for every s and every i is free of the co-ordinates of $\underline{Y}$.

Introducing some non-zero numbers $w_i, i \in U$ and writing $d_{ij} = \sum_s p(s)(b_{si} I_{si} - 1)(b_{sj} I_{sj} - 1)$ and $\beta_i = \sum_j d_{ij} w_j$ it is well-known that provided $\sum_s p(s) b_{si} I_{si} = 1 \; \forall \, i \in U$, t_b has its design based expectation $E_p(t_b) = \sum p(s) t_b = Y \; \forall \; \underline{Y}$ i.e t_b is unbiased for Y and its design variance is

$$V_p(t_b) = \sum y_i^2 C_i + \sum \sum y_i y_j C_{ij} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

writing $C_i = \sum_s p(s) b_{si}^2 I_{si} - 1$ and $C_{ij} = \sum_s p(s) b_{si} b_{sij} I_{sij} - 1$, $I_{sij} = I_{si} I_{sj}$. Alternatively, vide Chaudhuri (2010) $V_p(t_b) = -\sum_i \sum_j w_i w_j d_{ij} \left(\frac{y_i}{w_i} - \frac{y_j}{w_j} \right)^2 + \sum \frac{y_i^2}{w_i} \beta_i \; \ldots\ldots\ldots\ldots\ldots\ldots..(3)$

A particular case of t_b is $t_H = \sum \frac{y_i}{\pi_i}$, the Horvitz & Thompson's (1952) estimator with a variance $V = V_p(t_H) = \sum y_i^2 \left(\frac{1 - \pi_i}{\pi_i} \right) + \sum_{i \neq} \sum_j y_i y_j \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(4)$

for which an alternative form is

$$V' = V_p(t_H) = \sum_{i<} \sum_j (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 + \sum \frac{y_i^2 \alpha_i}{\pi_i}$$

writing $\alpha_i = 1 + \frac{1}{\pi_i} \sum_{j \neq i} \pi_{ij} - \sum_i \pi_i \; \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots......(5)$

We shall consider two separate cases, namely (I) Qualitative when y_i takes only one of the 2 possible values 1 or 0 and (II) Quantitative when every y_i may take any real value. In case (I) in each of the formulae (2) – (5) above each y_i^2 is to replaced by y_i for $i \in U$ and we shall refer to the revised formulae as $(2)' - (5)'$.

Introducing constants C_{si} and C_{sij}'s free of $\underline{Y}$ subject respectively to $\sum p(s) C_{si} I_{si} = C_i$, $\sum_s p(s) C_{sij} I_{sij} = C_{ij}$ it follows from (2) and (3) that

$$v = v_p(t_b) = \sum y_i^2 C_{si} I_{si} + \sum_i \sum_j y_i y_j C_{sij} \; \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots...(6)$$

has $E_p v_p(t_b)$ = V_p (t_b) i.e v is unbiased for V_p (t_b) and

$$v' = v'_p\left(t_b\right) = -\sum_i \sum_j w_i w_j d_{sij} I_{sij} \left(\frac{y_i}{w_i} - \frac{y_j}{w_j}\right)^2 + \sum \frac{y_i^2}{w_i} \beta_i I_{si} / \pi_i \quad \text{………………...(7)}$$

with d_{sij}'s as constants free of $\underline{Y}$ subject to $\sum_s p(s) d_{sij} I_{sij} = d_{ij}$, satisfies $E_p v'_p\left(t_b\right) = V'_p\left(t_b\right)$ so that $v'_p\left(t_b\right) = v'$ is unbiased for $V'_p\left(t_b\right)$. Of course in case (I) every y_i^2 should be replaced by y_i in the formulae (6) - (7) and when so done these formulae labelled as $(6)' - (7)'$.

Next we consider the main problem when y_i values may not be directly ascertained from the sampled persons labelled in U because they relate to sensitive and stigmatizing issues. Then a standard practice is to elicit randomized responses (RR) from the persons sampled in suitable ways. Section 2 below presents a few standard RR devices.

2. Certain RR Devices

Case I. (i) Warner's (1965) RR device as reformulated by Chaudhuri (2001) is as follows. A person labelled i if sampled is presented a box of identical cards differing only in being marked A or A^C in proportions $p : (1 - p)$, $\left(0 < p \neq \frac{1}{2} < 1\right)$. If i bears a stigmatizing characteristic A, then y_i = 1; otherwise y_i = 0. Randomly choosing a card from the box before putting it back the person responds I_i; this I_i = 1 if the card type matches the person's trait A or A^C; else I_i = 0. Writing E_R, V_R as the operators for expectation and variance generically for any RR device it follows that

$$E_R\left(I_i\right) = p y_i + (1 - p)(1 - y_i) = (1 - p) + (2p - 1) y_i \text{ and}$$

$$V_R\left(I_i\right) = E_R\left(I_i\right)\left(1 - E_R\left(I_i\right)\right) = p(1 - p) \quad \text{since} \quad I_i^2 = I_i \quad \text{and} \quad y_i^2 = y_i.$$

Writing $r_i = \dfrac{I_i - (1 - p)}{(2p - 1)}$, one gets $E_R\left(r_i\right) = y_i$ and $V_R\left(r_i\right) = \dfrac{p(1 - p)}{(2p - 1)^2} = V_i$, say.

We shall assume $E_p E_R = E_R E_p = E$, say, and $E_p V_R + V_p E_R = E_R V_p + V_R E_p = V$, say.

Consequently, writing $\underline{R} = \left(r_1, .., r_i, .., r_N\right)$, and $R = \sum r_i$,

$e_b = t_b\big|_{\underline{Y}=\underline{R}} = \sum r_i b_{si} I_{si}$ and $e_H = t_H\big|_{\underline{Y}=\underline{R}} = \sum \frac{r_{i'}}{\pi_i}$ one gets $E(e_b) = E_p(t_b) = E_R(R) = Y$ and also $E(e_H) = E_p(t_H) = E_R(R) = Y$ i.e both e_b and e_H are unbiased for Y. Further,

$$\begin{aligned}V(e_b) &= V_p[E_R(e_b)] + E_p[V_R(e_b)]\\ &= V_p(t_b) + E_p\left[\sum V_i b_{si}^2 I_{si}\right]\\ &= \sum y_i C_i + \sum_{i\neq}\sum_j y_i y_j C_{ij} + \sum V_i(1+C_i)\\ &= E_R E_p\left[\sum r_i C_{si} I_{si} + \sum_{i\neq}\sum_j r_i r_j C_{sij} I_{sij}\right] + E_R E_p\left[\sum V_i\left(\frac{1}{\pi_i} + C_{si}\right) I_{si}\right].\end{aligned}$$

So, $v(e_b) = \sum r_i C_{si} I_{si} + \sum_{i\neq}\sum_j r_i r_j C_{si} I_{sij} + \sum V_i\left(\frac{1}{\pi_i} + C_{si}\right) I_{si}$ ………………………….(8)

is an unbiased estimator for $V(e_b)$. Alternatively,

$$\begin{aligned}V(e_b) &= V_R[E_p(e_b)] + E_R[V_p(e_b)]\\ &= V_R\left(\sum r_i\right) + E_R\left[-\sum_{i\neq}\sum_j w_i w_j d_{ij}\left(\frac{r_i}{w_i} - \frac{r_j}{w_j}\right)^2 + \sum \frac{r_i^2}{w_i}\beta_i\right]\\ &= \sum V_i + \left[-\sum_{i\neq}\sum_j w_i w_j d_{ij}\left(\frac{V_i + y_i}{w_i^2} + \frac{V_j + y_j}{w_j^2} - \frac{2y_i y_j}{w_i w_j}\right) + \sum \frac{V_i + y_i}{w_i}\beta_i\right]\\ &= -\sum_{i\neq}\sum_j w_i w_j d_{ij}\left(\frac{y_i}{w_i} - \frac{y_j}{w_j}\right)^2 + \sum \frac{y_i}{w_i}\beta_i + \sum V_i\left[1 - \frac{\beta_i}{w_i}\right]\end{aligned}$$

Let $v'(e_b) = -\sum_{i\neq}\sum_j w_i w_j\left(\frac{r_i}{w_i} - \frac{r_j}{w_j}\right)^2 d_{sij} I_{sij} + \sum \frac{r_i}{w_i}\frac{\beta_i}{\pi_i} I_{si} + \sum V_i \frac{I_{si}}{\pi_i}$

Then, $E_R v'(e_b) = -\sum_{i\neq}\sum_j w_i w_j\left[\left(\frac{y_i}{w_i} - \frac{y_j}{w_j}\right)^2 + \frac{V_i}{w_i^2} + \frac{V_j}{w_j^2}\right] d_{sij} I_{sij} + \sum \frac{y_i}{w_i}\beta_i\frac{I_{si}}{\pi_i} + \sum V_i\frac{I_{si}}{\pi_i}$

$$\begin{aligned}E\, v'(e_b) &= E_p E_R[v'(e_b)]\\ &= -\sum_{i\neq}\sum_j w_i w_j d_{ij}\left(\frac{y_i}{w_i} - \frac{y_j}{w_j}\right)^2 + \sum \frac{y_i}{w_i}\beta_i + \sum V_i\left(1 - \frac{\beta_i}{w_i}\right)\end{aligned}$$

Then, $v'(e_b)$ is unbiased for $V(e_b)$.

Again,

$$V(e_H) = V_p(E_R(e_H)) + E_p[V_R(e_H)]$$

$$= V_p\left(\sum_{i \in s} \frac{y_i}{\pi_i}\right) + E_p[V_R(e_H)]$$

$$= \sum y_i \left(\frac{1-\pi_i}{\pi_i}\right) + \sum_{i \neq}\sum_{j} y_i y_j \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) + \sum \frac{V_i}{\pi_i}$$

So, $v(e_H) = \sum r_i \left(\frac{1-\pi_i}{\pi_i}\right) \frac{I_{si}}{\pi_i} + \sum_{i \neq}\sum_{j} r_i r_j \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) \frac{I_{sij}}{\pi_{ij}} + \sum \frac{V_i}{\pi_i} \frac{I_{si}}{\pi_i}$

is an unbiased estimator for $V(e_H)$. Alternatively,

$$V(e_H) = V_R(E_p(e_H)) + E_R[V_p(e_H)]$$

$$= \sum V_i + E_R\left[\sum_{i<}\sum_{j} \left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j}\right)^2 - (\pi_i \pi_j - \pi_{ij}) + \sum \frac{\alpha_i}{\pi_i} r_i^2\right]$$

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

An unbiased estimator for this $V(e_H)$ is

$$v'(e_H) = \sum_{i<}\sum_{j} (\pi_i \pi_j - \pi_{ij}) \left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j}\right)^2 \frac{I_{sij}}{\pi_{ij}} + \sum \alpha_i \frac{r_i}{\pi_i} \frac{I_{si}}{\pi_i} + \sum \frac{V_i}{\pi_i}(1-\alpha_i-\pi_i) \frac{I_{si}}{\pi_i} \quad \text{...........(9)}$$

While employing Warner's (1965) RR device some respondents may be found to opt for giving out the actual facts treating the item not stigmatizing at all. Let a part s_1 of the sample s chosen with probability $p(s)$ yield the true values y_i for $i \in s_1$ but in the remainder s_2 of the sample s only r_i for $i \in s_2$ is available on applying the Warner's (1965) RR technique as narrated already.

Then we may consider in estimating $Y = \sum y_i$, the three entities namely $t_b = \sum y_i b_{si} I_{si} = \sum_{i \in s} y_i b_{si}$, $e_b = \sum_{i \in s} r_i b_{si}$ and $e_b' = \sum_{i \in s_1} y_i b_{si} + \sum_{i \in s_2} r_i b_{si}$. Writing $e_b = \sum_{i \in s_1} r_i b_{si} + \sum_{i \in s_2} r_i b_{si}$, we may note $E_R\left(e_b \middle| y_i,\, i \in s_1\right) = e_b'$ ………………………………(10)

$E_R(e_b') = t_b = E_R(e_b)$. Denoting the conditional expectation-operator $E_R\left(\bullet \middle| y_i,\, i \in s_1\right)$ by E_{CR} let us note following Chaudhuri & Saha (2005),

$$E_R(e_b - e_b')^2 = E_R\left[(e_b - t_b) - (e_b' - t_b)\right]^2$$

$$= V_R(e_b) + V_R(e_b') - 2E_R(e_b' - t_b)E_{CR}(e_b - t_b)$$

$$= V_R(e_b) - V_R(e_b') \text{ giving}$$

$$V_R(e_b') = V_R(e_b) - E_R(e_b - e_b')^2 \le V_R(e_b)$$

So, $V(e_b') = E_p V_R(e_b') + V_p E_R(e_b')$

$$= E_p V_R(e_b) - E_p E_R(e_b - e_b')^2 + V_p E_R(e_b) \text{ because } E_R(e_b') = E_R(e_b).$$

Thus, $V(e_b') = V(e_b) - E_p E_R(e_b - e_b')^2$ ……………………………………………………...(11)

So, one unbiased estimator for $V(e_b')$ is $v_1(e_b') = v(e_b) - (e_b - e_b')^2$ ……………………..(12)

Noting, $E_R(e_b - e_b')^2 = E_R\left[\sum_{i \in s_1}(r_i - y_i)b_{si}\right]^2 = \sum_{i \in s_1} V_i b_{si}^2$ it follows that an alternative unbiased estimator of $V(e_b')$ is $v_2(e_b') = v(e_b) - \sum_{i \in s_1} V_i b_{si}^2$.

From (11) it follows that if some sampled people opt to give out the true responses while some others produce randomized responses then a greater efficiency in estimation may be achieved on going for utilization of the known direct responses combined with the randomized responses gathered from two complementary parts of the sample.

There is an alternative approach in handling optional RR's. We report Chaudhuri & Dihidar's (2009) work in this context in brief. In trying to unbiasedly estimate the proportion θ of people bearing a stigmatizing characteristic a person sampled *i*, say, may be approached with a request either to (1) give out the genuine truth about bearing *A* or its complement A^C or alternatively to (2) implement Warner's (1965) RRT offering a box of cards in proportions $p:(1-p)$, $(0<p\neq\frac{1}{2}<1)$ marked A rather than A^C.

Let $y_i = 1/0$ if i bears A/A^C,

$I_i = 1/0$ if i finds a 'match' in the card-type versus his/her real feature and $C_i(0<C_i<1)$ be an unknown probability that i responds using the option (1) above rather than (2).

Letting

$z_i = y_i$ with probability C_i

$= I_i$ with probability $(1 - C_i)$

it follows that

$$E_R(z_i) = C_i y_i + (1-C_i)\left[p y_i + (1-p)(1-y_i)\right].$$

To estimate y_i it is an easy course to eliminate C_i by getting an independent response z_i' from i allowing a second box to execute Warner's (1965) RRT with a different proportion $p'(p'\neq p,\ 0<p'<1)$ of A/A^C-marked cards.

Then, one may work out $r_i = \dfrac{(1-p')z_i - (1-p)z_i}{(p-p')}$ so as to get $y_i = E_R(r_i)$ and also, $V_R(r_i) = E_R(r_i - 1)r_i$ so that $v_i = r_i(r_i - 1)$ is an unbiased estimator for $V_i = V_R(r_i)$. Then, corresponding to $t_b = \sum_{i\in s} y_i b_{si}$ one may employ $f_b = \sum_{i\in s} r_i b_{si}$ to unbiasedly estimate $Y = \sum y_i$. Then one may unbiasedly estimate $V(f_b) = E_p V_R(f_b) + V_p E_R(f_b) = E_R V_p(f_b) + V_R E_p(f_b)$ by

$$\hat{V}_1(f_b) = \sum_{i\in s} v_i\left(b_{si}^2\right) + \left[\left(\sum_{i\in s} r_i C_{si} + \sum_{i\neq}\sum_{j} r_i r_j C_{sij}\right)\right]$$ and also by

$$\hat{V}_2(f_b) = \sum_{i\in s} v_i(b_{si}) + \left[\sum_{i\in s} r_i C_{si} + \sum_{i\neq j}\sum_{\in s} r_i r_j C_{sij}\right].$$

Again, if the stigmatizing variable refers to real numbers like days of drunken driving, numbers of induced abortions, amount spent on gambling etc, then also an optional RR approach may work as follows, vide Chaudhuri & Dihidar (2009).

Suppose a person labelled i may, with an unknown probability C_i give out the true value of y_i or with the complementary probability $(1-C_i)$ give an RR on executing a trick as follows.

Suppose the person i is offered a box carrying cards marked $a_1, .., a_j, .., a_m$ and a second box with cards marked $b_1, b_2, ..., b_L$ with a request to independently take one card from each and report the value $I_i = a_j y_i + b_k$, say and independently repeat this exercise to likewise report a second value $I_i' = a_u y_i + b_v'$, say, using a 3rd box with cards marked $b_1',...,b_L'$.

Letting $z_i = y_i$ with probability C_i

$= I_i$ with probability $(1-C_i)$

and $z_i' = y_i$ with probability C_i

$= I_i'$ with probability $(1-C_i)$

and defining $\mu_a = \frac{1}{m}\sum_{j=1}^{m} a_j$, $\mu_b = \frac{1}{L}\sum_{1}^{L} b_k$, $\mu_b' = \frac{1}{L}\sum_{1}^{L} b_k' \neq \mu_b$ one may work out

$$E_R(z_i) = C_i y_i + (1-C_i)(y_i\mu_a + \mu_b)$$

$$E_R(z_i') = C_i y_i + (1-C_i)(y_i\mu_b + \mu_b')$$

yielding $r_{1i} = (\mu_b' z_i - \mu_b z_i')/(\mu_b' - \mu_b)$.

Repeating this exercise entirely once again one may work out a second independent observation r_{2i} distributed identically as r_{1i} to derive (A) $r_i = \frac{1}{2}(r_{1i} + r_{2i})$ with $E_R(r_i) = y_i$ and (B) $v_i = \frac{1}{4}(r_{1i} - r_{2i})^2$ with $E_R(v_i) = V_i = V_R(r_i)$.

The rest follows as in earlier cases, vide Chaudhuri (2011).

References

Chaudhuri, Arijit (2001). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomized finite population. J. Stat. Plan. Inf. 94, 37-42.

________ (2010). Essentials of Survey Sampling, Prentice Hall of India, New Delhi.

________(2011). Randomized Response and Indirect Questioning Techniques in surveys. Chapman & Hall, CRC Press, Taylor & Francis Group, Boca Raton, FL.US.

Chuadhuri, Arijit and Dihidar, Kajal (2009). Estimating means of stigmatizing qualitative and quantitative variables from discretionary responses randomized or direct. Sankhyā B71, 123-136.

Chaudhuri, Arijit & Saha, Amitava (2005). Optional versus compulsory randomized response techniques in complex surveys. J. Stat. Plan. Inf. 135, 516-527.

Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from finite universe. J. Amer. Stat. Assoc, 47, 663-685.

Warner, S.L. (1965). RR: a survey technique for eliminating evasive answer bias. J. Amer. Stat. Assoc. 60, 63-69.