Sankhyā : The Indian Journal of Statistics
2009, Volume 71-B, Part 1, pp. 123-136

Estimating Means of Stigmatizing Qualitative and Quantitative Variables from Discretionary Responses Randomized or Direct

Arijit Chaudhuri
Indian Statistical Institute, Kolkata, India
Kajal Dihidar
Indian Statistical Institute, Kolkata, India

Abstract

The problem addressed here is to unbiasedly estimate the proportion of people bearing a sensitive attribute like habitual tax evasion, gambling, uncontrolled alcoholism etc. in a community and also the means of the amounts involved in meeting the costs or savings/earnings on or through such dubious indulgences. Relevant data are supposed to be gathered from persons sampled in a wide variety of ways permitting direct or randomized responses depending on their personal judgments and views. Unbiased variance estimators are derived as well.

AMS (2000) *subject classification.* Primary 62D05.
Keywords and phrases. Optional randomized responses, unbiased estimation, varying probability sampling.

1 Introduction

Warner (1965) showed us an enlightening way to aptly generate adequate and reliable data by introducing his Randomized Response Technique (RRT), suited to unbiased estimation of the proportion of people indulging in stigmatizing practices like drunk driving, induced abortion, spousal abuse and the like. In such cases procuring direct responses (DR) is rather hard. This spawned an ever increasing variety of alternative devices, extensions to quantitative data, unequal probability sampling even without replacement from the elementary beginning with Simple Random Sampling With Replacement (SRSWR). Optional randomization is also permitted in two different ways; the first permits intentional disclosure of truth overlooking the

stigma, vide Chaudhuri and Mukerjee (1985, 1988) and Chaudhuri and Saha (2005), and the other, vide Mangat and Singh (1994), Singh and Joarder (1997), Gupta et al. (2002), Arnab (2004) and Pal (2008) by allowing a respondent to either reveal the true characteristic or to follow a prescribed randomized response device while keeping the alternative opted for a secret.

In Section 2, we present three specific and typical procedures covering qualitative and quantitative characteristics separately. For brevity we avoid further illustrations. Section 3 provides simulated numerical examples. Section 4 briefly illustrates a motivating application. Section 5 includes discussions and concluding remarks.

2 Optional Randomized Responses: Their Uses

We consider the qualitative case first. Suppose, for a finite population $U = (1, \ldots, i, \ldots, N)$, the value of a sensitive variable y on a person labeled i in U is y_i which is 1 if the individual i bears a stigmatizing feature A, say, and is 0 if he/she bears the complementary characteristic A^c. Let us suppose, an unknowable probability C_i $(0 \leq C_i \leq 1)$ has been assigned by nature to the ith person that he/she gives out the true value of y_i if sampled and addressed. With probability $(1 - C_i)$ he/she is supposed to give out a randomized response (RR) I_i and, independently, a second RR I_i' through the following procedure:

The sampled person i is given two boxes containing identical cards marked A or A^c in proportions $p_j : (1-p_j), j = 1, 2$, such that he/she independently draws one card each from the two boxes labeled $j = 1, 2$, and puts them back. Then, $I_i = 1$ if 'card' type drawn from the first box matches the feature A/A^c, and is 0 otherwise. I_i' is similarly defined for the draw from the second box. Let

$$\begin{aligned} z_i &= y_i \text{ with probability } C_i \\ &= I_i \text{ with probability } (1 - C_i) \end{aligned}$$

and

$$\begin{aligned} z_i' &= y_i \text{ with probability } C_i \\ &= I_i' \text{ with probability } (1 - C_i) \end{aligned}$$

be two independent random variables. Also, the values p_1, p_2 are known to both the respondent and the investigator.

Denoting by E_R and V_R the expectation and variance operators respectively for the above types of randomized response, we get $E_R(z_i) = C_i y_i + (1-$

$C_i)[p_1y_i+(1-p_1)(1-y_i)]$ and $E_R(z_i') = C_iy_i+(1-C_i)[p_2y_i+(1-p_2)(1-y_i)]$ leading to $E_R[(1-p_2)z_i-(1-p_1)z_i'] = (p_1-p_2)y_i$ and $r_i = \frac{(1-p_2)z_i-(1-p_1)z_i'}{p_1-p_2}$. When the constants are chosen so that $p_1 \neq p_2$, $E_R(r_i) = y_i$ for each i in U.

Since $y_i^2 = y_i$, $I_i^2 = I_i$, $(I_i')^2 = I_i'$, it follows that

$$V_R(r_i) = E_R(r_i^2) - E_R(r_i) = E_R[r_i(r_i-1)].$$

Thus $r_i(r_i-1)$ is an unbiased estimator for $V_R(r_i) = V_i$, say, while r_i unbiasedly estimates y_i for every i, if sampled. Writing E_p, V_p as operators for expectation and variance with respect to sampling from U according to any arbitrary sampling design p which assigns to a sample s from U the selection probability $p(s)$, we get

$$E = E_pE_R = E_RE_p \quad \text{and} \quad V = E_pV_R + V_pE_R = E_RV_p + V_RE_p$$

as the overall expectation and variance operators are assumed to have a commutative property.

Let $Y = \sum_{i=1}^{N} y_i$ and let $\theta = \frac{Y}{N}$ be the proportion bearing the feature A in the population. Our interest is to employ an unbiased estimator for θ based on survey data, i.e. r_i for $i \in s$, along with an unbiased variance estimator for that estimator.

We write $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$, $\underline{R} = (r_1, \ldots, r_i, \ldots, r_N)$ generically. Appealing to the standard literature on Survey Sampling, vide Cochran (1977), Chaudhuri and Stenger (2005) and noting

$$e = e(s, \underline{R}) = \frac{1}{N}\sum_{i\in s}\frac{r_i}{\pi_i}, \quad \sum_{s\ni i} p(s) = \pi_i \tag{2.1}$$

which is assumed positive for each i, it follows that $E(e) = \theta$, implying e is an unbiased estimator for θ. Further, let $\pi_{ij} = \sum_{s\ni i,j} p(s)$, which is assumed positive for all i, j $(i \neq j)$, and let $\alpha_i = 1 + \frac{1}{\pi_i}\sum_{j=1, j\neq i}^{N} \pi_{ij} - \sum_{i=1}^{N}\pi_i$. From the general results of Chaudhuri and Pal (2002) and those of Chaudhuri et al. (2000), we have

$$v = v(s, \underline{R}) = \frac{1}{N^2}\left[\sum_{i<}\sum_{j\in s}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{\alpha_i r_i^2}{\pi_i^2} + \sum_{i\in s}\frac{v_i}{\pi_i}\right] \tag{2.2}$$

as an unbiased estimator for $V(e)$. It is obvious that $\alpha_i = 0$ for all i for a sampling scheme with a constant size, say n, for every sample s.

The RR technique presented above is an optional version of the compulsory RR (CRR) device introduced by Warner (1965) modified by Mangat and Singh (1994) to allow an option for direct revelation and by Chaudhuri (2001) to allow the original device to be applicable to a general sampling scheme rather than SRSWR alone.

Next let us extend the above RR device to accommodate a correction to Warner's introduced by Mangat and Singh (1990).

A person labeled i, if sampled, is given a box with a known and verifiable proportion T $(0 < T < 1)$ of cards marked 'T' and the rest marked 'ORR'. The instruction is to randomly draw a card and without divulging its mark to give out the truth without saying so and to follow exactly the earlier ORR device in case a 'T-marked' or 'ORR-marked' card respectively happens to be chosen. The respondent is never to disclose whether a box has been used at all in giving out the response. Let us write the response yielded as

$$\begin{aligned} t_i &= y_i \text{ with probability } T \\ &= z_i \text{ with probability } (1-T). \end{aligned}$$

Further, let

$$\begin{aligned} t_i' &= y_i \text{ with probability } T \\ &= z_i' \text{ with probability } (1-T). \end{aligned}$$

Of course t_i and t_i' are independent variables and

$$E_R(t_i) = Ty_i + (1-T)[C_i y_i + (1-C_i)\{p_1 y_i + (1-p_1)(1-y_i)\}]$$

and

$$E_R(t_i') = Ty_i + (1-T)[C_i y_i + (1-C_i)\{p_2 y_i + (1-p_2)(1-y_i)\}].$$

Then, $E_R[(1-p_2)t_i - (1-p_1)t_i'] = (p_1-p_2)y_i$. Hence we get $r_i = \frac{(1-p_2)t_i-(1-p_1)t_i'}{p_1-p_2}$ for each i, $E_R(r_i) = y_i$, $v_i = r_i(r_i-1)$ is an unbiased estimator for $V_R(r_i) = V_i$ and generically e and v in (2.1) and (2.2) yield respectively an unbiased estimator for θ and an unbiased estimator for $V(e)$. Admittedly, r_i and v_i may both be negative and though $\theta \in [0,1]$, e may go beyond $[0,1]$ and v also may be negative but $E(v)$ being a variance cannot be so.

The well-known 'unrelated model' (the so-called URL) developed by Horvitz et al. (1967) and Greenberg et al. (1969) in its CRR form may

be supposed to have its ORR version with Chaudhuri's (2001) modification for applicability with general sampling schemes to be treated as follows:

Let x denote an innocuous variable taking values x_i for i in U such that, for example, it is 1 if i 'prefers Fine Arts to Music' and 0 otherwise. Let the ith person, if sampled, without divulging his/her option report the y_i-value with probability C_i and with probability $(1 - C_i)$ use two boxes with respective proportions $p_j : (1 - p_j), j = 1, 2$, of cards marked '$y$' and '$x$' respectively to yield the responses as, using the first box

$$\begin{aligned} z_i &= y_i \text{ with probability } C_i \\ &= I_i \text{ with probability } (1 - C_i) \end{aligned}$$

and independently, using the second box,

$$\begin{aligned} z_i' &= y_i \text{ with probability } C_i \\ &= I_i' \text{ with probability } (1 - C_i) \end{aligned}$$

with obvious connotations for I_i and I_i' such that we may work out

$$E_R(z_i) = C_i y_i + (1 - C_i)[p_1 y_i + (1 - p_1)x_i]$$

and

$$E_R(z_i') = C_i y_i + (1 - C_i)[p_2 y_i + (1 - p_2)x_i].$$

Then, $E_R[(1-p_2)z_i-(1-p_1)z_i'] = (p_1-p_2)y_i$, and choosing $p_1 \neq p_2$ it follows that $r_i = \frac{(1-p_2)z_i-(1-p_1)z_i'}{p_1-p_2}$, $E_R(r_i) = y_i$ and $v_i = r_i(r_i - 1)$ is an unbiased estimator for $V_R(r_i) = V_i$, say.

Hence one gets e and v of (2.1) and (2.2) analogously.

Introducing Mangat and Singh's (1990) modification using T and hence t_i, t_i' generically, is a simple matter. We omit further elaborations.

Any other RR device covering qualitative characteristics involving 1/0 response can be similarly covered. But other devices do not yield 1/0 response only amenable to the above and demand separate treatments. In our opinion every RR Technique (RRT, say) demands a separate treatment. A general formulation seems possible. We avoid this to save complications.

Finally we present our procedure covering quantitative characteristics permitting any real values of y_i for i in U. The parameter θ now denotes the finite population mean of y to be estimated.

Suppose a person labeled i, if sampled, may secretly exercise the option with unknown probability C_i to give out the response as the true value

y_i. With probability $(1 - C_i)$ he/she is to draw randomly from one box a card numbered one of $a_1, \ldots, a_j, \ldots, a_m$ with a mean $\mu_a = \frac{1}{m}\sum_{j=1}^{m} a_j = 1$, say, a_j and independently and randomly from a second box one of the numbers $b_1, \ldots, b_k, \ldots, b_L$ with a mean $\mu_b = \frac{1}{L}\sum_{k=1}^{L} b_k$, as say b_k. The person is independently to repeat a similar exercise with one box similar to the earlier first box with a_j's with mean $\mu_a = 1$ but a second similar box with numbers $b'_1, \ldots, b'_k, \ldots, b'_L$ with mean $\mu'_b = \frac{1}{L}\sum_{k=1}^{L} b'_k$, such that $\mu'_b \neq \mu_b$. No use of a box is to be disclosed to the interviewer. Then with probability $(1 - C_i)$ the two responses from the ith person are respectively, defined as

$$I_i = a_j y_i + b_k \quad \text{and} \quad I'_i = a_k y_i + b'_a,$$

with b_k and b'_a as drawn from the boxes with b_k's and b'_k's. Then let

$$\begin{aligned} z_i &= y_i \text{ with probability } C_i \\ &= I_i \text{ with probability } (1 - C_i) \end{aligned}$$

and independently, using the second box,

$$\begin{aligned} z'_i &= y_i \text{ with probability } C_i \\ &= I'_i \text{ with probability } (1 - C_i). \end{aligned}$$

Then, we get

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

and

$$E_R(z'_i) = C_i y_i + (1 - C_i)[y_i \mu_a + \mu'_b] = C_i y_i + (1 - C_i)[y_i + \mu'_b].$$

It follows, on taking $\mu'_b \neq \mu_b$, that $E_R(\mu'_b z_i - \mu_b z'_i) = (\mu'_b - \mu_b) y_i$ and hence

$$r_{1i} = \frac{\mu'_b z_i - \mu_b z'_i}{\mu'_b - \mu_b} \quad \text{satisfies} \quad E_R(r_{1i}) = y_i.$$

But in order to facilitate unbiased variance estimation we need to repeat the above exercises once again generating independently one random variable z^*_i

distributed identically as z_i and another, say z_i'' distributed identically as z_i'. Then, we may generate a new random variable

$$r_{2i} = \frac{\mu_b' z_i^* - \mu_b z_i''}{\mu_b' - \mu_b} \quad \text{with} \quad E_R(r_{2i}) = y_i$$

where r_{2i} is independent of $r_{1i}\ \forall i \in U$. Then, $r_i = \frac{1}{2}(r_{1i}+r_{2i})$ has $E_R(r_i) = y_i$ and $v_i = \frac{1}{4}(r_{1i} - r_{2i})^2$ has $E_R(v_i) = V_R(r_i) = V_i$, say.

Hence one may construct e to estimate $\frac{1}{N}\sum y_i$ and v to estimate $V(e)$, using formulae analogous to (2.1) and (2.2). Thus, in dealing with the quantitative case four responses are needed while for the qualitative case only two responses suffice for each sampled person.

The RRT employed is obviously not simple enough but may be applicable to respondents adequately intelligent and motivated.

3 Illustrative Simulation-based Findings

We consider a fictitious population of $N = 117$ people with last month's household expenses (E) in an appropriate currency as values of y which is 1 if a person is a habitual tax-evader and 0 otherwise and values of x which is 1 if the person prefers cricket to football and 0 otherwise as is considered in Chaudhuri et al. (2009). Corresponding to each of the persons in that population, we consider one more variable about the person's last month's expenses on purchase of alcohol (F). These values are displayed in Table 1 below. Using E values as size measures for these N people and the corresponding normed size-measures namely p_i's $(0 < p_i < 1, i = 1, \ldots, 117)$ we draw from them a sample of $n = 25$ people employing the following scheme as considered by Chaudhuri and Pal (2002). By dint of Brewer's (1963) scheme on the 1st draw we choose a person labeled i with a probability $\frac{p_i(1-p_i)}{1-2p_i}$ and on the second draw we choose a person labeled $j(\neq i)$ from the remaining $N - 1 = 116$ people with the probability $\frac{p_j}{1-p_i}$. Writing $D = \sum_{i=1}^{N} \frac{p_i}{1-2p_i}$, from Brewer (1963) we know the inclusion probability of i and that of the pair $(i, j), i \neq j$ in this sample in 2 draws as respectively

$$\pi_i(2) = 2p_i \quad \text{and} \quad \pi_{ij}(2) = \left[\frac{2p_ip_j}{1+D}\right]\left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right).$$

TABLE 1. A FICTITIOUS POPULATION OF 117 PERSONS.

Serial No. i	y_i	x_i	E_i	F_i	Serial No. i	y_i	x_i	E_i	F_i
1	1	1	2891.31	492.31	61	0	1	2636.53	452.58
2	1	1	4261.13	722.69	62	1	0	1344.76	232.38
3	1	0	2262.45	0	63	1	1	1544.81	0
4	1	1	2530.49	424.09	64	1	1	1255.77	0
5	1	1	2430.49	413.75	65	1	0	1328.88	228.24
6	1	1	4226.83	722.85	66	1	1	3258.28	560.34
7	1	0	3270.41	559.66	67	1	1	2740.52	464.74
8	1	1	1179.95	204.70	68	1	0	4298.5	732.49
9	1	1	1902.73	0	69	1	1	2185.70	377.70
10	0	0	1482.09	0	70	1	0	251.27	42.54
11	1	1	1480.36	250.44	71	1	1	3065.67	523.67
12	0	1	250.9	47.80	72	0	1	1194.98	0
13	1	0	2255.33	0	73	0	1	179.98	0
14	0	1	2525.85	424.70	74	1	1	3845.06	651.45
15	1	1	1241.19	215.12	75	1	0	1188.66	0
16	1	0	1256.66	0	76	0	1	189.36	25.82
17	1	1	2194.89	374.59	77	0	0	1247.3	0
18	1	1	3187.48	540.80	78	0	1	5004.93	855.39
19	0	1	193.65	33.38	79	1	0	1505.03	249.29
20	1	0	1669.54	285.67	80	1	1	3240.26	554.56
21	1	1	3074.11	523.67	81	1	1	3254.33	548.27
22	1	1	4187.81	700.05	82	1	0	334.97	56.20
23	1	0	1264.92	227.93	83	1	1	1242.27	208.06
24	1	1	3196.59	541.03	84	1	1	4181.9	0
25	1	1	3354.57	568.83	85	1	1	187.78	30.37
26	1	1	2717.12	459.06	86	1	1	3242.91	543.94
27	1	1	2927.63	500.67	87	1	1	4334.62	734.94
28	1	0	4147.14	700.42	88	1	1	2575.97	436.85
29	1	1	3385.06	571.10	89	1	1	2608.09	446.20
30	1	1	2644.63	0	90	1	1	4703.93	809.93
31	1	0	2495.64	0	91	1	1	1940.05	337.61
32	1	1	4400.64	756.44	92	1	1	2724.16	459.22
33	1	1	3284.96	562.61	93	1	1	3199.71	536.66
34	1	1	1334.98	226.23	94	1	1	1241.56	203.21

TABLE 1. CONTINUED

Serial No. i	y_i	x_i	E_i	F_i	Serial No. i	y_i	x_i	E_i	F_i
35	1	0	1408.34	235.96	95	0	1	1173.01	192.27
36	1	1	1241.83	208.51	96	1	0	1435.06	247.81
37	1	1	4649.75	790.65	97	0	0	251.42	0
38	1	1	2243.53	374.68	98	1	1	3236.45	548.97
39	1	0	1120.97	184.68	99	1	0	1309.49	225.07
40	1	0	1296.67	220.31	100	1	1	3247.36	0
41	1	1	2878	0	101	1	0	1271.32	209.80
42	1	1	1268.51	0	102	1	1	208.24	27.95
43	1	0	1258.95	212.02	103	1	1	246.96	39.35
44	1	1	2990.47	506.01	104	0	0	1474.4	255.89
45	1	0	1299.93	222.41	105	1	1	2430.23	417.67
46	0	1	205.55	35.79	106	1	1	1148.49	191.86
47	1	1	1245.97	216.11	107	1	1	640.08	101.62
48	1	1	1241.24	209.14	108	1	1	3942.96	670.74
49	0	1	195.59	34.77	109	1	1	2202.25	0
50	1	0	2260.59	379.27	110	0	1	241.63	31.99
51	0	1	242.99	39.17	111	1	1	4191.92	706.51
52	0	1	195.08	36.56	112	1	0	4269.03	726.91
53	1	1	3194.31	542.93	113	1	1	2742.73	466.04
54	0	0	2307.38	0	114	0	1	542.3	0
55	1	1	4842.01	823.37	115	1	0	1546.3	254.72
56	1	1	2904.35	0	116	0	0	1478.00	260.68
57	1	1	3154.77	544.98	117	0	1	789.00	134.62
58	1	1	2191.78	372.80					
59	1	1	2241.53	375.12					
60	1	1	1241.82	0					

These first two draws are next followed by $n-2=23$ draws from the remaining $N-2=115$ people in the population by simple random sampling without replacement (SRSWOR).

From Seth's (1966) works we know that the inclusion probability of i and that of the pair $(i,j), i \neq j$ in a resulting sample of size n are

$$\pi_i(n) = \frac{1}{(N-2)}\left[(n-2) + (N-n)\pi_i(2)\right]$$

and

$$\begin{aligned}\pi_{ij}(n) &= \pi_{ij}(2) + \left(\frac{n-2}{N-2}\right)(\pi_i(2) + \pi_j(2) - 2\pi_{ij}(2)) \\ &\quad + \left(\frac{n-2}{N-2}\right)\left(\frac{n-3}{N-3}\right)(1 - \pi_i(2) - \pi_j(2) + \pi_{ij}(2))\,.\end{aligned}$$

In our formulae for e and v these $\pi_i(n)$ and $\pi_{ij}(n)$ will be used. We also take $cv = 100\dfrac{\sqrt{v}}{e}$ to be the coefficient of variation – the smaller it is the better the estimate.

For this population, $\theta = \dfrac{1}{117}\displaystyle\sum_{i=1}^{117} y_i = 0.81$ and $\bar{F} = \dfrac{1}{117}\displaystyle\sum_{i=1}^{117} F_i = 304.47$.

For every person i, we independently draw a random number from $(0, 1)$ rounded up to 2 decimal places and call them as C_i, for $i = 1, \ldots, 117$. Then Tables 2, 3, 4 and 5 show some of our survey results based respectively on the RRT's by Warner (1965), Mangat and Singh (1990), Greenberg et al.'s (1969) URL model and the quantitative model considered in our present work with brief specifications as below.

Comment: These illustrations reveal that in spite of an undesirable possibility of finding situations yielding negative values of e and of v, such contingencies arise rather infrequently to our relief. Moreover accuracy in estimation is rather tolerably well-maintained with values of cv turning out to be within 30% and often turning out much below this level.

Table 2. Some results based on ORR with Warner's (1965) RRT
$p_1 = 0.4$, $p_2 = 0.3$.
Total Number of replicated samples = 1000.
Number of samples giving negative values of e = 79.
Number of samples giving negative value of v = 0.

Replicate sl. no.	Value of e	Value of $\sqrt{v}$	cv
36	0.80	0.28	35.0
395	0.83	0.26	31.3
671	0.98	0.37	37.8
858	0.87	0.26	29.9

4 An Instructive Case in Point

An investigator while embarking on a randomized response technique out of an apprehension that a respondent given to stigmatizing propensities may not take it as a friendly gesture to agree to implement a randomization device explained to him/her may take a bolder step leaving an option to give out the truth if so desired. But he/she may not be brave enough to ask for straight-forward truthful responses. So giving him/her an option either for a DR or RR without divulging which course is actually adopted seems quite worthy of attention. How this device works has been exemplified in Section 3. This, we believe, may often be put to practice.

TABLE 3. SOME RESULTS BASED ON ORR WITH MANGAT AND SINGH'S (1990) RRT
$p_1 = 0.4$, $p_2 = 0.3$, $T = 0.2$.
TOTAL NUMBER OF REPLICATED SAMPLES = 100.
NUMBER OF SAMPLES GIVING NEGATIVE VALUES OF $e = 5$.
NUMBER OF SAMPLES GIVING NEGATIVE VALUE OF $v = 0$.

Replicate sl. no.	Value of e	Value of $\sqrt{v}$	cv
25	0.92	0.28	30.4
55	0.86	0.29	33.7
61	0.72	0.08	11.1
79	0.91	0.26	28.6

TABLE 4. SOME RESULTS BASED ON ORR WITH GREENBERG ET AL.'S (1969) RRT
$p_1 = 0.45$, $p_2 = 0.37$.
TOTAL NUMBER OF REPLICATED SAMPLES = 100.
NUMBER OF SAMPLES GIVING NEGATIVE VALUES OF $e = 3$.
NUMBER OF SAMPLES GIVING NEGATIVE VALUE OF $v = 0$.

Replicate sl. no.	Value of e	Value of $\sqrt{v}$	cv
2	0.88	0.06	6.8
85	0.99	0.31	31.3

5 Discussion and Concluding Remarks

Mangat and Singh (1994), Singh and Joarder (1997) and Gupta et al. (2002) considered essentially the ORR approach of this paper. But they

restricted themselves to Simple Random Sampling With Replacement (SRSWR) and the RR's on quantitative variables derived by Eichorn and Hayre's (1983) scrambling device involving multiplication of the true value by a random variable with a specified distribution and known mean and variance. In each it was assumed that there exists a section of the population with an unknown proportion C $(0 \leq C \leq 1)$ of people willing to give out the true value rather than opt to give an RR. Arnab (2004) and Pal (2008) applied the same approach allowing unequal probability sampling. The latter permitted each respondent to have his/her own probability of opting for a direct response. The former vehemently criticised those who assumed a common value C as above for the probability to opt for a direct response. But he presents no results allowing this probability to vary from person to person.

TABLE 5. SOME RESULTS BASED ON ORR WITH QUANTITATIVE RRT OF THE PRESENT WORK

$\underline{a} = (0.935, 0.759, 0.764, 1.124, 1.172, 1.048, 0.817, 1.196, 1.223, 0.923)$

$\underline{b} = (-42, 57, 195, -78, 90, -21, -84, 31, 229, 42, 67, -17)$

$\underline{b'} = (134, 252, -56, -27, 9, 5, -21, 64, 246, 77, -117, 83)$

TOTAL NUMBER OF REPLICATED SAMPLES = 100.

Replicate sl. no.	Value of e	Value of $\sqrt{v}$	cv
1	405.4	59.9	14.8
4	340.9	59.9	17.6
15	262.2	48.9	18.6
59	289.1	51.9	18.0
72	504.4	71.7	14.2
91	290.4	61.5	21.2

Each of the above researchers discussed estimating a common C or person specific C_i's. In several of the above ORR-related works comparison of estimators based on ORR's versus the corresponding CRR's has been presented on deriving variance formulae for estimators.

In the present work our emphasis is on unbiased estimation of the variances of estimators for the proportion or the population mean. Variance formulae are avoided especially because C_i's are left unestimated. The question whether an ORR is a better option than a CRR is left unanswered. Further research is invited along this direction.

Finally, we reiterate that SRSWR is impractical in large-scale surveys. In a general survey, sampling involves selection with unequal probabilities

and out of several items of interest only a few may be stigmatizing. To cover both the innocuous and the sensitive ones a common sample may be serviceable in applying our recommendations.

Our numerical applications through simplistic simulations in Section 3 show how our illustrated methods may fare in practice. It is well-known that the literature on sample surveys through DR's hardly gives clues to compare among estimators based on general sampling designs. So, it is futile to venture a way out to cover RR-based results. Hence, we view it gratifying enough to work out procedures to provide standard error estimates as well as estimated coefficients of variation as we have done.

Acknowledgement. We gratefully acknowledge the help received from the referees which led to an improved version of the manuscript, including the addition of Section 4.

References

ARNAB, RAGHUNATH (2004). Optional randomized response techniques for complex designs. *Biometrical Journal.* **46**, 114-124.

BREWER, K. R. W. (1963). A model of systematic sampling with unequal probabilities. *Aust. J. Statist.* **5**, 5-13.

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

CHAUDHURI, ARIJIT, ADHIKARY, ARUN KUMAR and DIHIDAR, SHANKAR (2000). Mean square error estimation in multi-stage sampling. *Metrika.* **52**, 115-131.

CHAUDHURI, ARIJIT, CHRISTOFIDES, T. C. and SAHA, AMITAVA (2009). Protection of privacy in efficient application of randomized response techniques. *Statistical Methods and Applications.* **18**, 389-418.

CHAUDHURI, ARIJIT and MUKERJEE, RAHUL (1985). Optionally randomized response techniques. *Calcutta Statist. Assoc. Bull.* **34**, 225-229.

CHAUDHURI, ARIJIT and MUKERJEE, RAHUL (1988). *Randomized response: Theory and techniques.* Marcel Dekker. New York.

CHAUDHURI, ARIJIT and PAL, SANGHAMITRA (2002). On certain alternative mean square error estimators in complex surveys. *J. Statist. Plann. Inference.* **104**, 363–375.

CHAUDHURI, ARIJIT and SAHA, AMITAVA (2005). Optional versus compulsory randomized response techniques in complex surveys. *J. Statist. Plann. Inference.* **135**, 516-527.

CHAUDHURI, ARIJIT and STENGER, HORST (2005). *Survey Sampling: Theory and Methods.* Taylor and Francis. New York.

COCHRAN, W. G. (1977). *Sampling Techniques.* John Wiley & Sons, New York.

EICHORN, B. H. and HAYRE, L. S. (1983). Scrambled RR method for obtaining sensitive quantitative data. *J. Statist. Plann. Inference.*, **7**, 307-316.

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

Gupta, Sat, Gupta, Bhisam and Singh, Sarjinder (2002). Estimation of sensitivity level of personal interview survey questions. *J. Statist. Plann. Inference.*, **100**, 239-247.

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

Mangat, N. S. and Singh, Ravindra (1990). An alternative randomized response procedure. *Biometrika.* **77**, 439-442.

Mangat, N. S. and Singh, Sarjinder (1994). Optional randomized response model. *J. Indian Statist. Assoc.* **32**, 71-75.

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

Seth, G. R. (1966). On estimators of variance of estimate of population total in varying probabilities. *J. Indian Society of Agricultural Statistics.* **18**, 52-56.

Singh, Sarjinder and Joarder, A. H. (1997). Optional randomized response technique for sensitive quantitative variable. *Metron.* **55**, 151-157.

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

Arijit Chaudhuri
Applied Statistics Unit
Indian Statistical Institute
203 B.T. Road, Kolkata 700 108, India
E-mail: arijitchaudhuri@rediffmail.com

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

Paper received August 2009; revised September 2009.