Metrika (1992) 39:341–357

Randomized Response: Estimating Mean Square Errors of Linear Estimators and Finding Optimal Unbiased Strategies

A. Chaudhuri [1]

Summary: General procedures are described to generate quantitative randomized response (RR) required to estimate the finite population total of a sensitive variable. Permitting sample selection with arbitrary probabilities a formula for the mean square error (MSE) of a linear estimator of total based on RR is noted indicating the simple modification over one that might be based on direct response (DR) if the latter were available. A general formula for an unbiased estimator of the MSE is presented. A simple approximation is proposed in case the RR ratio estimator is employed based on a simple random sample (SRS) taken without replacement (WOR). Among sampling strategies employing unbiased but not necessarily linear estimators based on RR, certain optimal ones are identified under two alternative models analogously to well-known counterparts based on DR, if available. Unlike Warner's (1965) treatment of categorical RR we consider quantitative RR here.

Key words and phrases: Finite population, Linear estimator, Mean square error estimation, Randomized response, Unbiased optimal strategies, Varying probability sampling.

1 Introduction

Suppose $U = (1, \dots, i, \dots, N)$ denotes a finite population of N individuals. Our problem is to estimate $Y = \sum Y_i$, where Y_i is the value for the unit labelled i of a sensitive variable y. For example, y may denote amount spent on gambling or number of days of drunken driving last month etc. A sample s is to be chosen with an arbitrary probability $p(s)$ employing a design p. For a sampled person i it is assumed that the true value Y_i cannot be determined because it relates to a sensitive issue. Instead, following Warner's (1965) pioneering work it is decided to obtain an RR denoted R_i for every i in s. As described by Chaudhuri (1987) and Chaudhuri and Mukherjee (1988) two procedures to procure RR are as follows.

[1] Dr. Arijit Chaudhuri, Indian Statistical Institute, 203, Barrackpore Trunk Road, Calcutta – 700035, India.

Procedure 1: Two vectors $\boldsymbol{a} = (a_1, \ldots, a_F)$, $\boldsymbol{b} = (b_1, \ldots, b_T)$ of F and T real numbers respectively are fixed taking F and T as two arbitrary positive integers. Known means (variances) for them are respectively μ_a, μ_b (σ_a^2, σ_b^2). A sampled person i is required to independently draw two random numbers j $(1 \le j \le F)$ and k $(1 \le k \le T)$, choose corresponding two numbers a_j and b_k from $\boldsymbol{a}$, $\boldsymbol{b}$ and give the RR as

$$Z_i = a_j Y_i + b_k \ .$$

This is done 'independently' by respective persons sampled. Of course the respondent must not disclose to the interviewer the particular (j, k) and hence (a_j, b_k) actually chosen.

But denoting by E_r the operator for expectation with respect to this or any other random experiment to produce an RR, we may note that

$$E_r(a_j) = \frac{1}{F} \sum a_j = \mu_a \ , \quad E_r(b_k) = \frac{1}{T} \sum b_k = \mu_b \ .$$

So, it follows that

$$E_r(Z_i) = \mu_a Y_i + \mu_b \ .$$

Since μ_a and μ_b are known, as $\boldsymbol{a}$ and $\boldsymbol{b}$ are pre-assigned, it follows that

$$R_i = (Z_i - \mu_b)/\mu_a \ , \quad \text{assuming} \quad \mu_a \neq 0 \ ,$$

may be taken as a transformed RR such that

$$E_r(R_i) = Y_i \ , \quad i = 1, \ldots, N \ . \tag{1.1}$$

Also, if we denote by V_r and C_r the operators for taking variance and covariance with respect to randomization experiment, then one gets

$$V_r(R_i) = Y_i^2(\sigma_a^2/\mu_a^2) + (\sigma_b^2/\mu_a^2) \ , \quad i = 1, \ldots, N$$

$$C_r(R_i, R_j) = 0 \ , \quad i \neq j \ .$$

Procedure 2: (cf. Eriksson 1973). First anticipate the possible range of Y_i, $i = 1, \ldots, N$, and then choose a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ of J (arbitrarily chosen positive integer) real numbers θ_j with their range covering that of Y_i's, $i = 1$,

$\ldots, N$. Next choose a number C $(0<C<1)$ and J numbers q_j $(0<q_j<1, j=1, \ldots, J, \sum q_j = 1-C)$. These choices of θ_j, q_j, C are disclosed to the respondents. A sampled person i, 'independently' of one another is then requested to report a number Z_i, $i \in s$. The number Z_i is determined by the respondent labelled i through a random experiment. This experiment is required to produce one of $J+1$ distinct outcomes with respective probabilities q_j, $j=1,\ldots,J$ and C. For example, one may use a box containing tickets marked "True value" and numbers θ_j respectively in proportions C and q_j $(j=1,\ldots,J)$. Accordingly, Z_i is assigned one of the values θ_j $(j=1,\ldots,J)$ or Y_i. Thus, from a sampled person labelled i, RR is

$$\begin{aligned} Z_i &= \theta_j \quad \text{with probability} \quad q_j \ , \quad j=1,\ldots,J \\ &= Y_i \quad \text{with probability} \quad C \ . \end{aligned}$$

Then, $E_r(Z_i) = C Y_i + \sum q_j \theta_j$.

The respondent is to report only Z_i but not the outcome of the experiment to the interviewer. Since C, q_j, θ_j are pre-assigned, one may consider the transformed RR in this case as

$$R_i = (Z_i - \sum q_j \theta_j)/C \ .$$

It may be checked that

$$E_r(R_i) = Y_i \ , \quad V_r(R_i) = \frac{1-C}{C}\left[Y_i^2 - 2Y_i \frac{\sum q_j \theta_j}{1-C} + \frac{1}{C}\frac{\sum q_j \theta_j^2}{(1-C)} \right]$$

$i = 1,\ldots,N$ and $C_r(R_i, R_j) = 0$, $i \neq j$.

For both the procedures, $V_r(R_i)$ is a quadratic in Y_i with known coefficients.

In general, therefore, we shall assume that it is possible to adopt a 'random device' which a respondent i may implement to make a randomized response, which if necessary may be suitably transformed to yield a quantity R_i such that

$$E_r(R_i) = Y_i \ , \quad V_r(R_i) = \alpha_i Y_i^2 + \beta_i Y_i + \psi_i = V_i \ , \quad \text{say}$$

with α_i, β_i, ψ_i as known numbers, and

$$C_r(R_i, R_j) = 0 \ , \quad i \neq j \ .$$

Assuming that $(1+\alpha_i) \neq 0$, one may estimate V_i 'unbiasedly' by

$$v_i = (\alpha_i R_i^2 + \beta_i R_i + \psi_i)/(1+\alpha_i)$$

because one may check that

$$E_r(v_i) = V_i \ , \quad i = 1, \ldots, N \ .$$

For convenience we shall suppose that for the population U, the following random vector

$$\boldsymbol{R} = (R_1, \ldots, R_i, \ldots, R_N)$$

incorporating the potential RR's is defined, corresponding to the vector

$$\boldsymbol{Y} = (Y_1, \ldots, Y_i, \ldots, Y_N)$$

of unknown true values of y. In the next section we consider estimators of Y, their MSE's and more importantly unbiased estimators of the latter.

2 Linear RR Estimators and Estimators of their MSE's

For a sample s drawn according to a design p let us first define the indicator functions:

$$I_{si} = 1\,(0) \quad \text{if} \quad i \in s \quad (s \not\ni i)$$

$$I_{sij} = 1\,(0) \quad \text{if} \quad i,j \in s \quad (s \not\ni i,j) \ .$$

Let b_{si}, d_{sij} denote real numbers free of $\boldsymbol{Y}$ and $\boldsymbol{R}$ such that $b_{si} = 0$ if $s \not\ni i$ and $d_{sij} = 0$ if $s \not\ni i,j$. Furthermore, let for a sample s drawn according to p,

$$t_b = \sum Y_i b_{si}$$

which could be taken as an estimator for Y if DR's were available but not usable in the present context. However, first supposing DR's were available let us denote the MSE of t_b as

$$M(t_b) = E_p(t_b - Y)^2 = \sum_i \sum_j d_{ij} Y_i Y_j$$

writing $d_{ij} = E_p(b_{si} - 1)(b_{sj} - 1)$.

Let d_{sij} satisfy the condition

$$E_p(d_{sij}) = d_{ij} \ .$$

Then, $m(t_b) = \sum \sum d_{sij} Y_i Y_j$ becomes a design-unbiased estimator of $M(t_b)$.

Further, let

$$\pi_i = \sum_s p(s) I_{si} (>0) \ , \quad \text{denote the inclusion-probability of } i$$

$$\pi_{ij} = \sum_s p(s) I_{sij} (>0) \ , \quad \text{denote the inclusion-probability of } i,j \ .$$

A well-known example of t_b is the Horvitz-Thompson (1952) estimator

$$\bar{t} = \sum \frac{Y_i}{\pi_i} I_{si} \ , \quad \text{taking} \quad b_{si} = \frac{I_{si}}{\Pi_i} \ .$$

For this, $d_{ij} = E_p \left(\frac{I_{si}}{\Pi_i} - 1 \right) \left(\frac{I_{sj}}{\Pi_j} - 1 \right) = \frac{\Pi_{ij}}{\Pi_i \Pi_j} - 1$ and

$$M(\bar{t}) = \sum_i \sum_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) Y_i Y_j \ ,$$

noting $E_p(I_{si}) = \Pi_i$, $E_p(I_{si} I_{sj}) = E_p(I_{sij}) = \Pi_{ij}$.

This $M(\bar{t})$ is also the variance of $\bar{t}$ because $E_p(\bar{t}) = Y$.

As was first noted by Vijayan (1975) and later, more fully discussed by Rao and Vijayan (1977) and Rao (1979), a procedure to find uniformly non-negative quadratic unbiased estimator (UNNQUE) for $M(t_b)$ is to proceed as follows.

Given a t_b as above, it is often possible to find non-zero numbers W_i ($i = 1, \ldots, N$) such that

"the value of $M(t_b)$ equals zero"

if Y_i is assigned the value CW_i, $i = 1, \ldots, N$ taking C as a non-zero constant. We shall refer to this condition on t_b as 'condition A'. For every such t_b Rao and Vijayan (1977) have shown that it is possible to write $M(t_b)$ in the form

$$M(t_b) = -\sum_{i<j}\sum d_{ij} W_i W_j \left(\frac{Y_i}{W_i} - \frac{Y_j}{W_j}\right)^2$$

for an arbitrary $\boldsymbol{Y} = (Y_1, \ldots, Y_i, \ldots, Y_N)$.

If t_b is taken as $\bar{t}$, for example, in the case where every sample has a fixed number (say, n) of units, each distinct, then, the choice

$$W_i = \pi_i$$

leads to (i) $M(\bar{t}) = 0$, if $Y_i = C\pi_i$, $i = 1, \ldots, N$ as is easy to see noting that $\sum \pi_i = n$ and to (ii) the form

$$M(\bar{t}) = -\sum_{i<j}\sum \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) \pi_i \pi_j \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 .$$

This of course is the familiar Yates-Grundy form. From Vijayan (1975), Rao et al. (1977) and Rao (1979) it follows that a UNNQUE for $M(t_b)$ in the above case, is 'necessarily of the form'

$$m(t_b) = -\sum_{i<j}\sum d_{sij} W_i W_j \left(\frac{Y_i}{W_i} - \frac{Y_j}{W_j}\right)^2 .$$

In case t_b is taken as $\bar{t}$, a possible choice of d_{sij} is $d_{sij} = \dfrac{1}{\pi_i \pi_j} - \dfrac{1}{\pi_{ij}}$, yielding

$$m(\bar{t}) = -\sum_{i<j}\sum \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}\right) I_{sij} \pi_i \pi_j \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 .$$

This is also the familiar Yates-Grundy variance estimator of $\bar{t}$. One may consult Rao et al. (1977) and Rao (1979) for further examples of t_b subject to 'condition A'.

Bearing these in mind concerning estimation based on Y_i for $i \in s$, we propose analogous procedures as follows when instead of DR only RR is available. We propose the use of the linear estimator for Y given by

$$e_b = e_b(s, \boldsymbol{R}) = \sum R_i b_{si}$$

simply writing R_i for Y_i in the DR estimator t_b.

It follows that (i) $E_r(e_b) = \sum Y_i b_{si} = t_b$

(ii) $M(e_b) = E_pE_r(e_b - Y)^2 = E_pE_r(e_b - t_b)^2 + E_p(t_b - Y)^2 = E_p V_r(e_b) + M(t_b)$

may be taken as the MSE of e_b about Y.

Restricting throughout to t_b subject to 'condition A', we have

$$M(e_b) = \sum V_i E_p(b_{si}^2) - \sum_{i<j}\sum d_{ij} W_i W_j \left(\frac{Y_i}{W_i} - \frac{Y_j}{W_j}\right)^2 .$$

As an 'unbiased' estimator for $M(e_b)$ we propose

$$m(e_b) = \sum v_i b_{si}^2 - \sum_{i<j}\sum d_{sij} W_i W_j \left[\left(\frac{R_i}{W_i} - \frac{R_j}{W_j}\right)^2 - \left(\frac{v_i}{W_i^2} + \frac{v_j}{W_j^2}\right)\right]$$

on noting that

$$E_pE_r m(e_b) = M(e_b) .$$

In particular, we may recommend the use of the RR analogue of Horvitz-Thompson estimator, namely,

$$\bar{e} = \sum \frac{R_i}{\pi_i} I_{si} .$$

For this it is easy to work out

$$M(\bar{e}) = \sum \frac{V_i}{\pi_i} - \sum_{i<j}\sum \left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right) \pi_i\pi_j \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2$$

$$m(\bar{e}) = \sum \frac{v_i}{\pi_i^2} I_{si} - \sum_{i<j}\sum \left(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{ij}}\right) I_{sij}\pi_i\pi_j \left[\left(\frac{R_i}{\pi_i} - \frac{R_j}{\pi_j}\right)^2 - \left(\frac{v_i}{\pi_i^2} + \frac{v_j}{\pi_j^2}\right)\right]$$

and verify that $E_pE_r m(\bar{e}) = M(\bar{e})$.

As another example we consider t_b of the form

$$t_1 = X\frac{\bar{y}}{\bar{x}} ,$$

called the ratio estimator based on a design p for which

$$p(s) = \frac{Q(s)}{M_1} .$$

Here X_i (>0) are known numbers called size-measures, closely associated with Y_i ($i = 1, \ldots, N$), with a total $X = \sum X_i$, $Q(s) = \frac{\sum X_i I_{si}}{X}$, n is the sample-size, $M_g = \binom{N-g}{n-g}$, $g = 0, 1, 2, \bar{y}, \bar{x}$ are sample means of Y_i's and X_i's. When Y_i, $i \in s$ are not available but R_i, $i \in s$ may be gathered, then we recommend the use of

$$e_1 = X\frac{\bar{r}}{\bar{x}} ,$$

writing $\bar{r}$ for the sample mean of R_i's.

For this sampling design selection schemes are given by Lahiri (1951), Midzuno (1952) and Sen (1953) and hence called LMS design, we have

$$M(t_1) = E_p\left(X\frac{\bar{y}}{\bar{x}} - Y\right)^2 = \sum \sum d_{ij} Y_i Y_j$$

with $d_{ij} = \frac{1}{M_1} \sum_s \frac{I_{sij}}{Q(s)} - 1.$

Since $M(t_1)$ equals zero in cae $Y_i = CX_i$, $C \neq 0$, the Rao and Vijayan (1977) alternative form $M(t_1)$ is

$$M(t_1) = \sum_{i<j} \sum \left[1 - \frac{1}{M_1} \sum_s \frac{I_{sij}}{Q(s)}\right] X_i X_j \left(\frac{Y_i}{X_i} - \frac{Y_j}{X_j}\right)^2 .$$

For simplicity we shall write

$$a_{ij} = X_i X_j \left(\frac{Y_i}{X_i} - \frac{Y_j}{X_j}\right)^2 .$$

A simple unbiased estimator for $M(t_1)$ is then

$$m(t_1) = \sum_{i<j}\sum \frac{I_{sij}}{Q(s)}\left[\frac{N-1}{n-1} - \frac{1}{Q(s)}\right] a_{ij}$$

on noting that $\sum_s I_{sij} = M_2$ and $\frac{M_1}{M_2} = \frac{N-1}{n-1}$.

Analogously, it follows that

$$M(e_1) = \sum \frac{V_i}{M_1} \sum_s \frac{I_{si}}{Q(s)} + \sum_{i<j}\sum \left[1 - \frac{1}{M_1}\sum_s \frac{I_{sij}}{Q(s)}\right] a_{ij} \tag{2.1}$$

and

$$m(e_1) = \sum \frac{v_i I_{si}}{Q^2(s)} + \sum_{i<j}\sum \frac{X_i X_j}{Q(s)} I_{sij}\left[\frac{N-1}{n-1} - \frac{1}{Q(s)}\right] \times \left[\left(\frac{R_i}{X_i} - \frac{R_j}{X_j}\right)^2 - \left(\frac{v_i}{X_i^2} + \frac{v_j}{X_j^2}\right)\right]. \tag{2.2}$$

In case e_1 is based on simple random sampling without replacement (SRSWOR) in n draws for which

$$p(s) = \frac{1}{M_0}, \quad \text{for every sample,}$$

(2.1) and (2.2) change respectively into

$$M'(e_1) = \frac{1}{M_0}\left[\sum V_i \sum_s \frac{I_{si}}{Q^2(s)} - \sum_{j<j}\sum a_{ij} b_{ij}\right],$$

writing

$$b_{ij} = \sum_s \, [I_{sij} - Q(s)(I_{si} + I_{sj}) + Q^2(s)]/Q^2(s)$$

and

$$m'(e_1) = \frac{1}{M_0}\left[\frac{N}{n}\sum v_i I_{si}\left(\sum_s \frac{I_{si}}{Q_s(s)}\right) - \frac{N(N-1)}{n(n-1)}\sum_{i<j}\sum I_{sij} X_i X_j b_{ij} \times \left\{\left(\frac{R_i}{X_i} - \frac{R_j}{X_j}\right)^2 - \left(\frac{v_i}{X_i^2} + \frac{v_j}{X_j^2}\right)\right\}\right].$$

As it is laborious to compute b_{ij} we replace the term

$$-\frac{1}{M_0}\sum\sum_{i<j} a_{ij}b_{ij} = E_p(t_1-Y)^2 \quad \text{in} \quad M'(e_1)$$

by its well-known Cochran (1977) approximation which on writing $f=\frac{n}{N}$, is

$$T=\frac{N}{f}\left(\frac{1-f}{N-1}\right)\sum\left(Y_i-\frac{Y}{X}X_i\right)^2 ,$$

and approximate $M'(e_1)$ by

$$M''(e_1)=\frac{1}{M_0}\left(\sum_s \frac{1}{Q^2(s)}\sum_i V_i I_{si}\right)+T .$$

An unbiased estimator for this easily follows as

$$\begin{aligned} m''(e_1) &= \frac{N}{f}(1-f) \\ &\times\left[u(s)-\frac{1}{N-1}\left\{\sum v_i\frac{I_{si}}{f}+\frac{\left(\sum_s v_i I_{si}\right)\sum X_i^2 I_{si}}{Q^2(s)}-2\frac{\sum v_i X_i I_{si}}{Q(s)}\right\}\right] \\ &+\left(\sum_s v_i I_{si}\right)\frac{X^2}{Q^2(s)} . \end{aligned}$$

Here, $u(s)=\frac{1}{n-1}\sum\left(R_i-\frac{\bar{r}}{\bar{x}}X_i\right)^2 I_{si}$;

it may be checked that

$$E_pE_r m''(e_1)=M''(e_1) .$$

In the context of randomized response surveys 'ratio estimator' was earlier employed by Abul-Ela and Abdel-Hamied (1985) who applied Greenberg et al.'s (1971) scheme of sampling. Using our notation the scheme is as follows.

Two independent simple random samples (SRS) of sizes n_1 and n_2 are both taken with replacement (WR). A sampled person i drawn in the jth ($j = 1,2$) sample is requested to implement a 'pre-determined' random device so as to give out the 'true value' Y_i of the sensitive variable y with a pre-assigned probability P_j and with probability $1-P_j$ the value X_i of an 'unrelated or correlated' variable x which is innocuous or at least not as sensitive as y. The resulting RR from jth sample for ith person, say, Z_{ji} has the expectation $E_r(Z_{ji}) = P_j Y_i + (1-P_j)X_i$, $j = 1,2$.

The means $\bar{Y}$ and $\bar{X}$ of y, x are then estimated from the sample means $\bar{z}_j$ of Z_{ji} by, respectively,

$$\hat{\bar{Y}} = [(1-P_2)\bar{z}_1 - (1-P_1)\bar{z}_2]/(P_1 - P_2) \ , \quad \text{taking} \quad P_1 \neq P_2$$

$$\hat{\bar{X}} = [P_2\bar{z}_1 - P_1\bar{z}_2]/(P_2 - P_1)$$

and Y is estimated by the 'ratio estimator'

$$e_2 = X\hat{\bar{Y}}/\hat{\bar{X}} \ , \quad \text{assuming } X \text{ known.}$$

Abul-Ela and Abdel-Hamied (1985) examined its efficiency but did not consider estimating its MSE. Our problem here mainly is estimating MSE. Further differences are in methods of (i) sampling – we need only one sample and (ii) generating RR. Also we need X_i-values fully. So, the two treatments are not amenable to comparison in greater details.

3 Optimal Strategies

In estimating Y using Direct responses when available often a super-population model is postulated concerning Y treating it as a random vector rather than a constant. We shall illustrate two models. In one, Y_i's are supposed to be distributed 'independently' with "means and variances"

$$E_m(Y_i) = \mu_i \ , \quad V_m(Y_i) = \sigma_i^2 \ , \quad i = 1, \ldots, N \ .$$

Postulating this model, the following results from Godambe and Thompson (1977) are well-known.

Among all estimators $t = t(s, Y)$ for Y subject to $E_p(t) = Y$ based on any p, the 'optimal' one is given by

$$t_\mu = \sum \frac{Y_i - \mu_i}{\pi_i} I_{si} + \sum \mu_i \quad \text{(the sum is over } i\text{)}$$

with the property that

$$E_m E_p (t-Y)^2 \geq \sum \sigma_i^2 \left(\frac{1}{\pi_i} - 1 \right) = E_m E_p (t_\mu - Y)^2 \ . \tag{3.1}$$

This t_μ cannot be used if μ_i is 'unknown' as it should be the case in general. But if

$$\mu_i = \beta X_i$$

with β (>0) unknown but X_i (>0) known, then if one employs only a design p_n for which every sample s with $p(s) > 0$ contains only distinct units, n in number, and if in addition, one may employ a still restricted design p_{nX}, say, for which

$$\pi_i = n \frac{X_i}{X} \ , \quad i = 1, \ldots, N \ ,$$

then t_μ based on p_{nX} reduces to the Horvitz-Thompson estimator

$$\bar{t} = \sum \frac{Y_i}{\pi_i} I_{si}$$

which becomes 'optimal' in the sense that

$$E_m E_{p_{nx}} (t-Y)^2 \geq \sum \sigma_i^2 \left(\frac{1}{\pi_i} - 1 \right) = E_m E_{p_{nx}} (\bar{t} - Y)^2 \ . \tag{3.2}$$

Thus, in the 'restricted class' of 'strategies' (p_{nX}, t), the sub-class $(p_{nx}, \bar{t})$ is optimal.

If furthermore,

$\sigma_i \propto X_i$, in addition to $\mu_i \propto X_i$, and one

restricts to designs p_n for which

$\pi_i \propto X_i \infty \sigma_i$, denoted now as $p_{nx\sigma}$,

then a strategy $(p_{nx\sigma}, \bar{t})$ is optimal among strategies (p_n, t) in the sense that

$$E_m E_{p_n}(t-Y)^2 \geq \frac{(\sum \sigma_i)^2}{n} - \sum \sigma_i^2 = E_m E_{p_{nx\sigma}}(\bar{t}-Y)^2 \ . \tag{3.3}$$

If DR's are unavailable and one may use only R_i, $i \in s$, as in Sections 1 and 2, then treating 'not necessarily linear' estimators

$e = e(s, \boldsymbol{R})$ free of R_j for $j \notin s$,

subject to

$E_p(e) = \sum R_i$

considering $\bar{e} = \sum \frac{R_i}{\pi_i} I_{si}$ and

$e_\mu = \sum \frac{R_i - \mu_i}{\pi_i} I_{si} + \sum \mu_i$ based on $\boldsymbol{R}$,

it is possible to modify the results (3.1)–(3.3) into the results (3.1)′–(3.3)′ stated below.

$$E_m E_p E_r (e-Y)^2 \geq \sum \frac{E_m V_i}{\pi_i} + \sum \sigma_i^2 \left(\frac{1}{\pi_i} - 1 \right) = E_m E_p E_r (e_\mu - Y)^2 \tag{3.1$'$}$$

$$E_m E_{p_{nx}} E_r (e-Y)^2 \geq \sum \frac{E_m V_i}{\pi_i} + \sum \sigma_i^2 \left(\frac{1}{\pi_i} - 1 \right) = E_m E_{p_{nx}} E_r (\bar{e} - Y)^2 \tag{3.2$'$}$$

$$E_m E_{p_n} E_r (e-Y)^2 \geq \sum \frac{E_m V_i}{\pi_i} + \frac{(\sum \sigma_i^2)}{n} - \sum \sigma_i^2 = E_m E_{p_{nx\sigma}} E_r (\bar{e} - Y)^2 \ . \tag{3.3$'$}$$

In order to check the results (3.1)′–(3.3)′ one needs to consult the relevant materials in Cassel, Särndal and Wretman (1977), Godambe and Joshi (1965), Godambe and Thompson (1977) and Ho (1980), assume that E_p, E_m, E_r commute and writing

$$h = h(s,R) = e - \bar{e}\ , \quad h_\mu = h_\mu(s,R) = e_\mu - \bar{e}$$

check the following:

(i) $V_r(h_\mu) = 0$

(ii) $V_m(E_r h_\mu) = 0$

(iii) $E_m E_r e_\mu = \mu$

(iv) $E_m E_p E_r(e-Y)^2 = E_m E_p V_r(\bar{e}) + E_m E_p V_r(h) + E_p V_m(E_r\bar{e}) + E_p V_m(E_r h) + E_p(E_m E_r e - \mu)^2 - V_m(Y)$

(v) $E_m E_p E_r(e-Y)^2 = E_m E_p V_r(\bar{e}) + E_p V_m(E_r \bar{e}) + E_p(E_m E_r \bar{e} - \mu)^2 - V_m(Y)$

(vi) $E_m E_p E_r(e_\mu - Y)^2 = E_m E_p V_r(\bar{e}) + E_p V_m(E_r\bar{e}) - V_m(Y)$

$$= \sum \frac{E_m V_i}{\pi_i} + \sum \sigma_i^2 \left(\frac{1}{\pi_i} - 1\right) .$$

Another optimality result concerning DR as follows is available from Godambe and Thompson (1973) under the following alternative model.

Suppose $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_i, \ldots, \phi_N)$ is a real vector of known numbers ϕ_i $(0<\phi_i<1,\ \sum \phi_i = n)$ such that writing $D_i = \dfrac{Y_i}{\phi_i}$, the vector

$$\boldsymbol{D} = (D_1, \ldots, D_i, \ldots, D_N)$$

has an exchangeable distribution i.e. every vector $(D_{i1}, \ldots, D_{ij}, \ldots, D_{iN})$ for a permutation $(i_1, \ldots, i_N)$ of $(1, \ldots, N)$ has the same probability distribution. Denoting by E_π the operator for expectation over this distribution and denoting by $p_{n\phi}$ a sampling design p_n for which

$$\pi_i = \phi_i$$

then, it follows that $(p_{n\phi}, \bar{t})$ is optimal among strategies (p_n, t), subject to $E_p(t) = Y$ in the sense that

$$E_\pi E_{p_n}(t-Y)^2 \geq E_\pi E_{p_{n\phi}}(\bar{t}-Y)^2\ . \tag{3.4}$$

The special case of this model is called the 'random permutation model' for which the vector Y is a vector of fixed constants but a probability distribution for D is postulated by assigning a common probability $\frac{1}{N!}$ to each vector of the form $(D_{i1},\ldots,D_{iN})$ above, with $(i1,\ldots,iN)$ a permutation of $(1,\ldots,N)$. Retaining the same notation E_π for this case the equivalent result (3.4) was proved by Thompson [cf. Rao (1971)] strengthening earlier results by Kempthorne (1969) and Rao (1971). Postulating a similar 'random permutation model' we present below a counterpart of Thompson's result with a few modifications to cover the case of RR surveys when DR's are unavailable.

Let $B=(B_1,\ldots,B_i,\ldots,B_N)$ be a vector of known real numbers with $B=\sum B_i$, $\bar{B}=B/N$ and $D_i'=(Y_i-B_i+\bar{B})/\phi_i$, $i=1,\ldots,N$. Let $D'=(D_1',\ldots,D_N')$ be subject to the 'random permutation model' and the notation E_π be extended for the distribution of D'. Letting $e_B=\sum\frac{(R_i-B_i+\bar{B})}{\pi_i}I_{si}$ and $e_{B\phi}$ as e_B replacing π_i in the latter by ϕ_i, one has then the

Theorem:

$$E_\pi E_{p_{n\phi}} E_r(e-Y)^2 \geq E_\pi E_{p_{n\phi}} E_r(e_{B\phi}-Y)^2 \ .$$

Proof (in outlines only): Letting $h_B=e-e_B$ one has

$$E_p(h_B)=0 \ , \quad E_p C_r(e_B,h_B)=0 \ .$$

It follows on writing V_π for variance over the 'random permutation' modelling, that

$$E_\pi E_p E_r(e-Y)^2 = E_\pi E_p V_r(e)+E_\pi E_p(E_r e-E_\pi Y)^2-V_\pi(Y)$$

$$E_\pi E_p E_r(e_B-Y)^2 = E_\pi E_p V_r(e_B)+E_\pi E_p(E_r e_B-E_\pi Y)^2-V_\pi(Y)$$

$$E_p V_r(e)=E_p V_r(e_B)+E_p V_r(h_B)\geq E_p V_r(e_B)$$

$$E_{p_{n\phi}}(E_r e_{B\phi})=E_r(E_{p_{n\phi}} e_{B\phi})=E_r(\textstyle\sum R_i)=Y \ ,$$

$$E_\pi(Y)=E_\pi \sum \phi_i D_i' = \frac{1}{N!}\sum{}' \sum \phi_i D_{ji}' = \frac{n}{N}\sum D_i' \ .$$

Here $\sum'$ denotes sum over all possible permutations of $(i1, \ldots, iN)$. On checking from Godambe and Thompson's (1973) and Thompson's result (3.4) that

$$E_\pi E_{p_n}(E_r e - E_\pi Y)^2 \geq E_\pi E_{p_{n\phi}}(E_r e_B - E_\pi Y)^2$$

because $E_r e$ can be taken as t, $E_r e_B = \sum \frac{(Y_i - B_i + \bar{B})}{\pi_i} I_{si}$ the result follows on simplification.

Acknowledgement: Thanks are due to two referees for helpful suggestions. The work was done when visiting University of Mannheim, Germany.

References

Abul-Ela AA, Abdel-Hamied SM (1985) A Randomized Response Ratio Estimate from Quantitative Data. Proc Soc Stat Sec Amer Stat Assoc, pp 300–305

Cassel CM, Särndal CE, Wretman JH (1977) Foundations of Inference in Survey Sampling. Wiley, New York

Chaudhuri A (1987) Randomized Response Surveys of Finite Populations: A Unified Approach with Quantitative Data. Jour Stat Plan and Inf. 15:157–165

Chaudhuri A, Mukherjee R (1988) Randomized Response: Theory and Techniques. Marcel Dekker Inc, New York

Cochran WG (1977) Sampling Techniques. Wiley, New York

Eriksson SA (1973) A New Model for Randomized Response. Rev Inter Stat Inst 41:101–113

Godambe VP, Joshi VM (1965) Admissibility and Bayes Estimation in Sampling Finite Populations, I. Ann Math Stat 36:1707–1722

Godambe VP, Thompson ME (1973) Estimation in Sampling Theory with Exchangeable Prior Distributions. Ann Stat 1:1212–1221

Godambe VP, Thompson ME (1977) Robust Near Optimal Estimation in Survey Practice. Bull Inter Stat Inst 47:3, 129–146

Greenberg BG, Kuebler R, Abernathy JR, Horvitz DG (1971) Application of the Randomized Response Technique in Obtaining Quantitative Data. Jour Amer Stat Assoc 66:243–250

Ho EWH (1980) Model – Unbiasedness and the Horvitz-Thompson Estimator in Finite Population Sampling. Aust Jour Stat 22:218–225

Horvitz DG, Thompson DJ (1952) A Generalization of Sampling Without Replacement from a Finite Universe. Jour Amer Stat Assoc 47:663–685

Kempthorne O (1969) Some Remarks on Statistical Inference in Finite Sampling. In: Johnson NL, Smith H Jr (eds) New Developments in Survey Sampling. Wiley, New York, pp 671–695

Lahiri DB (1951) A Method of Sample Selection Providing Unbiased Ratio Estimators. Bull Int Stat Inst 33:2, 133–140

Midzuno H (1952) On the Sampling System with Probabilities Proportionate to Sum of Sizes. Ann Inst Stat Math 3:99–107

Rao CR (1971) Some Aspects of Statistical Inference in Problems of Sampling from Finite Populations. In: Godambe VP, Sprott DA (eds) Foundations of Statistical Inference. Holt, Rinehart and Winston, Toronto, pp 177–202

Rao JNK (1979) On Deriving Mean Square Errors and Their Non-Negative Unbiased Estimators in Finite Population Sampling. Jour Ind Stat Assoc 17:125–136

Rao JNK, Vijayan K (1977) On Estimating the Variance in Sampling with Probability Proportional to Aggregate Size. Jour Amer Stat Assoc 72:579–584

Sen AR (1953) On the Estimator of the Variance in Sampling with Varying Probabilities. Jour Ind Soc Agri Stat 5:2, 119–127

Vijayan K (1975) On Estimating the Variance in Unequal Probability Sampling. Jour Amer Stat Assoc 70:713–716

Warner SL (1965) Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Jour Amer Stat Assoc 60:63–69

Yates F, Grundy PM (1953) Selection without Replacement from within Strate with Probability Proportional to Size. Jour Roy Stat Soc B15:252–261

Received 7 September 1990
Revised version 27 March 1992