COMMUN. STATIST.-THEORY METH., 19(3), 1119-1125 (1990)

VARIANCE ESTIMATION WITH RANDOMIZED RESPONSE

Arijit Chaudhuri

Virginia Polytechnic Insititute
& State University
Blacksburg, VA 24061

Arun Kumar Adhikary

University of Nairobi
Nairobi, Kenya

Key Words and Phrases: finite survey population; multi-stage sampling; probability proportional to size sampling; randomized response; variance estimation.

ABSTRACT

Presenting a general procedure of eliciting a randomized response (RR) from selected persons in order to estimate the total of a sensitive variable related to a finite survey population, we consider two estimators along with variance estimators treating the case of sampling with probabilities proportional to (known) size measures (PPS) with replacement (WR), drawing analogies with multi-stage sampling and note their relative efficacies.

1. INTRODUCTION

Let the values Y_i of a sensitive variable y, defined on a finite survey population $I = (1, ..., i, ..., N)$ of N identifiable and labelled persons be supposedly unavailable through a direct response (DR) survey when one intends to estimate $Y = \sum_{i=1}^{N} Y_i$ on choosing a sample s from I with a probability $p(s)$ according to a design p. Instead, let a randomized response (RR) R_i be available in independent manners from the respective persons i, on request if sampled, in such a way that their expectations, variances, covariances (E_R, V_R, C_R) respectively satisfy $E_R(R_i) = Y_i$, $V_R(R_i) = \alpha_i Y_i^2 + \beta_i Y_i + \theta_i = V_i^2$ (say).

$C_R(R_i,R_j) = 0$, for $i \neq j$ such that $\alpha_i(>0)$, β_i, θ_i are known for every $i = 1,\ldots,N$. for example, a sampled person labelled i may be requested to choose independently at random two tickets numbered $a_j(>0)$ and b_k out of two boxes, presented by the investigator, respectively containing the tickets numbered (1) $A_1, \ldots, A_m$ with mean $\bar{A}$ and variance σ_A^2 and (2) $B_1, \ldots, B_T$ with mean $\bar{B}$ and variance σ_B^2 and report RR as $Z_i = a_jY_i + b_k$ leading to $E_R(Z_i) = \bar{A}Y_i + \bar{B}$ giving $R_i = (Z_i - \bar{B})/\bar{A}$ with $E_R(R_i) = Y_i$, $V_R(R_i) = (\sigma_A^2Y_i^2 + \sigma_B^2)/(\bar{A})^2$ for each i. In the general case it follows that we have $E_R(E_i^2) = V_i^2$ where

$$E_i^2 = (\alpha_i R_i^2 + \beta_i R_i + \theta_i)/(1 + \alpha_i)$$

which is thus an unbiased estimator for V_i^2. Suppose that normed size-measures $p_i(0 < p_i < 1, \sum_1^N p_i = 1)$ are available for the individuals $i = 1, \ldots, N$ and a PPSWR sample s is taken in n draws. Two estimators for Y based on $\underline{R} = (R_1, \ldots, R_i, \ldots, R_N)$ from such a sample are considered in the next section along with their unbiased variance estimators followed by a discussion of their uses in section 3.

2. ESTIMATORS FOR TOTAL AND VARIANCE ESTIMATORS

Method I. Each time a person is selected in the n draws an independent R_i is obtained and the estimator for Y is taken as

$$e_1 = \frac{1}{n}\sum_{k=1}^{n}\frac{r_k}{p_k} \quad \text{giving} \quad E_R(e_1) = \frac{1}{n}\sum_{k=1}^{n}\frac{y_k}{p_k} = t_1$$

say and

$$E(e_1) = E_pE_R(e_1) = E_p(t_1) = Y \tag{2.1}$$

writing $y_k(r_k, p_k)$ as the value of y(RR, normed size-measure) for the person selected on the kth draw, $E(E_p)$ as over-all (design) expectation so that e_1 is unbiased (randomization and design-based) for Y.

Writing $V(V_p)$ as the over-all (design) variance operator and d_k^2 for $V_R(R_i)$ if kth draw yields i, the variance of e_1 may be taken as

$$V(e_1) = E_pE_R(e_1 - Y)^2 = E_pE_R[(e_1 - t_1) + (t_1 - Y)]^2$$

$$= E_p V_R(e_1) + V_p(t_1) = E_p\left(\frac{1}{n^2}\sum_{1}^{n}\frac{d_k^2}{p_k^2}\right) + V_p\left(\frac{1}{n}\sum_{1}^{n}\frac{Y_k}{p_k}\right) \tag{2.2}$$

$$= \frac{1}{n}\sum_{1}^{N}\frac{V_i^2}{p_i} + \frac{1}{n}\left(\sum_{1}^{N}\frac{Y_i^2}{p_i} - Y^2\right)$$

It is easy to check that $E_p E_R(\nu_1) = V(e_1)$ where

$$\nu_1 = \frac{1}{n(n-1)}\sum_{1}^{n}\left(\frac{r_k}{p_k} - \frac{1}{n}\sum_{k=1}^{n}\frac{r_k}{p_k}\right)^2 . \tag{2.3}$$

Thus ν_1 may be taken as an unbiased estimator for $V(e_1)$. The results (2.1) - (2.3) parallel the corresponding results given by Raj (1968) in multi-stage sampling where the primary sampling units (PSU) are chosen by PPSWR method, each time a PSU is drawn it is independently sub-sampled in subsequent stages and sampling in subsequent stages is so implemented that independent unbiased estimators $\hat{Y}_i$, say, for PSU totals Y_i are available.

Method II. A practical limitation of Method I is that a person repeatedly sampled may refuse to implement the randomization experiment more than once. Similarly, it may be infeasible and it is expensive to independently sub-sample a chosen PSU more than once in case of multi-stage sampling. A cost-effective and practicable but less efficient alternative may be to employ instead the following estimator e_2 which requires only one RR from each sampled person, namely,

$$e_2 = \frac{1}{n}\sum_{i=1}^{N}\frac{R_i}{p_i} f_{si}, \text{ where } f_{si} = \text{frequency of i in s}.$$

Then, $E_R(e_2) = \frac{1}{n}\sum_{1}^{N}\frac{Y_i}{p_i} fs_i = t_2$ (say),

$$E(e_2) = E_p E_R(e_2) = E_p(t_2) = Y$$

and the variance of e_2 is

$$V(e_2) = E_p E_R(e_2 - Y)^2 = E_p V_R(e_2) + V_p(t_2)$$

$$= E_p\left(\frac{1}{n^2}\sum_{1}^{N}\frac{V_i^2}{p_i^2} f_{si}^2\right) + V_p(t_2) = \frac{1}{n^2}\sum_{1}^{N}\frac{V_i^2}{p_i^2}\left[np_i(1-p_i) + n^2p_i^2\right]$$

$$+ \frac{1}{n}\left(\sum_1^N \frac{Y_i^2}{p_i} - Y^2\right) = \frac{1}{n}\left(\Sigma \frac{Y_i^2}{p_i} - Y^2\right) + \frac{1}{n} \Sigma \frac{V_i^2}{p_i} + \frac{n-1}{n} \Sigma V_i^2$$

$$= V(e_1) + \frac{n-1}{n} \Sigma V_i^2 \qquad (2.4)$$

Then,

$$\nu_2 = \frac{1}{n(n-1)} \sum_1^N \left(\frac{R_i}{p_i} - e_2\right)^2 fs_i + \frac{1}{n} \sum_1^N \frac{E_i^2}{p_i} fs_i \qquad (2.5)$$

may be taken as an unbiased estimator for $V(e_2)$ because

$$E(\nu_2) = E_p E_R(\nu_2) = \frac{1}{n(n-1)} \left[E_p \sum_1^N fs_i \left(\frac{Y_i^2 + V_i^2}{p_i^2}\right)\right.$$

$$\left. - n\, E_p E_R(e_2^2)\right] + \frac{1}{n} E_p \left(\sum_1^N fs_i \frac{V_i^2}{p_i}\right)$$

$$= \frac{1}{(n-1)} \left[\Sigma \frac{Y_i^2}{p_i} - Y^2 + \Sigma \frac{V_i^2}{p_i} - V(e_2)\right] + \Sigma V_i^2$$

$$= \frac{1}{(n-1)} \left[(n-1)V(e_2) - (n-1)\Sigma V_i^2\right] + \Sigma V_i^2 = V(e_2).$$

The estimator e_2 and ν_2 were earlier mentioned briefly by Chaudhuri (1987) and Chaudhuri and Mukerjee (1988).

3. A THEORECTICAL JUSTIFICATION FOR METHOD II.

Adhikary, Chaudhuri and Vijayan (1984) and Chaudhuri (1987) among others investigated optimal strategies for estimating Y using suitably gathered RR relating to sensitive variables. In developing theoretical results they postulated super-population models concerning $\underline{Y} = (Y_1, \ldots, Y_i, \ldots, Y_N)$. permitted non-linear functions of $\underline{R} = (R_1, \ldots, R_i, \ldots, R_N)$ based on R_i, $i \in s$ as estimators for Y and frequently utilized technical analyses employed by Godambe and Joshi (1965) and Godambe and Thompson (1977) and while so doing required that the operators E_p and E_R should commute in the sense that for an estimator $e = e(s,\underline{R})$ for Y based on $\underline{R}$ but free of R_j for $j \notin s$ it is necessary to have (i) $E_p E_R(e) = E_R E_p(e)$ and (ii) $E_p E_R(e - Y)^2 = E_R E_p(e - Y)^2$. Several optimal strategies derived by them relate to classes of strategies which are required to satisfy this "commutativity property". We shall see below that the Method II satisfies "commutativity" while Method I does not. We note that

$E_p(e_1) = \sum_1^N R_i$ and so $E_R E_p(e_1) = E_R\left(\sum_1^N R_i\right) = Y = E(e_1) = E_p E_R(e_1)$

but

$$E_R E_p(e_1 - Y)^2 = E_R E_p[(e_1 - R) + (R - Y)]^2 = E_R V_p(e_1) + V_R(R)$$

$$= E_R\left[\frac{1}{n}\left(\sum_1^N \frac{R_i^2}{p_i} - R^2\right)\right] + \Sigma V_i^2 = \frac{1}{n}\left(\Sigma \frac{Y_i^2}{p_i} - Y^2\right) + \frac{1}{n}\Sigma \frac{V_i^2}{p_i} + \left(1 - \frac{1}{n}\right)\Sigma V_i^2$$

$$= V(e_1) + \frac{n-1}{n}\Sigma V_i^2 \neq V(e_1) = E_p E_R(e_1 - Y)^2 \qquad (3.1)$$

and hence the 'non-commutativity'. On the other hand, $E_p(e_2) = \Sigma R_i$ giving $E_R E_p(e_2) = E_R(\sum_1^N R_i) = Y = E(e_2) = E_p E_R(e_2)$ and also,

$$E_R E_p(e_2 - Y)^2 = E_R E_p[(e_2 - R) + (R - Y)]^2 = E_R V_p(e_2) + V_R(R)$$

$$= E_R\left[\frac{1}{n^2}\Sigma \frac{R_i^2}{p_i^2}\{np_i(1 - p_i)\} - \frac{1}{n^2}\sum\sum_{i\neq j} \frac{R_i}{p_i}\frac{R_j}{p_j} n\, p_i p_j\right] + \Sigma V_i^2$$

$$= \frac{1}{n} E_R\left(\sum_i \frac{R_i^2}{p_i} - \Sigma R_i^2 - \sum\sum_{i\neq j} R_i R_j\right) + \sum_i V_i^2$$

$$= \frac{1}{n}\left(\sum_i \frac{Y_i^2}{p_i} + \sum_i \frac{V_i^2}{p_i} - \sum_i Y_i^2 - \sum_i V_i^2 - \sum\sum_{i\neq j} Y_i Y_j\right) + \sum_i V_i^2$$

$$= \frac{1}{n}\left(\Sigma \frac{Y_i^2}{p_i} - Y^2\right) + \frac{1}{n}\Sigma \frac{V_i^2}{p_i} + \left(1 - \frac{1}{n}\right)\Sigma V_i^2 = V(e_2) = E_p E_R(e_2 - Y)^2$$

and hence the "commutativity". Consequently, the strategy I of employing "PPSWR sampling and e_1" is not admitted as a competitor within the class of strategies wherein optimal ones are sought by Adhikary et al (1984) and Chaudhuri (1987) demanding "commutativity" for every member of the class while the Strategy II of "PPSWR sampling and e_2" is admitted within it. Hence a theoretical use of the inefficient (relative to Method I) Method II which was noted in Section 2 above to have its practical advantage over Method I.

Remark I. If optimal strategies are investigated in multi-stage sampling and if a similar "commutativity" among operators E_1 and E_L of expectations in the first and later stages of sampling is required of a class of strategies within which optimal ones are to be sought then the one due to Raj (1968) mentioned

in Section 2 has also to be similarly discarded and a modification as in Method II should be helpful.

Remark II. Writing V_L for the variance operator for later stages of sampling an important well-known advantage of Raj (1968) strategy is that in variance estimation no estimator for $V_L(\dot{Y}_i) = W_i$, say, is required just as for Method I, ν_1 does not contain any 'unbiased estimator' term for V_i^2. But for Method II, in ν_2 there is a term involving E_i^2 subject to $E_R(E_i^2) = V_i^2$. Similar will be a requirement of a term like $\dot{W}_i$ subject to $E_L(\dot{W}_i) = W_i$ in an analogous modification of Method II and that will be an insurmountable problem in the context of multi-stage sampling if sampling, for example, at any of the later stages is 'systematic' with a single start. But in the context of RR survey this causes no problem so long as the RR procedure is followed in the manner described above in Section 1 admitting E_i^2 with $E_R(E_i^2) = V_i^2$.

Remark III. It may be easily checked (details omitted to save space) that the 'commutativity' property referred to above is satisfied for the strategies derived from the well-known ones due to Horvitz and Thompson (1952) and Rao, Hartley and Cochran (1962) among many others on replacing Y_i in the latter by R_i, where for each sampled individual only one RR is required. It is possible, though we are as yet unable to prove so, that the only reason responsible for 'non-commutativity' is "repeated" procurement of independent RR's from individuals repeatedly sampled.

ACKNOWLEDGEMENT

The authors are grateful to the referee for helpful suggestions that led to an appreciable improvement upon an earlier draft.

BIBLIOGRAPHY

Adhikary, A.K., Chaudhuri, A. and Vijayan, K. (1984). Optimum sampling strategies for randomized response trials. Int. Statist. Rev. 52. 115-125.

Chaudhuri, A. (1987). Randomized response surveys of finite populations: A unified approach with quantitative data. Jour. Statist. Plan. Inf. 15, 157-165.

Chaudhuri, A. and Mukerjee, R. (1988). Randomized response: Theory and techniques. Marcel Dekker, N.Y.

Godambe, V.P. and Thompson, M.E. (1977). Robust near optimal estimation in survey practice. Bull. Int. Statist. Inst. XLVII. Book 3, 129-146.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from finite universes. Jour. Amer. Statist. Assoc. 47, 663-685.

Raj, Des. (1968). Sampling theory. McGraw-Hill Inc., N.Y.

Rao, J.N.K., Hartley, H.O. and Cochran, W.A. (1962). On a simple procedure of unequal probability sampling without replacement. Jour. Roy. Statist. Soc. Ser. B. 24, 482-497.

Received June 1989; Revised December 1989.

Recommended by P. S. R. S. Rao, University of Rochester, Rochester, NY.

Refereed by James Drew, GTE Labs., Waltham, MS.