Pak. J. Statist,
1993 Vol.9(2)A, pp 101-104

# MEAN SQUARE ERROR ESTIMATION IN RANDOMIZED RESPONSE SURVEYS

By

Arijit Chaudhuri
Computer Science Unit
Indian Statistical Institute, Calcutta, India

*(Received: June, 1991   Accepted: Dec, 1992)*

## ABSTRACT

*In estimating, from a sample chosen with varying probabilities, the finite population total of a sensitive variable, a general technique of deriving randomized responses is first presented. A very general mean square error estimation formula is then developed.*

## KEY WORDS

Finite population; Mean square error estimation; Randomized response; Sensitive issues; Varying probability sampling.

## 1. INTRODUCTION

Suppose from a finite population of $N$ units labelled $i = 1, \ldots, N$ a sample $s$ is chosen with an arbitrary probability $p(s)$ corresponding to a design $p$. The problem considered is to estimate the total $Y = \sum_1^N Y_i$ when $Y_i$ are the values of a senstitive variable $y$. Admitting that it is unwise to ask for direct responses (DR) let it be considered judicious to seek randomized responses (RR) from the sampled persons. We shall suppose that from every sampled person labelled $i$ an RR is obtained in independent manners as $r_i$ in such a way that writing $E_r$ as the operator for expectation over randomization in gathering a response we have

$$E_r(r_i) = Y_i$$

$$E_r(r_i - Y_i)^2 = \alpha_i Y_i^2 + \beta_i Y_i + \delta_i = V_i$$

where $\alpha_i(> 0)$, $\beta_i$ and $\delta_i$ are known numbers for every $i = 1, \ldots, N$. For a simple example, a device may be so employed that every sampled person $i$ may be requested to independently choose at random two numbers from each of two sets of numbers

$$\underline{A} = (A_1, \ldots, A_k, \ldots, A_M)'$$

with mean $\mu_A$ and variance $\sigma_A^2$ and

$$\underline{B} = (B_1, \ldots; B_j, \ldots, B_T)'$$

with mean $\mu_B$ and variance $\sigma_B^2$ and record a response, say,

$$Z_i = A_k Y_i + B_j$$

not divulging the right hand side elements to the investigator. Then, $r_i = (Z_i - \mu_B)/\mu_A$ will meet the above stipulations with

$$\alpha_i = \sigma_A^2/\mu_A^2, \beta_i = 0, \delta_i = \sigma_B^2/\mu_A^2 \# \forall\, i = 1, \ldots, N.$$

For alternative devices one may consult Chaudhuri and Mukerjee (1988). Presuming that this trick works, survey data at hand will be

$$d' = (i, r_i | i \varepsilon s) \text{ instead of } d = (i, Y_i | i \varepsilon s)$$

which might have been procured through a DR survey. Based on $d$, if available, an estimator or a predictor for $Y$ is often taken in the form

$$t = t(d) = a_s + \sum_{i \epsilon s} b_{si} Y_i$$

with $a_s$ and $b_{si}$ as constants free of $Y_i'$s. Since $d$ is unavailable we recommend employing

$$e = e(d') = a_s + \sum_{i \epsilon s} b_{si} r_i$$

for the purpose noting that

$$E_r(e) = t \text{ for every } s.$$

In choosing a suitable $t$ often one postulates a plausible model characterizing the $Y_i'$ values which are then treated as random variables. Writing $E_p, E_m$ as operators for expectation with respect to design $p$ and model respectively and $E_G$ to denote either $E_p$ or $E_m$ or $E_pE_m$ or $E_mE_p$, the mean square error (MSE) for $t$ may be denoted as

$$M = E_G(t - Y)^2$$

which one plans to keep under control. When employing $e$ the relevant MSE would be

$$M_r = E_r E_G (e - Y)^2 = E_G E_r (e - Y)^2.$$

As an estimator for $M$ it is the usual practice to employ a quadratic form

$$m = \sum_{i \epsilon s} C_{si} Y_i^2 + \sum_{i \neq j} \sum d_{sij} Y_i Y_j + \Theta_s$$

with $C_{si}, d_{sij}$ and $\Theta_s$ as known constants so chosen that the magnitude of

$$E_G(m) \text{ is kept close to } M.$$

Our purpose here is to propose an estimator $m_r$ for $M_r$ such that $E_r E_G(m_r) = E_G E_r(m_r)$ is close to $M_r$.

## 2. THE PROPOSED MSE ESTIMATOR AND ITS RATIONALE

Our proposed $m_r$ is given by

$$m_r = \sum_{i \epsilon s} (b_{si}^2 - C_{si}) v_i + \sum C_{si} r_i^2 + \sum_{i \neq j} \sum d_{sij} \cancel{Y_i} r_i r_j + \Theta_s$$

where $v_i = \frac{1}{(1+\alpha_i)} (\alpha_i r_i^2 + \beta_i r_i + \delta_i)$, $i = 1, \ldots, N$. Then noting that $E_r(v_i) = V_i$ it follows that

$$\begin{aligned} M_r = E_G E_r (e - Y)^2 &= E_G E_r \left[ \sum_{i \epsilon s} b_{si} (r_i - Y_i) + (t - Y) \right]^2 \\ &= E_G \left[ \sum_{i \epsilon s} b_{si}^2 V_i \right] + E_G (t - Y)^2 = E_G (\sum_{i \epsilon s} b_{si}^2 V_i) + M \end{aligned}$$

and

$$\begin{aligned} E_G E_r (m_r) &= E_G \left[ \sum_{i \epsilon s} (b_{si}^2 - C_{si}) V_i + m + \sum_{i \epsilon s} C_{si} V_i \right] \\ &= E_G (\sum_{i \epsilon s} b_{si}^2 V_i) + E_G(m). \end{aligned}$$

This justifies the use of $m_r$ as an estimator for $M_r$.

## 3. CONCLUDING REMARK

It is needless to illustrate particular choices of $m$ which are too numerous as one may check from the literature especially as is recently being developed through the works of Deng and Wu (1987), Wolter (1985), Rao and Vijayan (1977), Kumar et al. (1985), Royall and Cumberland (1978), Kott (1990 a,b), Särndal, Swensson and Wretman (1989) among others. In each case an RR counter-part is plausibly available as shown in section 2.

## 4. REFERENCES


(1) Chaudhuri, A. and Mukerjee, R. (1988). Randomized Response: Theory and Techniques. Marcel Dekker Inc. N.Y.

(2) Deng Li-Yuan and Wu, C.F.J. (1987). Estimation of varance of the regression estimator. *Jour. Amer. Stat. Assoc.* 82, 568-76.

(3) Kott, P.S. (1990a). Estimating the conditional variance of a design consistent regression estimator. *Jour. Stat. Plan. Inf.* 24, 3, 287-96.

(4) Kott, P.S. (1990b). Recently proposed variance estimators for the simple regression estimator. *Jour. Official Stat.* 6, 4, 451-54.

(5) Kumar, P.,Gupta, V.K. and Agarwal, S.K. (1985). On variance estimation in unequal probability sampling. *Aust. J. Stat.* 27, 195-201.

(6) Rao, J.N.K. and Vijayan, K. (1977). On estimating the variance in sampling with probability proportional to aggregate size. *Jour. Amer. Stat. Assoc.* 72, 579-84.

(7) Royall, R.M. and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Jour. Amer. Stat. Assoc.* 73, 351-58.

(8) Särndal, C.E., Swensson, B. and Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-37.

(9) Wolter, K.M. (1985). Introduction to variance estimation. Springer-Verlag. N.Y.