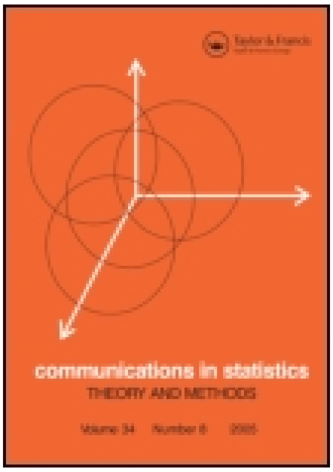


Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

# Note on umvsu-estimation under randomized response model

Parimal Mukhopadhyay [a]

[a] Indian Statistical Institute , 203 B.T.Road, Calcutta, 700035

Published online: 27 Jun 2007.

COMMUN. STATIST.—THEORY METH., 26(10), 2415-2420 (1997)

# A NOTE ON UMVU-ESTIMATION UNDER RANDOMIZED RESPONSE MODEL

Parimal Mukhopadhyay
Indian Statistical Institute
203 B.T.Road
Calcutta 700035



*Key words and Phrases* Ordered design;uniformly minimum variance unbiased estimation;symmetric population function

**ABSTRACT**

Considering a class of randomized response trials for eliciting sensitive information from a sample survey and a class of ordered sampling designs, a uniformly minimum variance unbiased estimator of population variance (of the sensitive character) has been obtained. This note indicates that a theorem (theorem 3.9) of Cassel, Sarndal and Wretman (1977) and the results in the present note can be extended to estimation of any symmetric function of population values in the field of direct response surveys and randomized response surveys respectively.

## 1 Introduction and Preliminaries

Let $\mathcal{P}$ be a finite population of units labelled $1,\ldots,i,\ldots,N$. Associated with $i$ is a real quantity $y_i$, value of a character '$y$' of interest ($i=1,\ldots,N$). We assume that $y$ is a sensitive character (eg. containing some social stigma) about which the respondents may not answer truthfully to a direct question. In such circumstances generally some randomized response trials (RRT) are performed by virtue of which the respondent answers a value $z_i$ when his true value is $y_i$. Such trials are performed independently from person to person. We are interested in estimating the population variance

$$V(\underset{\sim}{y}) = a_1 \sum_{i=1}^{N} y_i^2 - a_2 \sum\sum_{i \neq i' = 1}^{N} y_i y_i' \tag{1.1}$$

where

$$a_1 = \frac{1}{N}, \qquad a_2 = \frac{1}{N(N-1)}$$

by survey sampling, for which a sample $s$ is selected with probability

$$p(s) \left( p(s) \geq 0, \sum_{s \in \mathcal{S}} p(s) = 1, \mathcal{S} = \{s\} \right).$$

$$\pi_k = \sum_{s \ni k} p(s), \pi_{kk'} = \sum_{s \ni (k,k')} p(s)$$

denoting the first-order and second-order inclusion probabilities respectively. We assume that the sampling design used in selecting s is an ordered design such that

(i) each $s$ is a sequence of $n$ distinct labels and has nominal size $n$.

(ii) all $n!$ sequences $s$ that are permutations of the same set of labels have equal probability of being selected.

(iii) $\pi_{ij} \geq 0 \forall i \neq j = 1, \ldots, N$.

Let $\mathbf{S}=(K_1, \ldots, K_n)$ be a random vector taking values $s = (k_1, \ldots, k_n)$ where $K_i$ is a random variable denoting the label of the units selected at the ith draw in $\mathbf{S}$. Data obtained from the survey are $\mathcal{D}=((K_1, z_{K1}), \ldots, (K_n, z_{Kn}))$ available through two types of randomisation,one due to sampling design and the other due to RRT.An estimator $e = (e, \underset{\sim}{z})$ is a function defined $\mathcal{S}XR_N$ such that for a given $s$ its value depends on $\underset{\sim}{z}(= (z_1, \ldots, z_N))$ only through $i \in s$. Let $E_p, E_R$ denote expectation operator with respect to sampling design $p$ and $RR$ respectively. An estimator $t$ is unbiased for $V(\underset{\sim}{y})$ if

$$E_p E_R(t(s, \underset{\sim}{z})) = V(\underset{\sim}{y}) \forall \underset{\sim}{y} \in R_N \tag{1.2}$$

Consider Erikson's (1973) $RR$ model as follows. When the $i$th individual (with true value $y_i$) is selected in the sample he chooses a quantity at radom from the set $(a_{1i}, \ldots, a_{M_i i})$ and a quantity $b_i$ at random from the set $(b_{1i}, \ldots, b_{Q_i i})$ and responds

$$Z_i = a_i y_i + b_i \tag{1.3}$$

It can be seen that

$$E_R(r_i) = y_i,\ E_R(R_i^2) = y_i^2,\ E_R(R_{ij}) = y_i y_j \tag{1.4}$$

where

$$r_i = \frac{z_i - \bar{a}_i}{\bar{b}_i},$$

$$R_i^2 = \frac{z_i^2 - \tilde{b}_i - 2\bar{a}_i \bar{b}_i r_i}{\tilde{a}_i},$$

$$R_{ij} = \frac{z_i z_j - \bar{a}_i \bar{b}_j r_i - \bar{a}_j \bar{b}_i r_j - \bar{b}_i \bar{b}_j}{\bar{a}_i \bar{a}_j},$$

$$\bar{a}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} a_{ji},$$

$$\bar{b}_i = \frac{1}{Q_i} \sum_{j=1}^{Q_i} b_{ji},$$

$$\tilde{a}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} a_{ji}^2,$$

$$\tilde{b}_i = \frac{1}{Q_i} \sum_{j=1}^{Q_i} b_{ji}^2.$$

Let $\underset{\sim}{e} = (e_1, \ldots, e_N) \in R_N$ be a vector of known constants. For a given sample $s = \{i_1, \ldots, i_n\}$, define the quantities:

$$\frac{n(R_{ik}^2 - e_{ik}^2)}{N\pi_{ik}} = T_{ik}^2, \quad (k = 1, \ldots,, n)$$

$$\frac{n(n-1)(R_{i_k i_e} - e_{i_k i_e})}{N(N-1)\pi_{i_k i_e}} = T_{i_k i_e} \ (k \neq e = 1, \ldots, n)$$

and the set of unlabelled data

$$\underset{\sim}{T}_s = (T_{i1}^2, \ldots, T_{i_n}^2, T_{i_1 i_2}, \ldots, T_{i_{n-1} i_n} \mid \text{labels dropped})$$

Clearly, the members of $\underset{\sim}{T}_s$ may not be distinct. We shall consider in this note the class of estimators that depend on $\mathcal{D}$ only through $\underset{\sim}{T}s$. We call this class of estimators as $\Delta$.

## 2 An UMVU-Estimator

**Lemma1.** Let $p$ be any fixed effective sampling design.If $\bar{a}_i(\neq 0), \tilde{a}_i(\neq 0), \bar{b}_i, \tilde{b}_i$ are known constants and $g(\underset{\sim}{T}_s)$ is a real valued statistic such that..

$$E_pE_R\{g(\underset{\sim}{T}_s)\} = 0$$

for every $RR$-model (1.3) and every $\underset{\sim}{y} \in R_N$, then $g(\underset{\sim}{T}_s) = 0$ identically.

**Proof.** The lemma follows from lemma 3.7 of Cassel, Sarndal and Wretman (1977)(whose proof is similar to a proof in Das(1962)) and theorem 2.1 of Adhikari,Chaudhuri and Vijayan(1984)(which can be easily extended to Eriksson's model (1.3)).

**Theorem 1.** Under the $RR$-models satisfying (1.3), for any given ordered design satisfying conditions (i) -(iii) and for any constant vector $\underset{\sim}{e} \in R_N$, the generalised differene estimator of population variance,

$$t'_{GD}(\underset{\sim}{e}) = \frac{1}{N}\sum_{i\epsilon s} T_i^2 - \frac{1}{n(n-1)}\sum\sum_{i\neq j\epsilon s} T_{ij} + V(\underset{\sim}{e})$$
$$= w(\underset{\sim}{T}_s) + V(\underset{\sim}{e}) \tag{2.1}$$

is UMV in the class of all unbiased estimators of $V(\underset{\sim}{y})$ in $\Delta$.

**Proof.** It follows by virtue of the assumptions (i)-(iii) on the class of sampling designs, that the random variables $(T_{i1}^2, \ldots, T_{i_n}^2)$ have exchangeable joint distribution with common expectation $E_pE_R(T_{i_k}^2) = \sum_{i=1}^{N}(y_i^2 - e_i^2)/N$. Similarly for all permutations of $(i_1, \ldots, i_n)$, the random variables $(T_{i_1i_2}, \ldots, T_{i_{n-1}i_n})$ have the same joint distribution with common expectation $E_pE_R(T_{i_ki_e}) = \frac{1}{N(N-1)}\sum\sum_{i\neq j=1}^{N}(y_iy_j - e_ie_j)$. Hence $E_pE_R(T_{i_k}^2 - T_{i_ki_l}) = V(\underset{\sim}{y}) - V(\underset{\sim}{e})$. Again, it follows from lemma 3.6 of Cassel, Sarndal and Wretman(1977)[with a slight modification] that a MVUE of $V(\underset{\sim}{y}) - V(\underset{\sim}{e})$ in the class of all estimators in $\Delta$ must be symmetric in $\underset{\sim}{T}_s$.Now $w(\underset{\sim}{T}_s)$is a symmetric function in $\underset{\sim}{T}_s$ and is unbiased for $V(\underset{\sim}{y}) - V(\underset{\sim}{e})$. Lemma 1 states that under randomisation due to sampling design (satisfying(i)-(iii)) and $RR$-models satisfying (1), distribution of $\underset{\sim}{T}_s$ is complete. Hence $w(\underset{\sim s}{T})$ is the unique unbiased estimator of $V(\underset{\sim}{y}) - V(\underset{\sim}{e})$. Consequently $t'_{GD}(\underset{\sim}{e})$ is the UMVUE of $V(\underset{\sim}{y})$ in the class of all unbiased estimators in $\Delta$.

**Corollary 1.** Under the $RR$-models satisfying (1.3), for any given ordred design satisfying (i)-(iii), the Horvitz-Thompson estimator $\sum_{i\in s} \frac{R_i^2}{N\pi_i} - \sum\sum_{i\neq j\in s} \frac{R_{ij}}{N(N-1)\pi_{ij}}$ is $UMVUE$ for $V(\underset{\sim}{y})$ in the class of all unbiased estimators that depend on $\mathcal{D}$ only through $\underset{\sim}{T}_s^0(=\underset{\sim}{T}_s$ with $\underset{\sim}{e} = \underset{\sim}{e})$.

**Remarks.** Clearly, result on $UMV$ in theorem 3.9 of Cassel and others(1977) can be extended to estimation of any symmetric function of population values, say, population moments, population cumulants,etc. in direct-response surveys. Again results in theorem 1 (which is extension of theorem 1 in Tracy and Mukhopadhyay(1994)) can be extended to any symmetric function of population values in randomized response surveys satisfying (1.3).

**Note 2.1.** We note that the conditions (i)-(iii) on sampling designs is not required for lemma 3.7 of Cassel, Sarndal and Wretman (1977). Thus theorem 2.1 of Adhikari,Chaudhuri and Vijayan (1984) can be modified as follows.

**Lemma 2.2.** Let p be any given FES(n) design. Then for the class of $RR$-models (1.3) $E_pE_R\{g(\underset{\sim}{Z}_s)\} = 0\forall \underset{\sim}{y} \in R_N \Rightarrow g(\underset{\sim}{Z}_s) = 0\forall s : p(s) \geq 0$ and$\forall \underset{\sim}{y} \in R_N$. where $g(\underset{\sim}{z}_s)$ is a function that depends on $\mathcal{D}$ only through the sets of unlabelled responses $\underset{\sim}{Z}_s$.

## BIBLIOGRAPHY

Adhikari,A.K.,Chaudhuri,A.,and Vijayan,K.(1984): Optimum sampling strategies for randomized response trials. Int. Statist. Rev.,**52**, 115-129.

Cassel,C.M., Sarndal,C.E. and Wretman,J.H.(1977): *Foudation of Inference in Survey Sampling*, John Wiley, New York.

Das, A.C.(1962): On MVU estimates of parameters of a finite population based on varying probability sampling. Cal. Statist. Assoc.Bull. **11**,39-48.

Eriksson, S.A.(1973);A new model for randomized response. Int. Statist Rev.,**41**, 101-103.

Tracy,D.S. and Mukhopadhyay,P.(1994): On UMVU-estimation under randomized response models. Statistics,**25**, 173-175.

Received August, 1995; Revised April, 1997.