*International Statistical Review*, (1984), **52**, 2, pp. 115–125. Printed in Great Britain


# Optimum Sampling Strategies for Randomized Response Trials

**Arun Kumar Adhikari[1], Arijit Chaudhuri[2] and K. Vijayan[3]**

[1] *Indian Statistical Institute, Calcutta, 700 035, India; current address, Operations Research Group, Baroda, Gujarat 391 740, India.* [2] *Indian Statistical Institute, Calcutta 700 035, India.* [3] *Department of Mathematics, University of Western Australia, Nedlands, WA 6009, Australia*

## Summary

**Godambe (1980) initiated the study of survey strategies when responses are randomized, within the framework of the unified theory (Godambe, 1955). For this randomized response situation we show that the optimality properties satisfied by the Horvitz–Thompson estimator in the conventional case are also satisfied by a version of the Horvitz–Thompson estimator appropriate for the randomized response case. It is also shown that the nonexistence of any best estimator for the population total continues to be true even when the responses are randomized.**

*Key words:* Horvitz–Thompson estimator; Minimum variance; Randomized response.

## 1 Introduction and notation

In sample surveys on sensitive issues like amount of income understated in income tax returns, respondents are normally reluctant to report the correct values. To overcome this difficulty, Warner (1965) suggested a randomized response approach which was generalized by Eriksson (1973). Unknown to the interviewer the respondent reports the actual values only with probability $c$ $(0<c<1)$, and otherwise would report a random value from a given set of numbers $X_1, X_2, \ldots, X_M$ that would cover the whole range of possible values for the characteristic of the survey. This way the exact value remains confidential and hence better response is expected. In this note we discuss the estimation problem in this randomized set-up.

As usual, we write $Y_i$ for the characteristic value of the $i$th unit in the population. Let us write $Z_i$ for his response, where

$$Z_i = \begin{cases} Y_i & \text{with probability } c, \\ X_j & \text{with probability } (1-c)/M \quad (j = 1, 2, \ldots, M). \end{cases}$$

We will write $\mathbf{Y}$ for the vector of population values and $\mathbf{Z}$ for the vector of responses. It is assumed here that, conceptually, the randomization can be done for every unit in the population beforehand.

We denote the sample design by $p$, the probability of obtaining sample $s$ by $p(s)$, and the randomization giving $\mathcal{Z}$ values by $R$. On the expectation operator $\mathcal{E}$ and variance operator $V$ we use suffixes $p$, $R$, or both, to denote whether the operation is with respect to design, randomization or both, respectively.

## 2 Necessary and sufficient form of unbiased estimates

If we assume that the sample respondents reported the actual $\mathcal{Y}$-values, an unbiased estimate of the population total $Y=\sum_i Y_i$, where the sum is over $i=1,\ldots,N$, could be written as $e(s,\mathbf{Y})$, where $e$ depends on $\mathbf{Y}$ only through the units in the sample $s$. For a given $e(s,\mathbf{Y})$ let us write

$$t(s,\mathbf{Z})=\frac{1}{c}e(s,\mathbf{Z})-\frac{1-c}{cM}\sum_{i=1}^{M}X_i\sum_{i\in s}\frac{1}{\pi_i}. \tag{1}$$

Then $t(s,\mathbf{Z})$ is an unbiased estimator of $Y$ when the responses are randomized. We may call $t(s,\mathbf{Z})$ a derived estimator from $e(s,\mathbf{Y})$. We will show below that any unbiased estimator under randomized response would necessarily be of the form (1). We need the following theorem.

THEOREM 2.1. *If for any real function $h(s,\mathbf{Z})$ the expectation is zero under Rp for all parametric values $\mathbf{Y}$, then $\mathscr{E}_p\{h(s,\mathbf{Z})\}$ is identically zero.*

*Proof.* Let us write $\mathscr{E}_p[h(s,\mathbf{Z})]=\phi(\mathbf{Z})$. We need to show that if $\mathscr{E}_R[\phi(\mathbf{Z})]$ is 0, then $\phi(\mathbf{Z})$ is 0 for every $\mathbf{Z}$. We will prove the result by induction.

First, let $N=1$. Then

$$\mathscr{E}_R\phi(Z_1)=c\phi(Y_1)+\frac{1-c}{M}\sum_{j=1}^{M}\phi(X_j)=0,$$

which implies that $\phi(Y_1)$ should be the same for all $Y_1$. As the range of $Y_1$ includes $X_1, X_2, \ldots, X_M$, we have $\phi(Z_1)\equiv 0$, and hence the claim is true for $N=1$.

Now assume that the claim is true for all $N\leqslant m$, and take the case $N=m+1$. As randomization is independent from unit to unit,

$$\mathscr{E}_R\phi(Z_1,Z_2,\ldots,Z_m,Z_{m+1})=c\mathscr{E}_R\phi(Z_1,Z_2,\ldots,Z_m,Y_{m+1})+\frac{1-c}{M}\sum_{j=1}^{M}\mathscr{E}_R\phi(Z_1,Z_2,\ldots,Z_m,X_j).$$

If we argue as before, $\mathscr{E}_R\phi(Z_1,Z_2,\ldots,Z_m,Y_{m+1})$ should be the same whatever the value of $Y_{m+1}$, and hence should equal zero. But then from the induction hypothesis

$$\phi(Z_1,Z_2,\ldots,Z_m,Y_{m+1})\equiv 0$$

for any value of $Y_{m+1}$, and the claim follows for $N=m+1$ and hence for every $N$.

THEOREM 2.2. *Any unbiased estimator $a(s,\mathbf{Z})$ of $Y$ is essentially of the form $t$ given by* (1).

*Proof.* Note that $\mathscr{E}_R a(s,\mathbf{Z})$ is an unbiased estimator of $Y$, and this could be used as an estimator when there is no randomization. Hence there is an estimator of the form (1). Under $p$ the expectation of this derived estimator and of $a(s,\mathbf{Z})$ are the same by Theorem 2.1. Using the form (1) then,

$$\mathscr{E}_p a(s,\mathbf{Z})=\frac{1}{c}\sum_{i=1}^{N}Z_i-\frac{1-c}{cM}\sum_{i=1}^{M}X_iN,$$

and hence

$$\mathscr{E}_p\left[ca(s,\mathbf{Z})+\frac{1-c}{M}\sum_{j=1}^{M}X_j\sum_{i\in s}\frac{1}{\pi_i}\right]=\sum_{i=1}^{N}Z_i,$$

from which it follows that

$$e(s, \mathbf{Y}) = ca(s, \mathbf{Y}) + \frac{1-c}{M} \sum_{j=1}^{M} X_j \sum_{i \in s} \frac{1}{\pi_i}$$

is an unbiased estimator of $Y$ in the conventional case and $a(s, \mathbf{Z})$ is the corresponding derived estimator.

## 3 Nonexistence of a best unbiased estimator

It is well known that for the conventional sampling set-up there does not exist a best unbiased estimator (Godambe & Joshi, 1965; Basu, 1971). Such a nonexistence result is true for the randomized response situation also. We use Basu's proof with some modification.

Let us define independent random variables $\alpha_1, \alpha_2, \ldots, \alpha_N$ each taking one of two values 0 and 1 with probabilities $1-c$ and $c$ respectively, which give rise to the randomized response

$$Z_i = \begin{cases} Y_i & \text{if } \alpha_i = 1, \\ X_j & \text{with probability } 1/M \text{ if } \alpha_i = 0 \quad (j = 1, 2, \ldots, M), \end{cases}$$

and another set of responses

$$Z_{i0} = \begin{cases} Y_{i0} & \text{if } \alpha_i = 1, \\ X_j & \text{with probability } 1/M \text{ if } \alpha_i = 0 \quad (j = 1, 2, \ldots, M). \end{cases}$$

The vector of the $Y_{i0}$'s (assumed to be known) may be denoted by $\mathbf{Y}_0$, and the vector of the $Z_{i0}$'s by $\mathbf{Z}_0$.

Now let $a(s, \mathbf{Z})$ be a best estimator if one exists. Consider the estimator

$$a^*(s, \mathbf{Z}) = a(s, \mathbf{Z}) - a(s, \mathbf{Z}_0) + \mathscr{E}_R[a(s, \mathbf{Z})],$$

which is unbiased for $Y$. When $\mathbf{Z} = \mathbf{Z}_0$,

$$a^*(s, \mathbf{Z}_0) = \mathscr{E}_R(a(s, \mathbf{Z}_0))$$

and hence

$$V_{Rp}(a^*(s, \mathbf{Z}_0)) < V_{Rp}(a(s, \mathbf{Z}_0))$$

unless $\mathscr{E}_p V_R(a(s, \mathbf{Z}_0)) = 0$, which is not possible. Hence $a(s, \mathbf{Z})$ is not uniformly minimum variance.

## 4 Optimality of Horvitz–Thompson estimator

For the conventional sampling set-up the Horvitz–Thompson estimator, defined as

$$e^*(s, \mathbf{Y}) = \sum_{i \in s} \frac{Y_i}{\pi_i},$$

is known to have many optimum properties (Godambe & Joshi, 1965; Godambe & Thompson, 1973; Cassell, Särndal & Wretman, 1976; C.R. Rao, 1971). All these optimality properties are true also in the random response case for its derived estimator

$$t^*(s, \mathbf{Z}) = \sum_{i \in s} \frac{Z_i^*}{c\pi_i},$$

where

$$Z_i^* = Z_i - \frac{1-c}{M}\sum_{j=1}^{M} X_j.$$

To establish this, we need the following lemma.

LEMMA 4.1. *Let $h(s, \mathbf{Z})$ have expectation zero under Rp and hence under p by Theorem 2.1. Then*

$$\mathscr{E}_p \operatorname{Cov}_R (t^*(s, \mathbf{Z}), h(s, \mathbf{Z})) = 0. \tag{2}$$

*Proof.* Now

$$\mathscr{E}_p\left(\sum_{i\in s}\frac{Z_i^*}{c\pi_i}h(s,\mathbf{Z})\right) = \sum_s\left[\sum_{i\in s}\frac{Z_i^*}{c\pi_i}h(s,\mathbf{Z})\right]p_s = \sum_{i=1}^{N}\sum_{s:i\in s}\frac{Z_i^*}{c\pi_i}h(s,\mathbf{Z})p_s$$
$$= \sum_{i=1}^{N}\frac{Z_i^*}{c\pi_i}\left(\mathscr{E}_p h(s,\mathbf{Z}) - \sum_{s:i\notin s} h(s,\mathbf{Z})p_s\right) = -\sum_{i=1}^{N}\frac{Z_i^*}{c\pi_i}\sum_{s:i\notin s} h(s,\mathbf{Z})p_s$$

from Theorem 2.1. On taking expectation under $R$ of both sides, and remembering that the randomization is done independently on each unit, we get

$$\mathscr{E}_R\mathscr{E}_p\left(\sum_{i\in s}\frac{Z_i^*}{\pi_i}h(s,\mathbf{Z})\right) = -\sum_{i=1}^{N}\frac{\mathscr{E}_R(Z_i^*)}{c\pi_i}\sum_{s:i\notin s}\mathscr{E}_R(h(s,\mathbf{Z}))p_s$$
$$= \sum_{i=1}^{N}\frac{\mathscr{E}_R(Z_i^*)}{c\pi_i}\sum_{s:i\in s}\mathscr{E}_R(h(s,\mathbf{Z}))p_s$$

as $\sum_s \mathscr{E}_R(h(s,\mathbf{Z}))p_s = 0$, and thus

$$\mathscr{E}_R\mathscr{E}_p\left(\sum_{i\in s}\frac{Z_i^*}{\pi_i}h(s,\mathbf{Z})\right) = \sum_s\sum_{i\in s}\frac{\mathscr{E}_R(Z_i^*)}{c\pi_i}\mathscr{E}_R(h(s,\mathbf{Z}))p_s$$
$$= \mathscr{E}_p\left\{\mathscr{E}_R\left(\sum\frac{Z_i^*}{c\pi_i}\right)\mathscr{E}_R h(s,\mathbf{Z})\right\}. \tag{3}$$

Hence

$$\mathscr{E}_p \operatorname{Cov}_R (t^*(s,\mathbf{Z}), h(s,\mathbf{Z})) = \mathscr{E}_p \operatorname{Cov}_R\left(\sum_{i\in s}\frac{Z_i^*}{c\pi_i}, h(s,\mathbf{Z})\right)$$
$$= \mathscr{E}_p\left\{\mathscr{E}_R\left(\left(\sum_{i\in s}\frac{Z_i^*}{c\pi_i}\right)h(s,\mathbf{Z})\right) - \mathscr{E}_R\left(\sum_{i\in s}\frac{Z_i^*}{c\pi_i}\right)\mathscr{E}_R(h(s,\mathbf{Z}))\right\} = 0,$$

from (3). Thus the lemma.

Now we can establish the following theorem from which the optimality results follow.

THEOREM 4.1. *For any unbiased estimator $a(s,\mathbf{Z})$ of $Y$ we have*

$$V_{Rp}(a(s,\mathbf{Z})) = V_p\{\mathscr{E}_R a(s,\mathbf{Z})\} + \mathscr{E}_p\{V_R(t^*(s,\mathbf{Z}))\} + \mathscr{E}_p\{V_R(a(s,\mathbf{Z}) - t^*(s,\mathbf{Z}))\}.$$

*Proof.* We have

$$V_{Rp}(a(s,\mathbf{Z})) = V_p\{\mathscr{E}_R a(s,\mathbf{Z})\} + \mathscr{E}_p\{V_R a(s,\mathbf{Z})\}$$
$$= V_p\{\mathscr{E}_R a(s,\mathbf{Z})\} + \mathscr{E}_p\{V_R t^*(s,\mathbf{Z}) + V_R(a(s,\mathbf{Z}) - t^*(s,\mathbf{Z}))$$
$$+ 2\operatorname{Cov}_R(t^*(s,\mathbf{Z}), a(s,\mathbf{Z}) - t^*(s,\mathbf{Z}))\}.$$

Because

$$\mathscr{E}_{Rp}\{a(s,\mathbf{Z})-t^*(s,\mathbf{Z})\}=0,$$

the result follows from Theorem 2.1.

We may note that, from Theorem 2.1 and (1), we have

$$\begin{aligned}V_{Rp}(a(s,\mathbf{Z}))-V_{Rp}(t^*(s,\mathbf{Z}))&>V_p\{\mathscr{E}_R(a(s,\mathbf{Z}))\}-V_p\{\mathscr{E}_R(t^*(s,\mathbf{Z}))\}\\&=V_p\{\mathscr{E}_R(a(s,\mathbf{Z}))\}-V_p(e^*(s,\mathbf{Y})).\end{aligned}$$

This observation leads to the following optimality theorem.

THEOREM 4.2. *Let $a(s,\mathbf{Z})$ be an unbiased estimator of $Y$. If $\mathscr{E}_R(a(s,\mathbf{Z}))$ has larger expected variance than $e^*(s,\mathbf{Y})$ under some superpopulation model, then $a(s,\mathbf{Z})$ has larger expected variance than $t^*(s,\mathbf{Z})$ under the same model.*

The following corollaries are immediate.

COROLLARY 4.1. *If $e^*(s,\mathbf{Y})$ together with an appropriate design has minimum expected variance amongst all unbiased strategies of $Y$ under some superpopulation model, so does $t^*(s,\mathbf{Z})$ under the same model.*

COROLLARY 4.2. *If $e(s,\mathbf{Y})$ is a linear estimator and has larger variance than $e^*(s,\mathbf{Y})$, then the corresponding derived estimator $t(s,\mathbf{Y})$ has larger variance than $t^*(s,\mathbf{Y})$.*

The optimality results of Godambe & Joshi (1965), Godambe & Thompson (1973, 1977), Cassel *et al.* (1976) and many others would then get extended for the randomized response case through Corollary 4.1. The comparisons made between Horvitz–Thompson and other strategies by Rao (1966), Vijayan (1966) and Chaudhuri & Arnab (1979) also have their analogues in the present case; some details are given by Chaudhuri & Adhikari (1981).

## Acknowledgements

A.K. Adhikari and A. Chaudhuri are thankful to Dr B.K. Sinha for critical comments on their earlier versions of this paper, and to the referees of those versions for helpful suggestions. This final version has been thoroughly rewritten by K. Vijayan, improving substantially on the results and proof techniques in a preliminary version of the paper written jointly by the other two authors. K. Vijayan's work was done while visiting the Department of Statistics and Actuarial Sciences, University of Waterloo, Ontario, Canada. He is thankful to Professors V.P. Godambe and M.E. Thompson for many helpful suggestions that improved the format of this note.

## References

Basu, D. (1971). An essay on the logical foundations of survey sampling I. In *Foundations of Statistical Inference*, Ed. V.P. Godambe and D.A. Sprott, pp. 203–242. Toronto: Holt, Rinehart and Winston.

Cassell, C.M., Särndal, C.E. & Wretman, J.H. (1976). Some results on generalised difference estimation and generalised regression estimation for finite populations. *Biometrika* **63,** 615–620.

Chaudhuri, A. & Arnab, R. (1979). On the relative efficiencies of sampling strategies under a superpopulation model. *Sankhyā* C **41,** 40–43.

Chaudhuri, A. & Adhikari, A.K. (1981). Sampling strategies with randomised response trials—their properties and relative efficiencies. Technical report, Indian Statistical Institute, Calcutta.

Eriksson, S.A. (1973). A new model for randomised response. *Rev. Int. Statist. Inst.* **41,** 101–113.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *J.R. Statist. Soc.* B **17,** 269–278.

Godambe, V.P. (1980). Estimation in randomised response trials. *Int. Statist. Rev.* **48,** 29–32.

Godambe, V.P. & Joshi, V.M. (1965). Admissibility and Bayes' estimation in sampling finite populations I. *Ann. Math. Statist.* **36,** 1707–1722.

Godambe, V.P. & Thompson, M.E. (1973). Estimation in sampling theory with exchangeable prior distributions. *Ann. Statist.* **1,** 1212–1221.
Godambe, V.P. & Thompson, M.E. (1977). Robust near optimal estimation in survey practice. *Bull. Int. Stat. Inst.* **47,** 129–146.
Rao, C.R. (1971). Some aspects of statistical inference in problems of sampling from finite populations. In *Foundations of Statistical Inference*, Ed. V.P. Godambe and D.A. Sprott, pp. 177–202. Toronto: Holt, Rinehart and Winston.
Rao, J.N.K. (1966). On the relative efficiency of some estimators in pps sampling for multiple characteristics. *Sankhyā* A. **28,** 61–70.
Vijayan, K. (1966). On Horvitz–Thompson and Des Raj estimators. *Sankhyā* A **28,** 87–92.
Vijayan, K. (1982). Optimum sampling strategies for randomised response trials, Technical report STAT-82-05, Department of Statistics and Actuarial Science, University of Waterloo.
Warner, S.L. (1965). Randomised response—A survey technique for eliminating evasive answer bias. *J. Am. Statist. Assoc.* **60,** 63–69.

**Résumé**

Godambe (1980) a initié l'étude des stratégies des enquêtes dans les cas de réponses randomisées, dans le cadre de la théorie unifiée par Godambe (1955). Pour cette situation des réponses randomisées on montre que les propriétés optimaux satisfaites par l'estimateur Horvitz–Thompson dans le cas classique sont aussi satisfaites par une version de cet estimateur approprié au cas des réponses randomisées. On note aussi que la manque d'un estimateur optimal pour la population totale est vraie même dans le cas des réponses randomisées.

[*Paper received March,* 1982, *revised November* 1983]

## Discussion of paper by A.K. Adhikari, A. Chaudhuri and K. Vijayan

**P.K. Sen**

*Department of Biostatistics 201H, University of North Carolina, Chapel Hill, NC 27514, USA*

It is a pleasure for me to contribute to the discussion of this interesting paper. Randomized response trials play a vital role in survey sampling, and the present paper has indeed made a valuable theoretical contribution in this area.

I have, however, a few observations and comments to make. First, the assumed randomized response model (for the $Z_i$) is somewhat less general than the usually adopted one. In a general context, one assumes typically that with the set $\mathscr{X} = \{x_1, \ldots, x_m\}$ of realizations, there is an associated set $Q = \{q_1, \ldots, q_m\}$, where the $q_j$ are nonnegative real numbers and $\sum_j q_j = 1$ with the sum over $j = 1, \ldots, m$, such that, for some $c$ $(0 < c < 1)$,

$$Z_i = \begin{cases} Y_i & \text{with probability } c, \\ x_j & \text{with probability } (1-c)q_j \quad (j = 1, \ldots, m), \end{cases} \tag{D1}$$

for $i = 1, \ldots, N$; without any loss of generality, we take $x_1, \ldots, x_m$ to be all distinct. If the $q_j$ are all rational numbers, then, of course, $M$ and $X_1, \ldots, X_M$ may be so chosen that $Mq_j$ of the $X_i$ are equal to $x_j$ $(j = 1, \ldots, m)$, and hence the uniform probability model considered by these authors can be worked out by allowing $X_1, \ldots, X_M$ to be not necessarily all distinct. On the other hand, this reduction to a discrete uniform distribution may not be generally possible. As a simple example, consider the case $\mathscr{X} = \{0, 1\}$ and $q_0 = 1 - q_1 = \pi^{-1}$, so that the probabilities are not rational numbers. In this case, the finite and discrete uniform distribution will not work out. The results in the current paper can easily be modified to suit this more general model. In fact, in (D1), $m$ may also be equal