Sampling 'Survey Populations'– Some Problems and Their Solutions

A Thesis Submitted to the Indian Statistical Institute

by

Kajal Dihidar

In Partial Fulfilment of the Requirements for the Degree
of
Doctor of Philosophy in Statistics

February, 2010

Indian Statistical Institute
Kolkata, India

Acknowledgement and Declaration

This thesis is being submitted to the Indian Statistical Institute in fulfillment of the primary requirements for the award of the degree of Doctor of Philosophy in Statistics.

I declare that no part of this thesis has ever been submitted elsewhere for any degree or diploma or certificate. However, this thesis includes materials from the following papers. These are

1. Chaudhuri, A., Bose, M. and Dihidar, K. (2005). Sample-size-restrictive adaptive sampling: an application in estimating localized elements. *Journal of Statistical Planning and Inference.* **134**, 254-267.

2. Chaudhuri, A., Dihidar, K. and Bose, M. (2006). On the feasibility of basing Horvitz & Thompson's estimator on a sample by Rao, Hartley & Cochran's scheme. *Communications in Statistics. Theory and Methods.* **35**, 2039-2044.

3. Chaudhuri, A., Bose, M. and Dihidar, K. (2009c). Rao-Hartley-Cochran sampling with competitive estimators. *6th Triennial Symposium Proceedings Volume, Calcutta Statistical Association Bulletin.* **61**, 227-242.

4. Chaudhuri, A., Bose, M. and Dihidar, K. (2009a). Estimating sensitive proportions by Warner's randomized response technique using multiple randomized responses from distinct persons sampled. Accepted for publication in *Statistical Papers.* Now available online through journal's website: DOI: 10.1007/s00362-009-0210-3.

5. Chaudhuri, A., Bose, M. and Dihidar, K. (2009b). Estimation of a sensitive proportion by Warner's randomized response data through inverse sampling. Accepted for publication in *Statistical Papers.* Now available online through journal's website: DOI: 10.1007/s00362-009-0234-8.

6. Bose, M., Chaudhuri, A., Dihidar, K. and Das, S. (2009). Model-cum-design based estimation of the prevalence rate of a disease in a locality using spatial smoothing. Accepted for publication in *Statistics.*

7. Dihidar, K. (2009). Modifying classical randomized response techniques with provision for true response. *Indian Statistical Institute Technical Report.* No. ISI/ASD/2008/8. (Submitted for publication).

I thank Professor Arijit Chaudhuri, Professor Mausumi Bose and Dr. Shyamal Das for permitting me to include my joint work with them in this thesis.

The work in this thesis was done under the supervision of Professor Mausumi Bose. She has not only served as my supervisor but has also played a major role in my academic development. Without her help, this thesis could not have been written. I am deeply indebted to her.

I am deeply grateful to Professor Arijit Chaudhuri for his excellent and inspiring help without which this thesis could not have taken this shape.

I thank Professor Samindra Sengupta, Department of Statistics, University of Calcutta, for some fruitful suggestions regarding the Chapters 5 and 6.

I thank Dr. Shyamal Das, Professor, Bangur Institute of Neurology, Kolkata for permitting me to use their data, obtained from a large scale sample survey for a study on the prevalence of major neurological disorders including stroke, in Chapter 4 of this thesis.

I also owe debts of gratitude to the Director, the Dean of Studies, the Head of Applied Statistics Unit and the Professor-in-Charge, Applied Statistics Division, Indian Statistical Institute, Kolkata for their kind help in enabling me to perform this work.

Finally, I acknowledge my sincere appreciation for the help and good wishes of all faculty members of my unit, my colleagues, friends, relatives and the members of my family. They have been an unending source of inspiration to me during the entire period of this research work.

Date:

(Kajal Dihidar)
Associate Scientist A
Applied Statistics Unit
Indian Statistical Institute
203, B.T. Road. Kolkata : 700108

Contents

Chapter 1

Introduction

1.1 Notation and Preliminaries

In this thesis we study some problems in survey sampling and propose their solutions. The thesis is divided into seven chapters, the first being an introductory one. Chapters 2–4 relate to surveys with direct responses, while Chapters 5–7 deal with randomized responses. We begin by introducing the motivations behind the principal problems studied in this thesis.

The first problem arose while developing a survey where the objective was to estimate the total number of workers in different industries in the rural unorganized sector in a district of the state of West Bengal, India. For this, the technique of adaptive sampling was found useful but this escalated the survey costs substantially as the final adaptive sample size was prohibitively large. So we studied this aspect and proposed an easily implementable modification to the traditional adaptive sampling method which effectively controls the final sample size. Corresponding estimators and variance estimators were derived too.

In the above study we adopted the Rao-Hartley-Cochran (1962)(RHC) scheme of sampling and used the traditional estimator (RHCE) and its variance estimator given by Rao, Hartley and Cochran (1962). We were then curious to see whether the celebrated estimator due to Horvitz and Thompson (HTE) could also be employed here when the sample was an RHC sample. On studying this problem and deducing the required inclusion probabilities, we found that this can indeed be done; the HTE based on an RHC sample turning out to be quite competitive compared to the

traditionally used RHCE.

The second problem came from a survey which was carried out to estimate the prevalence rate of a disease in a certain geographical area, namely the Kolkata Municipal Area in West Bengal. For this estimation, we proposed a modified estimator similar to the Hartley-Ross estimator and showed that suitable modeling, combined with our proposed design-based estimator, can lead to improved estimation.

Next, we focused on the randomized response (RR) techniques which are of immense use in practice when the variable under study is a stigmatizing or sensitive one. Here it is well known that in Warner's model, based on a simple random sample with replacement (SRSWR), a respondent is asked to generate an RR every time he/she is selected; the traditional estimator being the sample mean of these RR's. In the context of direct surveys with SRSWR, classical results due to Basu (1958), Pathak (1962) and others showed that estimators using the direct responses from the distinct sampled units perform better than the sample mean. We were curious to examine if a parallel result is also true in the RR scenario. We carried out this study for sampling with a fixed sample-size and also for inverse sampling. This in turn led to some more related questions in the area of RR based estimation.

We now introduce a number of terms and notation which we will use throughout. In the next section, we briefly discuss some sampling schemes and estimators, together with a few relevant references. More details and references are cited in the appropriate chapters that follow. In Section 1.3, we present a chapter-wise summary.

A finite collection of a known number of identifiable *units* will be called a *survey population*. This population will be denoted by U and we suppose that it consists of N units, labeled as $1, \ldots, i, \ldots, N$. Let y be a variable of interest, the unknown but fixed values of y for the units in U being $y_1, y_2, \ldots, y_N$. We assume that for a sampled unit i, y_i can be ascertained without error if y is not a sensitive variable. This is the usual case of *direct surveys* where a *direct response* can be obtained for y_i. Otherwise, that is, if y is a sensitive variable, then y_i cannot be ascertained by direct response and so one has to have recourse to *randomized responses* obtained on using a suitable *randomization device*.

The total of y over all the units in U is called the *finite population total*

and we will denote it by Y. So,

$$Y = \sum_{i=1}^{N} y_i \qquad (1.1.1).$$

Our objective is to estimate the population total Y (or the mean $\bar{Y} = Y/N$) from data obtained from a suitably chosen sample from U; a sample s is supposed to be chosen with probability $p(s)$. We assume that the sample size is a fixed pre-determined number n, unless we are studying a sample obtained by inverse sampling. We will write

$$\sum_{i \in s}$$

to denote the sum over all units i which are included in the sample s, when it is a set of distinct units. More generally, s may be a sequence of units of U, not necessarily distinct, but permitted to occur with varying frequencies.

In some cases, we will suitably use available *auxiliary variables*(or *covariates*) in sample selection and/or estimation of parameters, in order to get efficient estimators. These auxiliary non-stochastic variables will be denoted by $x_1, x_2, \ldots, x_k$, say, and it is assumed that for any auxiliary variable, its values are known for each i in U.

Let θ be a parameter of interest, for instance, θ may be Y. We will usually write $\hat{\theta}$ to denote an estimator for θ, for instance, $\hat{Y}$. Sometimes, we may use some other more convenient notation which we will define appropriately when needed. We will primarily study unbiased estimators, i.e., estimators which satisfy

$$E(\hat{\theta}) = \theta,$$

and then study its variance, namely, $V(\hat{\theta})$, where we write E and V for the expectation and variance operators with respect to a sampling design p. To specify such design based operators, we shall often employ a subscript P to E and V. We will also obtain an unbiased estimator for $V(\hat{\theta})$ which will usually be denoted by $\hat{V}(\hat{\theta})$ or $v(\hat{\theta})$; using other more convenient notation wherever necessary. The Mean Squared Error (MSE) of an estimator will also sometimes be studied to cover biased estimators $\hat{\theta}$ for θ.

In the sampling literature, two or more unbiased estimators are often compared on the basis of their variances. Again, a common measure for an estimator $\hat{\theta}$ is its coefficient of variation (cv), given by

$$cv = 100 \frac{\sqrt{v(\hat{\theta})}}{|\hat{\theta}|}. \qquad (1.1.2)$$

Estimators may sometimes be also compared using measures based on confidence intervals for the parameters they are meant to estimate. For this, the sampling is replicated a large number of times and the competitor estimators are computed for each such sample. The standardized pivotal, namely, $\tau = \frac{\hat{\theta}-\theta}{\sqrt{v(\hat{\theta})}}$ is assumed to be a standard normal deviate. Then,

$$\left(\hat{\theta} - 1.96\sqrt{v(\hat{\theta})}, \quad \hat{\theta} + 1.96\sqrt{v(\hat{\theta})}\right) \tag{1.1.3}$$

is used as a 95% confidence interval for θ based on the estimator $\hat{\theta}$.

Two measures based on this confidence interval are often used to compare the performance of the alternative estimators. One is the ACP, i.e., the Average Coverage Percentage, which is the percent of the replicated samples for which θ is covered by the interval given by (1.1.3). The second measure is the ARL, i.e., the average relative length, which is the relative length of the confidence interval ($=2 \times 1.96\sqrt{v(\hat{\theta})}/|\hat{\theta}|$) averaged over all the replicates. Another measure is the ACV, i.e., the Average Coefficient of Variation, which is the average over these replicates of the cv as in (1.1.2), is also used as a criterion for comparison. A good estimator is one with a high value of ACP, the closer this value is to 95%, the better the estimator. Again, with respect to ARL, a good estimator should have a small value of ARL. Similarly, a small value of ACV is also desirable for a good estimator.

In this thesis, depending on the problem under study, we have used either the variance and the estimates thereof, ACP, ARL or ACV to compare the efficiencies of competing estimators.

1.2 Review of relevant literature

We begin by listing some books on sample surveys which we have referred to. These are Cochran (1963, 1977), Murthy (1967), Raj (1968), Thompson (1997), Chaudhuri and Stenger (2005), Skinner, Holt and Smith (1989), Chambers, Chambers and Skinner (2003). In our review of survey sampling literature related to the material covered in this thesis, we begin with the estimator due to Horvitz and Thompson (1952). We have used this in our Chapters 2–6. Let $\sum_{s \ni i}$ and $\sum_{s \ni i,j}$ respectively denote the sum over all samples s which include unit i and which include both the units i and j. For an

arbitrary sampling design p, the inclusion probability of unit i is

$$\pi_i = \sum_{s \ni i} p(s), \quad 1 \le i \le N, \tag{1.2.1}$$

and the inclusion probability for any pair of units i, j is

$$\pi_{ij} = \sum_{s \ni i,j} p(s), \quad 1 \le i, j \le N, \ i \ne j. \tag{1.2.2}$$

Let

$$I_{si} = \begin{cases} 1 & \text{if} \quad i \in s \\ 0 & \text{if} \quad i \notin s \end{cases} \quad \text{and} \quad I_{sij} = I_{si} I_{sj}. \tag{1.2.3}$$

For a given sample s, and with $\pi_i > 0$ for all i, the Horvitz and Thompson (1952) estimator (HTE) and its variance are as given below. An unbiased estimator of the variance, as proposed by Horvitz and Thompson (1952) assuming $\pi_{ij} > 0$ for all $i \ne j$, is also shown. For ease of reference, we denote the HTE of Y by $\hat{Y}_{HT}$ and use a corresponding suffix HT for the expressions for variance and variance estimator too.

$$\begin{aligned} \hat{Y}_{HT} &= \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} I_{si} \frac{y_i}{\pi_i} \\ V_{HT}(\hat{Y}_{HT}) &= \sum_{i=1}^{N} y_i^2 \frac{1-\pi_i}{\pi_i} + \sum_{i(\ne j)=1}^{N} \sum_{j=1}^{N} y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}, \\ \hat{V}_{HT}(\hat{Y}_{HT}) &= \sum_{i=1}^{N} y_i^2 \frac{1-\pi_i}{\pi_i} \frac{I_{si}}{\pi_i} + \sum_{i(\ne j)=1}^{N} \sum_{j=1}^{N} y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I_{sij}}{\pi_{ij}}, \end{aligned} \tag{1.2.4}$$

Yates and Grundy (YG, 1953) gave an alternative form of the variance of Y_{HT}, together with an unbiased estimator for this variance, assuming $\pi_{ij} > 0$ for all $i \ne j$ and every s (with $p(s) > 0$) and in addition, having a fixed number of distinct units. For this reason, Brewer and Hanif (1983) call this variance estimator to be conditionally unbiased. These formulae due to Yates and Grundy (1953) are displayed below. As usual, we use the suffix YG to remind us that these are the Yates Grundy form of estimators.

$$\begin{aligned} V_{YG}(\hat{Y}_{HT}) &= \sum_{i(<j)=1}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2, \\ \hat{V}_{YG}(\hat{Y}_{HT}) &= \sum_{i(<j)=1}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \frac{I_{sij}}{\pi_{ij}}. \end{aligned} \tag{1.2.5}$$

Godambe (1960) showed that the HTE is admissible in the class of homogeneous linear unbiased estimators and subsequently, this result was

extended by Godambe and Joshi (1965) to the wider class of all unbiased estimators. This property makes the HTE attractive. Extensive research on many aspects of admissibility and uniform admissibility in finite population sampling are available in Ghosh (1987). However, neither $\hat{V}_{HT}(\hat{Y}_{HT})$ nor $\hat{V}_{YG}(\hat{Y}_{HT})$ is uniformly non-negative, but for designs satisfying

$$\pi_i \pi_j \geq \pi_{ij} \text{ for all } i \neq j, \tag{1.2.6}$$

$\hat{V}_{YG}(\hat{Y}_{HT})$ is uniformly non-negative. This makes the YG-variance estimator more popular among survey sampling practitioners than the HT variance estimator. Many researchers have studied this problem of negative variance estimation and some have suggested some remedies, for instance, Raj (1956), Jessen(1969), Rao (1979), Biyani (1980), Brewer (1990) and others. In this thesis, wherever we use the HTE, we will assume that $\pi_{ij} > 0$ for all $i \neq j$ and hence, also $\pi_i > 0$ for all i. Many other results on HTE are available , for instance in Brewer (2001), Hartley and Rao (1962), Asok and Sukhatme (1976) and others.

Now, we discuss some sampling schemes we have used in this thesis. Among the available sampling schemes, the simple random sampling scheme (SRS) is well known and well studied in the sampling literature; see for example, Cochran (1963, 1977), Chaudhuri and Stenger (2005) and others. This scheme, used either with replacement (SRSWR) or without replacement (SRSWOR), is popularly used in many survey sampling situations. For an SRSWR, we have

$$p(s) = \frac{1}{N^n} \tag{1.2.7}$$

for every sample s of size n. On the other hand, for an SRSWOR, we have

$$p(s) = \frac{1}{\binom{N}{n}}, \tag{1.2.8}$$

for every sample s of size n. It is known that in both cases, an unbiased estimator for Y is given by

$$\hat{Y} = N(\sum_{i \in s} y_i/n) = N\bar{y}, \tag{1.2.9}$$

where $\bar{y} = \sum_{i \in s} y_i/n$ is the sample mean. Then, we have

$$\begin{aligned} V(\hat{Y}) &= \frac{N}{n}\sum_{i=1}^{N}(y_i - \bar{Y})^2 \qquad \text{for SRSWR} \\ &= \frac{N(N-n)}{n(N-1)}\sum_{i=1}^{N}(y_i - \bar{Y})^2 \quad \text{for SRSWOR,} \end{aligned} \tag{1.2.10}$$

and the corresponding unbiased variance estimators are given by

$$\begin{aligned} \hat{V}(\hat{Y}) &= \frac{N^2}{(n-1)n}\sum_{i\in s}(y_i-\bar{y})^2 \qquad \text{for SRSWR} \\ &= \frac{N(N-n)}{n(n-1)}\sum_{i\in s}(y_i-\bar{y})^2 \quad \text{for SRSWOR.} \end{aligned} \tag{1.2.11}$$

Besides the schemes with constant $p(s)$ for all s, several varying probability sampling schemes are also popular in the literature. Among these, the scheme proposed by Rao, Hartley and Cochran (RHC, 1962) is a simple and practically useful one. The RHC scheme may be used for estimating Y when some size measures for the population units are available. For this scheme, Rao et al. (1962) gave an unbiased estimator (RHCE) for Y, together with a uniformly non-negative unbiased estimator for its variance.

To choose a sample of n units from U by the RHC method, the population units are first randomly grouped into n non-overlapping groups. SRSWOR is used to form these n groups out of the N units in U, with N_i units in the ith group, $1 \le i \le n$. Here the N_i's are positive integers satisfying $\sum_n N_i = N$, where $\sum_n$ denotes sum over the n groups.

Let x be an auxiliary variable with $\sum_{i=1}^{N} x_i = X$. We define the normed size-measure for unit i as

$$p_i = x_i/X; \quad \text{where } 0 < p_i < 1, \ \sum_1^N p_i = 1.$$

Let Q_i denote the sum of the normed sizes of the N_i units falling in the ith group. Then, a unit i_k, say, of the ith group is chosen with probability p_{i_k}/Q_i. Thus, only one unit is selected from each of the n groups, giving the sample of size n. This is independently repeated across the n groups, resulting in the desired sample of size n chosen by RHC method.

For simplicity in notation, we denote the value obtained from the chosen unit from ith group by y_i and its normed size measure as p_i. With this notation, the unbiased estimator of Y as given by RHC (1962), i.e., the RHCE of Y is shown below, together with its variance. For ease of reference, we use the suffix RHC with these expressions.

$$\hat{Y}_{RHC} = \Sigma_n y_i \frac{Q_i}{p_i}, \ \ V_{RHC}(\hat{Y}_{RHC}) = \frac{\sum_n {N_i}^2 - N}{N(N-1)} \sum_{i<}^{N}\sum_{j=1}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2. \tag{1.2.12}$$

Rao et al. (1962) also gave the optimal choice of the group sizes N_i for which $V(\hat{Y}_{RHC})$ attains its minimum. This optimal choice is as follows:

$$\begin{array}{rcl} N_i = [N/n] & = & m \text{ say}; \quad 1 \leq i \leq k; \\ = [N/n] + 1 & = & m+1; \quad k+1 \leq i \leq n, \end{array} \tag{1.2.13}$$

where $N = mk + (m+1)(n-k)$, with $1 < k \leq n$. In Chapters 2, 3 and 7, we adopt the RHC scheme with N_i values given by (1.2.13).

Rao et al. (1962) gave an unbiased estimator for V(RHCE). This estimator is as follows:

$$\hat{V}_{RHC}(\hat{Y}_{RHC}) = \frac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2} \Sigma_n \Sigma_n Q_i Q_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \tag{1.2.14}$$

where $\Sigma_n \Sigma_n$ denotes sum over non-overlapping pairs of the n groups. Note that the variance estimator in (1.2.14) is always non-negative. Later, Ohlsson (1989) gave the following alternative unbiased variance estimator.

$$\hat{V}_O(\hat{Y}_{RHC}) = \frac{\sum_n N_i^2 - N}{n(n-1)} \Sigma_n \Sigma_n \frac{Q_i}{N_i} \frac{Q_j}{N_j} \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 .$$

Using numerical illustrations based on simulated observations, he claimed that his variance estimator is possibly better than that given by Rao et al. (1962). In his illustrations he allowed the group sizes to deviate appreciably from Rao et al. (1962)'s optimal choice of group sizes as shown in (1.2.13). However, the variance estimator given by Rao et al. (1962) remains quite competitive with Ohlsson's (1989) variance estimator when the group sizes are chosen as the optimal sizes and the two estimators match when N is a multiple of n.

In this thesis, we use the optimal choice of N_i's as in (1.2.13) and so we do not use Ohlsson's (1989) estimator, but rather use the RHC variance estimator shown in (1.2.14).

Some authors have attempted to generalize the original RHC procedure; for instance, Chikkagoudar (1967) generalized the original RHC procedure in the sense that instead of choosing one element from each group, a sample of several units without replacement is drawn from every group; while Samiuddin and Asad (1981) proposed a method of group formation which is different from that of the RHC method, but where again a single unit is drawn from each group. Rao, Sinha and Srivenkataramana (2003) showed

that for the RHC scheme

$$p_i > p_j \Leftrightarrow \pi_i > \pi_j \quad \forall \quad i \neq j$$

and

$$p_i > p_j \Leftrightarrow \pi_{ik} > \pi_{jk} \quad \forall \quad k \neq i \neq j.$$

The RHC scheme of sampling has also been compared with other varying probability proportional to size (PPS) sampling schemes by several researchers. For example, Mukhopadhyay (1996) showed that the variance of RHCE is equal to a constant multiple of the variance of the PPSWR estimator, and when N is a multiple of n, the RHC procedure is better than the PPSWR procedure. Mukhopadhyay (1996) also compared RHC procedure with Deshpande's (1984) modified PPSWR procedure. Acknowledging that RHCE is inadmissible, Berg (1974) studied its Rao-Blackwellized version as an improvement but avoided over-elaboration owing to its complicated form. Hartley and Rao (1962) suggested the circular systematic sampling with varying probabilities after arranging the population units at a random permutation, and then obtained the asymptotic expression for the variance of their estimator. Considering this asymptotic expression for variance of Hartley and Rao's (1962) estimator, Mukhopadhyay (1996) showed that Hartley and Rao's (1962) scheme is superior to RHC scheme under certain assumptions.

Notwithstanding the above, the RHC scheme continues to be employed because of its simplicity and since the estimator of V(RHCE) is uniformly non-negative and also easily computable. In Chapters 2, 3 and 7 we have used the RHC scheme in our study.

Now, we focus on results in the area of adaptive sampling. A challenging survey sampling situation is one where the units bearing the characteristic of interest in the survey population are rare and localized. Thus, the population includes many units for which the variable of interest y has a zero value while units with non-zero values are few in number and clustered in some unknown pockets of the population; the non-zero values possibly being quite large in some cases. As a result, a sample chosen from this population by traditional sampling methods can often fail to capture enough of these rare and localized units and so, based on such a sample, it may be difficult to estimate the population total Y with an adequate level of efficiency. To overcome this problem, Thompson (1990) introduced the technique of

adaptive sampling for sampling from such populations. This technique was further extended and developed by Thompson (1992) and Thompson and Seber (1996). This method of sampling allows for adaptive augmentation of an initial sample to include neighbourhoods of those units in the sample where the material of interest is found. This technique has been successfully used in the exploration of mineral deposits in unknown regional pockets, in studies for estimating the number of rare plants and animals in a large geographical area, and in other similar surveys. For details of such examples and for more details on this topic, we refer to Thompson and Seber (1996).

The above mentioned references study adaptive sampling based on an initial sample drawn by simple random sampling. Later, Chaudhuri (2000) showed that if an initial sample is drawn by *any sampling method* which leads to an unbiased estimator for Y as well as one for its variance, then this initial sample may be extended to an adaptive sample and the corresponding unbiased estimators may be easily obtained. We give some details of this method below.

Let $N(i)$ denote a uniquely defined neighborhood of units corresponding to unit i. If y_i fails to satisfy a condition, say, C^*, then i itself is a 'Singleton Network' for i. If y_i satisfies C^*, then let $C(i)$ be a cluster of i containing i and all units in its neighborhood, and y_i is observed for the units in the neighbourhood. Again, C^* is checked for the units in $C(i)$ and if it is satisfied for any unit, units in its neighborhood are added to $C(i)$. This process of scanning and adding units continues, stopping only on reaching units with C^* unsatisfied. The units in the clusters with C^* unsatisfied are the *edge-units* of the cluster which of course are each a singleton network. The cluster $C(i)$ with all its edge-units dropped is called the network of i and is denoted by $A(i)$. Clearly, all the networks are mutually non-overlapping and they together exhaust the entire population.

By following this procedure, instead of covering only the original sample s, we have effectively to cover the units in $A(s)$, the union of $A(i)$ over i in s, which is an extension of s. This $A(s)$ is an *Adaptive sample* corresponding to s and this process of extending the sample from s to $A(s)$ is called *Adaptive sampling*.

Chaudhuri (2000) showed that if we denote the cardinality of $A(i)$ by m_i, and define

$$t_i = \frac{1}{m_i} \sum_{j \epsilon A(i)} y_j \tag{1.2.15}$$

then it follows that

$$T = \sum_{i=1}^{N} t_i = Y = \sum_{i=1}^{N} y_i.$$

So, estimating Y using the survey data $(s, y_i | i \epsilon s)$ is equivalent to estimating T using the observations $(s, t_i | i \epsilon s)$. Writing $\underline{Y} = (y_1, \cdots, y_i, \cdots, y_N)$ and $\underline{T} = (t_1, \cdots, t_i, \cdots, t_N)$, if one employs an estimator $\hat{Y} = t(s, \underline{Y})$ which is unbiased for Y then

$$\hat{T} = t(s, \underline{T}), \tag{1.2.16}$$

is also unbiased for T and hence for Y as well. This striking observation of Chaudhuri (2000) simplifies matters considerably and allows one to easily construct estimators for adaptive sampling, even in complex surveys, where the initial sample is chosen according to some varying probability sampling design.

Thompson and Seber (1996) have observed that when the original sample is a simple random sample without replacement (SRSWOR), increased efficiency is ensured for adaptive sampling. But for general sampling schemes, no general claim is warranted about gain in efficiency through adaptive sampling. Recently, Chaudhuri, Bose and Ghosh (2004) applied adaptive sampling to effectively estimate the numbers of rural earners, principally through specific small-scale single industries in the unorganized sector abounding in unknown pockets of a district in India. They showed by numerical methods that appreciable gain in efficiency could be achieved by an adaptive sample for estimating totals of rare and localized units. We study some aspects of adaptive sampling in Chapter 2.

Now we consider the estimation problem of the total Y based on certain ratios obtained from a sample selected from the population by SRSWOR. Suppose x is an auxiliary variable and for every sampled unit i, the values y_i and x_i are known, together with the population total $X = \sum_{i=1}^{N} x_i$. For such situations, the classical ratio estimator for Y is given by

$$\hat{Y}_R = X \frac{\bar{y}}{\bar{x}},$$

where $\bar{y}$ and $\bar{x}$ are the sample means of y and x, respectively. This estimator is biased and though for large n the bias is negligible as compared to the standard deviation, several 'ratio-type' estimators have been put forward that are unbiased or almost unbiased. We refer to Lahiri (1951), Raj (1954), Rao and Vijayan (1977) and others for some interesting results in this area.

Hartley and Ross (1954) obtained an exactly unbiased ratio estimator, starting from a biased estimator $X\bar{r}$ where, $\bar{r} = \frac{1}{n}\sum_{i\in s}\frac{y_i}{x_i}$, after eliminating the derived bias. This process of bias elimination can be applied to other biased estimators as well. In Chapter 4, we modify the Hartley and Ross (1954) estimator to propose a new exactly unbiased estimator suitable for application in a practical survey situation. We derive its variance and also obtain an unbiased estimator for its variance.

Another challenging problem faced while conducting surveys, is to gather reliable data on stigmatizing or sensitive topics such as drug addiction, induced abortion, drunken driving, habitual tax evasion, excessive gambling, etc. This is difficult since many respondents may be reluctant to truthfully answer direct questions on such topics. To overcome this difficulty, Warner (1965) pioneered a randomized response (RR) technique for estimating the proportion of persons in U belonging to some sensitive group A, say, on the basis of an SRSWR. Several authors have made significant contributions in this area, for instance, Horvitz et al. (1967), Horvitz et al. (1976), Greenberg et al. (1969), Mukerjee (1981), Mangat and Singh (1990), Mangat et al. (1995), Arnab (1999), Christofides (2003, 2005), Chaudhuri and Pal (2008) among others. For an excellent exposition on different developments in the area of randomized response, we refer to Chaudhuri and Mukerjee (1988). We give some details below.

In Warner's method, each respondent is provided with a randomization device by which he/she chooses one of two questions 'Do you belong to A', or 'Do you belong to A^c', with respective probabilities p and $1-p$; where $p \neq 1/2$. The respondent is asked to give a truthful 'yes' or 'no' answer to the question chosen by him/her; the interviewer does not see the question chosen and only records the 'yes' or 'no' answer. Thus, the randomization device protects the privacy of the respondent and so it is believed that he/she may be willing to cooperate by responding truthfully. Warner (1965)'s estimator for the proportion of persons belonging to group A and an estimator for its variance, utilizes the number of 'yes' and 'no' responses, with $p \neq 1/2$. Chaudhuri (2001) gave an equivalent version of Warner's (1965) randomized response device (RRD) as detailed below. It may be noted that Chaudhuri (2001)'s version allows one to use Warner's RRD no matter how a sample is chosen.

Consider a box consisting of cards marked either A or A^c in proportions

$p : (1-p)$, $0 < p \neq \frac{1}{2} < 1$. In Warner's RR method, a person chosen from U on the k^{th} draw ($k = 1, 2, \ldots$) chooses randomly one card from the box and generates a truthful randomized response (RR) as

$$\begin{aligned} I_k &= 1 \text{ if card type chosen on } k\text{th draw matches the person's true} \\ &\quad \text{category } A \text{ or } A^c \\ &= 0, \quad \text{otherwise}; \quad k = 1, \ldots, n. \end{aligned} \tag{1.2.17}$$

Suppose y is an indicator variable such that y_i is 1 or 0 corresponding to whether unit i bears the sensitive attribute A or its complement A^c. Let

$$\theta = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{1.2.18}$$

Clearly, this population mean θ is the proportion of people in U bearing the sensitive characteristic. Our objective is to unbiasedly estimate θ. Define

$$r_k = \frac{I_k - (1-p)}{(2p-1)}, \quad k = 1, \ldots, n.$$

Then writing E_R, V_R and C_R for RR based expectations, variances and covariances, we have

$$\begin{aligned} E_R(r_k) = y_k, \quad V_R(r_k) = \frac{p(1-p)}{(2p-1)^2} = \Phi_W, \text{ say,} \\ \text{and } C_R(r_k, r_{k'}) = 0 \ \forall \ k, k' (k \neq k'). \end{aligned} \tag{1.2.19}$$

Warner (1965) permits only SRSWR for his theory. Let $n' = \sum_{k=1}^{n} I_k$, $\lambda = P[I_k = 1] = 1 - p + \theta(2p-1)$ and $\hat{\lambda} = \frac{n'}{n}$. Then, with $p \neq \frac{1}{2}$,

$$\hat{\theta}_W = \frac{1}{n} \sum_{k=1}^{n} r_k = \bar{r}(n) = \frac{\hat{\lambda} + p - 1}{2p-1} \tag{1.2.20}$$

is Warner's (1965) unbiased estimator for θ. Analogous to $\bar{r}(n)$, we define

$$\bar{y}(n) = \frac{1}{n} \sum_{k=1}^{n} y_k.$$

Then using (1.2.19), the expectation and variance of $\hat{\theta}_W$ are given by

$$\begin{aligned} E(\hat{\theta}_W) &= E_P E_R(\hat{\theta}_W) = \theta, \\ V(\hat{\theta}_W) &= E_P V_R(\hat{\theta}_W) + V_P E_R(\hat{\theta}_W) \\ &= E_P(\tfrac{\Phi_W}{n}) + V_P(\tfrac{1}{n} \textstyle\sum_{k=1}^{n} y_k) \\ &= \tfrac{\Phi_W}{n} + V_P(\bar{y}(n)) \\ &= \tfrac{\Phi_W}{n} + \tfrac{\sigma^2}{n}, \end{aligned} \tag{1.2.21}$$

where E_P, V_P generically denote the expectation and variance operators with respect to the sampling scheme for the selection of respondents, and $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ with $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \theta$. Thus $\sigma^2 = \theta(1-\theta)$ which yields the formula

$$V(\hat{\theta}_W) = \frac{\Phi_W}{n} + \frac{\theta(1-\theta)}{n} \tag{1.2.22}$$

In Chapters 5 and 6, we study the estimation of θ using randomized responses obtained by Warner's randomization device.

There may be situations where both A and A^c may be sensitive. For such surveys, the unrelated-question model studied by Horvitz et al. (1967, 1976), Greenberg et al. (1969) and others may be adopted. In this model, instead of asking questions related to A or A^c, one asks a sampled person to respond 'yes' or 'no' truthfully about bearing attribute A, with probability p or bearing another attribute, say B, with probability $1-p$, B being unrelated, in the sense of statistical association, to A. The feature B and its complement B^c should both be innocuous, e.g., B might mean a person being born in January. Since this is unrelated to the sensitive attribute A, one can expect that the respondent will be more confident about privacy protection. However, in this approach, there is an added difficulty as the true proportion bearing B may also be unknown. In this case, two independent samples may be drawn; the randomization device being such that a respondent from sample i chooses the question about belonging to group A with probability p_i, $i = 1, 2$ with $p_1 \neq p_2$. Many researchers have studied this model, for instance, Moors (1971), Folsom et al.(1973), Lanke (1975) and Dowling and Shachtman(1975). Chaudhuri (2001) studied this for general schemes of sample selection while others restricted to SRSWR alone. We consider the unrelated question model in Chapter 7.

There are many other developments in the RR technique suggested by various authors. For instance, Abul-Ela et al. (1967), Bourke (1978), Tamhane (1981), Mukerjee (1981) and others extended these methods to polychotomous sensitive attributes and to multi-attribute situations; whereas extension to quantitative characteristics were made by Greenberg et al. (1971), Chow and Liu (1973), Eriksson (1973), Godambe (1980), Duffy and Waterton (1984) and others. Kuk (1990) proposed a scheme which is applicable to both qualitative and quantitative characteristics. In this procedure, unseen by the interviewer, each respondent generates values from

two known distributions G_1 and G_2 and reports the value from G_1 if he belongs to group 'A' and the value from G_2 otherwise. Thus, the resulting data is generated from a mixture of G_1 and G_2 and problem reduces to that of estimating a mixing proportion. Another advantage of this procedure is that it allows repetition of trials in a natural way, simply by asking each sampled individual to repeat the procedure several times, if they agree to do so. This way, one can get more data leading to a more precise estimate at no increase in sampling cost. We study Kuk's (1990) model in Chapters 5 and 6.

Christofides (2003) proposed an alternative randomized response technique which improves upon Warner's (1965) technique. In this new procedure, from a box containing cards marked $1, 2, \ldots, M$ mixed in proportions $p_1, p_2, \ldots, p_M$ with $\sum_{j=1}^{M} p_j = 1$, a card is drawn at random. Assume that the number drawn is k. If the respondent belongs to the group A^c, he is instructed to report the number k, otherwise, the number $M + 1 - k$. Of course, Warner's (1965) procedure is a special case obtained by specific choice of the parameters in Christofides (2003) model. This model has been studied in Chapters 5 and 6.

In Warner's original method, each time a person is sampled by SRSWR, a fresh RR is generated. In contrast to this method, Gould, Shah and Abernathy (1969), Liu and Chow (1976), Franklin (1989) and others recommended that a pre-assigned number of multiple responses are generated each time a person is selected by SRSWR. However, as remarked by them, such repeated requests for giving a truthful answer may impede respondent cooperation. We do not study this multiple response method in this thesis.

Mangat and Singh (1990) gave a modification of Warner's technique and obtained conditions when their proposed method is more efficient than Warner's (1965) method when the respondents may or may not be truthful in their answers. In their method, each sampled person is selected by SRSWR scheme and given two randomization devices, or two boxes. The first box contains cards marked '*True response*' or '*Randomized response*', in a known proportion, say $T : (1 - T)$; and the other box is the box used as the randomization device in Warner's method, as described earlier. The respondent is instructed to draw a card from the first box and if he/she draws a '*True*' marked card, then he/she should divulge truthfully whether

he/she bears the stigmatizing attribute or not. If he/she draws a '*Randomized response*' marked card, he/she should generate a RR as in Warner's method. As the interviewer does not see the steps of the trial implemented by the respondent, there is no loss in privacy. Mangat and Singh (1990) remarked that his new estimator is superior to Warner's usual estimator if T satisfies certain condition. We apply this modification, to some RR models proposed in the literature, *irrespective of how the sample is drawn*, in Chapter 7.

Chaudhuri and Mukerjee (1988) proposed a technique known as the *Forced Response Model* where the RRD contains three types of cards mixed in proportions $p_1 : p_2 : (1 - p_1 - p_2)$. A respondent is asked to draw a card from the box and report 'yes' if a card of the first type is drawn, 'no' if a card of the second type is drawn. If a card of the third type is drawn, the respondent is asked to truthfully report 'yes' or 'no' according as whether he/she bears the characteristic A or not. Thus a respondent reports forcefully 'yes' with probability p_1, or forcefully 'no' with probability p_2 or provides an honest response 'yes' or 'no' with probability $(1 - p_1 - p_2)$. We have applied Mangat and Singh's (1990) 2-stage RR device to this model in Chapter 7.

Singh and Joarder (1997) gave another model revising Warner's model where if a selected respondent by SRSWR scheme belongs to group A^c, then he/she is requested to report as in Warner's device, but if he/she belongs to group A, then he/she is requested to repeat the trial in Warner's randomization device if in the first trial he/she does not draw the statement according to his/her status. The details of the trial implemented by the respondent remain unknown to the interviewer. This model is called the *Unknown Repeated Trial* model. We apply Mangat and Singh's (1990) 2-stage RR to this model in Chapter 7.

Eichhorn and Hayre (1983) considered the scrambled randomized response technique to estimate the population mean of quantitative sensitive variable. His procedure involves the reporting of a randomized response by multiplying the actual value with a random number drawn from a known distribution having known mean and variance. We consider a version of this procedure in Chapter 7.

1.3 Summary of the thesis

We now give a summary of the seven chapters of this thesis. The terms and notation used are as introduced in Sections 1.1 and 1.2.

CHAPTER 1: This chapter consists of 3 sections. In the first section we define some general terms and notation which we use in subsequent chapters. In the next section we present a brief review of the literature related to the main topics covered in this thesis. The last section consists of a chapter-wise summary.

CHAPTER 2: In this chapter we study an inherent practical limitation in the technique of adaptive sampling. This problem arises because in adaptive sampling, as was explained in Section 1.2, one keeps on adding neighbouring units until one reaches an edge unit, and thus, the size of the final adaptive sample may far exceed that of the initial sample, resulting in a substantial additional pressure on the resources. Salehi and Seber (1997, 2002) addressed this issue and gave certain solutions. In Chapter 2, we propose an alternative simple remedy by introducing a built-in procedure to keep the final adaptive sample size in check. We propose that after ascertaining the sets $A(i)$ for $i \in s$, one restricts the determination of the values of y_j's for only some suitable subsets of $A(i)$ to be drawn by SRSWOR. We thus modify adaptive sampling to a *size-constrained* adaptive sampling which is easier to implement and can keep the final sample size in check. For this method, we obtain an unbiased estimator $\hat{Y}$ for Y, $V(\hat{Y})$ and an unbiased estimator for $V(\hat{Y})$.

We illustrate our method using data from Economic Census 1990 of India to estimate the total numbers of earners through ten small-scale cottage industries in the rural unorganized sector. We use our proposed method of adaptive sampling, based on an initial sample selected through a stratified two-stage sampling plan, with RHC scheme of sampling in both stages. In the process, we also revise the definitions of 'neighbourhood' and 'network' required for the method applied to make our presentation more realistic. Additionally, we use the generalized regression technique for possible improvement upon traditional estimators. We numerically illustrate how one may use adaptive sampling to get improved estimates and at the same time, keep the final sample size in check.

Some of the results presented in this Chapter have been published in

Chaudhuri, Bose and Dihidar (2005).

CHAPTER 3: Särndal (1996) pointed out that both the variance estimators for the HTE , as given by Horvitz and Thompson (1952) and the subsequent one given by Yates and Grundy (1953), as shown in (1.2.4) and (1.2.5), respectively, suffer from the shortcoming that the computation of π_{ij} is very difficult for many standard sampling schemes. This causes a problem if one intends to apply the admissible HTE for estimation in some surveys. For example, if a sample is drawn by the RHC scheme, and one needs to compute the HTE from this sample, it is difficult to do so and then estimate its variance. So, in the first part of this chapter, we examine the feasibility of basing the HTE on a sample drawn by the RHC scheme. For this, we derive algebraic expressions for the inclusion probabilities π_i and π_{ij} as given in (1.2.1) and (1.2.2), for an RHC sample. We see that these inclusion probabilities are all positive and so, one may easily employ the HTE based on a sample obtained via the RHC scheme to unbiasedly estimate Y, and then compute any of the above two variance estimators of it.

Next, in this chapter, we obtain some other alternative unbiased variance estimators for the RHCE and the HTE for estimating Y from an RHC sample. For ease of presentation, our estimators are shown for single-stage sampling, but they may easily be extended to multistage sampling. We then compare the accuracy in estimation by these alternative procedures by dint of simulation, using the data which was used in Chapter 2 and two-stage sampling with RHC and SRSWOR being used in the two stages. For our comparison, we use three criteria defined in Section 1.1, namely, the ACV, ACP and ARL. Our computations show that for estimating the population total from an RHC sample, the HTE often performs better than the RHCE, even though the sample is drawn by the RHC scheme. Thus, for estimating the population total from an RHC sample, along with the usual RHCE, the HTE also emerges as a viable option.

The results in the first part of this chapter have been published in Chaudhuri, Dihidar and Bose (2006) and another based on the second part has been published in Chaudhuri, Bose and Dihidar (2009c).

CHAPTER 4: In health care services planning, knowledge about prevalence of various diseases in an area is crucial. However, comprehensive

records of the diagnosis or treatment of these diseases are sometimes not available. This necessitates data collection by surveys and subsequent estimation of these prevalence rates from survey data, which may not always be straightforward. For example, stroke is a neurological disease which is a leading cause of human morbidity and mortality, but in India, systematic data on stroke is lacking and reliable figures for the prevalence rate of stroke in India are hard to come by. The results in Chapter 4 were developed in the course of a survey undertaken to estimate this rate.

To get an estimate of the stroke prevalence rate, a two-stage sample survey was conducted in the city of Kolkata, India in 2005. It was found that the usual Ratio estimator or the Hartley & Ross (1954) estimator was not effective in estimation for this study and a new estimator had to be developed. We here obtain the form of the new estimator for a two-stage sampling scheme with SRSWOR used in both stages as it is easy to implement in large scale surveys. An unbiased estimator for the variance of this estimator is also obtained. The method is illustrated with the data collected in the survey mentioned above. Finally, it is investigated how a suitable modeling may lead to improved estimators.

The results in this chapter have been accepted for publication as Bose, Chaudhuri, Dihidar and Das (2009).

CHAPTER 5: In this chapter we focus on unbiased estimation of the proportion of persons bearing a sensitive characteristic by Warner's device using multiple responses from distinct persons sampled. We may recall from Section 1.2 that for applying Warner's randomized response device, an SRSWR with a pre-determined number of draws is taken and the over-all sample mean of a linear transform of the gathered randomized responses (RR's) is used to unbiasedly estimate the required proportion; (see (1.2.17)). The problem studied in this chapter stems from the observation that in direct response surveys Basu(1958), Pathak (1962), Korwar and Serfling (1970) and others have shown that if one uses the responses from the distinct units sampled by SRSWR method, alternative unbiased estimators performing better than the classical estimator, namely the sample mean, are available. A natural step is to investigate the counter-part involving the randomized responses instead of direct responses, and in Chapter 5 we study this problem.

Mangat et al. (1995) gave an unbiased estimator based on only one ran-

domized response for every distinct unit sampled and studied the relative efficiencies. In this Chapter we propose some alternative unbiased estimators, together with unbiased estimators of their variances. We compare the efficiencies of these new estimators against certain known competitors. In this chapter, besides using Warner's device, we also study this problem with Kuk's(1990) and Christofides' (2003) randomized response devices.

A paper based on the results in this chapter for Warner's device has been accepted for publication as Chaudhuri, Bose and Dihidar (2009a).

CHAPTER 6: In Chapter 6, we continue with the same problem as in Chapter 5, but now, unlike the usual practice of taking an SRSWR sample of size n, we adopt an alternative inverse sampling plan. Here we pre-fix a number, say ν, of distinct persons we intend to cover, but we permit a random number of $n(\geq \nu)$ draws with equal probabilities with replacement that we may need to realize this. In the context of direct response surveys, it is again well known from Basu (1958), Chikkagoudar (1966) and Lanke (1975) that for estimating a population mean, if one uses the responses obtained from the distinct units sampled by such an inverse sampling method, then estimators performing better than the usual sample mean are available. In this chapter, we investigate if such a result also holds for randomized response surveys. For this problem, we propose some alternative estimators for the population proportion based on randomized responses collected from each person selected by this inverse sampling. We compare the efficiencies of these estimators and present some illustrative situations for a comparison, as is done in Chapter 5. For the sake of simplicity, we first consider randomized responses gathered by using Warner's (1965) randomized response device. Next we also cover the same problem by Kuk's (1990) and Christofides' (2003) randomized response devices.

A paper based on the results in this chapter for Warner's device has been accepted for publication as Chaudhuri, Bose and Dihidar (2009b).

CHAPTER 7: In this chapter, we continue with our study on randomized responses, but in contrast to Chapters 5 and 6, where we restricted our sampling to SRSWR, here we allow any general sampling scheme for choosing the respondents.

For this, we consider the technique of Mangat and Singh (1990) who used SRSWR and we show that *irrespective of how a sample is drawn*, this technique may be profitably used with the unrelated question model of

Horvitz et al. (1967), Greenberg et al. (1969), Singh and Joarder (1997)'s model and also the Forced Response Model of Chaudhuri and Mukerjee (1988). Next, as opposed to the models for a stigmatizing qualitative variable considered so far, we also consider modifying a model for the case where the stigmatizing variable is quantitative and the objective is to estimate the population mean for this variable. For each of the above modified models, relevant unbiased estimators for the parameters under study and unbiased variance estimators are derived for a general sampling scheme. On algebraically comparing the variances for the original and modified versions of these models, we obtain conditions on the relevant RR device parameters such that on constructing the devices with parameters satisfying these conditions, an improvement may be achieved by this modification. Finally, we give some numerical illustrations showing the estimated efficiencies for different estimators, when the respondents are chosen by the RHC scheme of sampling.

A paper based on the results in this chapter has been prepared as technical report No: ISI/ASD/2008/8 and submitted for publication as Dihidar (2009).

Chapter 2

Sample-size-restrictive adaptive sampling: an application in estimating localized elements

ABSTRACT: In some populations it is found that the value of the variable of interest for many sampling units is either zero or negligibly low while for some other units these values are substantial, these high valued units being highly localized in certain segments of the population. In such cases if one wants to use sample surveys to estimate the population characteristics for the variable of interest, then estimation may be inaccurate if a chosen sample fails to capture enough of the high valued units. In such situations, the technique of adaptive sampling is found useful. However, the size of an adaptive sample may often far exceed that of the initial sample. In this chapter we present a method of keeping the adaptive sample-size in check. To examine the efficacy of this method, we illustrate it by applying this method to estimate total numbers of rural earners through specific vocations in a given district in India, simultaneously for several vocations.

2.1 Introduction

We first briefly introduce the specific sampling problem which motivated the study described in this chapter. In India, it is important to estimate

the total numbers of persons earning through specific small-scale industries in the unorganized sector in the various districts. Though the earning of an individual in this sector is small, it is believed that their collective contribution to the nation's gross domestic product(GDP) is substantial. Hence it is of national interest to estimate this figure. However, obtaining a reliable estimate through a traditional sample survey is difficult since these earners are often concentrated in small regional pockets. Moreover, they frequently change occupations and locations. Consequently, traditional sampling designs may fail to capture these sparsely scattered industry-wise rural earners in the unorganized sector.

Chaudhuri, Bose and Ghosh (2004) showed that adaptive sampling may be effectively used in this situation. However, in that work it was evident, as expected, that the adaptive sample size may far exceed that of an initial sample.

We use the same data as used by Chaudhuri et al. (2004) to illustrate our proposed procedure for keeping the adaptive sample size in check. In Section 2.2, we modify the sampling procedure of Chaudhuri et al. (2004) and choose our initial sample through a stratified two-stage sampling design with RHC scheme used in both stages. For this design, we obtain the estimators for Y and $V(\hat{Y})$. In Section 2.3, we obtain some alternative estimators by using Cassel, Sarndal and Wretman's (1976) generalized regression (greg) technique as a possible improvement on the usual estimator in Section 2.2. In Section 2.4, we show how the EC data motivates the use of adaptive sampling and how we may suitably define the neighbourhood and network for adaptive sampling. In this section, we also show how to modify adaptive sampling to keep the size of the adaptive sample in check and then, we derive expressions for $\hat{Y}$, its variance and its unbiased variance estimator for this modified method. Using EC data in Section 2.5, we numerically illustrate how we may achieve improved estimation while restricting the sample size. Our comments and recommendations are given in the concluding Section 2.6.

2.2 A practical survey problem and choice of initial sample

For our study, we consider the district Birbhum in the state West Bengal, India, and use available data from the Economic Census(EC), 1990 and the Indian Population Census (IPC), 1991.

This district has 21 administrative blocks. We first group them into 3 strata of 7 each, taking account of the aggregated numbers of earners through all the 10 rural industries in strata formation. Next, from each stratum, 3 blocks are selected adopting Rao, Hartley and Cochran's (RHC, 1962) sampling scheme described in Section 1.2 and using the group sizes 2,2 and 3, as given by (1.2.13). The total numbers of earners in a block through all these 10 industries together is available from the EC, 1990 and this value is taken as the size-measure for the block. From each chosen block , 20% of its villages, rounded upward to integers, are again sampled following the RHC scheme. For this village selection, the village-population-size as known from the IPC, 1991 is taken as the size-measure.

We consider 10 rural industries and our objective is to estimate the total numbers of earners for each of the 10 industries separately. To avoid cumbersome notation, we use y to denote the variable of interest, namely, the number of earners through any one particular industry. Thus, y_i generically denotes the number of earners through this industry in the ith block. Then, our parameter of interest is Y as in (1.1.1) with $N = 21$, and this denotes the total number of earners through this industry in the entire district. Thus, the same notation serves for all the 10 industries, simply keeping in mind that when we are estimating the total number of earners for any one chosen industry, we start with the value of y_i for that industry.

For simplicity, we write y_i and p_i as the variate value and the normed size measure of the unit drawn from the ith group. Then, for estimating Y with this RHC scheme of sampling, the unbiased estimator is as in (1.2.12). If y_i cannot be observed, as in the present case where the ith block is composed of M_i villages, then m_i of the villages are again to be selected by the RHC method. Let y_{ij} denote the y-value for the jth village of the ith block. Moreover, $\Sigma_{mi}, M_{ij}, Q_{ij}$ will denote entities corresponding respectively to

Σ_n, N_i and Q_i. Then, an unbiased estimator for y_i will be

$$\hat{y}_i = \Sigma_{mi} \frac{Q_{ij}}{p_{ij}} y_{ij}.$$

Using this in (1.2.12), we get an unbiased estimator for Y as

$$\hat{Y} = \Sigma_n \frac{Q_i}{p_i} (\Sigma_{mi} \frac{Q_{ij}}{p_{ij}} y_{ij}), \text{ say.} \tag{2.2.1}$$

Now, from (1.2.14) and a result in Chaudhuri, Adhikary and Dihidar (2000) for multistage sampling, it follows that an unbiased estimator for $V(\hat{Y})$ is

$$\begin{aligned} \hat{V}(\hat{Y}) &= B\Sigma_n\Sigma_n Q_i Q_{i'} (\tfrac{\hat{y}_i}{p_i} - \tfrac{\hat{y}_{i'}}{p_{i'}})^2 + \Sigma_n \tfrac{Q_i}{p_i} B_i \Sigma_{mi}\Sigma_{mi} Q_{ij} Q_{ij'} (\tfrac{y_{ij}}{p_{ij}} - \tfrac{y_{ij'}}{p_{ij'}})^2 \\ \text{where} & \\ B &= (\Sigma_n N_i^2 - N)/(N^2 - \Sigma_n N_i^2); \\ B_i &= (\Sigma_{mi} M_{ij}^2 - M_i)/(M_i^2 - \Sigma_{mi} M_{ij}^2); \end{aligned} \tag{2.2.2}$$

and we write $\Sigma_n\Sigma_n$ as the sum over non-overlapping pairs of the n groups of blocks, $\Sigma_{mi}\Sigma_{mi}$ as that over the villages.

2.3 Use of generalized regression technique

In this section, we use the generalized regression (greg) technique of Cassel, Särndal and Wretman (1976) for a possible improvement upon the above estimator $\hat{Y}$ obtained in (2.2.1). We may also mention the study by Tam (1988), Särndal (1980) etc. for important contribution in this area. A greg estimator was not discussed by Chaudhuri et al. (2004). The motivation of greg is to improve upon the original estimator by using an auxiliary variable which has a good positive correlation with y, the variable of interest.

As the regressor, we use the total number of all non-agricultural workers in a village, as given in the IPC. We note that this is different from the size variable used for block selection which is the total number of earners through the 10 industries as obtained from EC. Let x_{ij} denote the number of non-agricultural workers in the jth village of ith block, $x_i = \Sigma_{M_i} x_{ij}$ and $X = \Sigma_N x_i$. We consider the following alternative models for motivating some greg estimators which may be considered for possible improvements

over $\hat{Y}$.

$$\begin{aligned} \text{Model} \quad M_1: \quad & y_{ij} = \beta_i x_{ij} + \text{ error} \\ \text{Model} \quad M_2: \quad & y_{ij} = \beta x_{ij} + \text{ error} \\ \text{Model} \quad M_3: \quad & y_i = \beta^* x_i + \text{ error} \end{aligned}$$

The common slopes in models M_2 and M_3 are used as usual for 'borrowing of strength' as is done in this context of small-area estimation. Of course, it is possible to further generalize these estimators, for instance, by incorporating quadratic terms. However, as the above models are traditionally and widely used in small-area estimation, we focus our attention to these versions only and obtain estimators based on them. Along similar lines, our techniques should also lead to estimators if we start from the generalized versions of the above models.

M_1 and M_2 are used to first get estimates of y_i using y_{ij} values from the second stage sample and corresponding x_{ij} values from IPC data. Let these estimators be denoted by $\hat{y}_i(1)$, $\hat{y}_i(2)$, respectively. Clearly,

$$\hat{y}_i(1) = \Sigma_{mi} \frac{Q_{ij}}{p_{ij}}(y_{ij} - \hat{\beta}_i x_{ij}) + \hat{\beta}_i x_i, \quad \text{and} \quad \hat{y}_i(2) = \Sigma_{mi} \frac{Q_{ij}}{p_{ij}}(y_{ij} - \hat{\beta} x_{ij}) + \hat{\beta} x_i. \tag{2.3.1}$$

These are then used to find estimates of Y. M_3 is used to derive regression estimates of Y by using the estimates of y_i from M_2 and the x_i values from the IPC. M_3 is derived from M_2, but we shall show below that M_3 motivates specific regression estimators for Y.

The estimators of Y as motivated by M_1, M_2 and M_3, respectively, are denoted by g_{11}, g_{12} and g_{22}, and they are given by

$$\begin{aligned} g_{11} &= \Sigma_n \tfrac{Q_i}{p_i}[\Sigma_{mi} \tfrac{Q_{ij}}{p_{ij}}(y_{ij} - \hat{\beta}_i x_{ij}) + \hat{\beta}_i x_i] \\ g_{12} &= \Sigma_n \tfrac{Q_i}{p_i}[\Sigma_{mi} \tfrac{Q_{ij}}{p_{ij}}(y_{ij} - \hat{\beta} x_{ij}) + \hat{\beta} x_i] \\ g_{22} &= g_{12} + \hat{\beta}^*(X - \Sigma_n \tfrac{Q_i}{p_i} x_i) \end{aligned} \tag{2.3.2}$$

In the above,

$$\hat{\beta}_i = \frac{\Sigma_{mi} y_{ij} x_{ij} R_{ij}}{\Sigma_{mi} x_{ij}^2 R_{ij}}, \quad \hat{\beta} = \frac{\Sigma_n \frac{Q_i}{p_i} \Sigma_{mi} y_{ij} x_{ij} R_{ij}}{\Sigma_n \frac{Q_i}{p_i} \Sigma_{mi} x_{ij}^2 R_{ij}}, \quad R_{ij} = \frac{(1 - \frac{p_{ij}}{Q_{ij}})}{(p_{ij} x_{ij}/Q_{ij})},$$

$$\hat{\beta}^* = \frac{\Sigma_n [\Sigma_{mi} \frac{Q_{ij}}{p_{ij}}(y_{ij} - \hat{\beta} x_{ij}) + \hat{\beta} x_i] x_i R_i}{\Sigma_n x_i^2 R_i}, \; R_i = (1 - \frac{p_i}{Q_i})/(\frac{p_i x_i}{Q_i}).$$

Here R_{ij}, R_i parallel the choice $\frac{1-\pi_i}{\pi_i x_i}$ in the greg estimator of Cassel et al. (1976). Using (2.3.1), we may write the estimators in (2.3.2) in more com-

pact form as

$$\begin{aligned} g_{11} &= \Sigma_n \tfrac{Q_i}{p_i} \hat{y}_i(1) \\ g_{12} &= \Sigma_n \tfrac{Q_i}{p_i} \hat{y}_i(2) \\ g_{22} &= \Sigma_n \tfrac{Q_i}{p_i} \hat{y}_i(2) h_i, \end{aligned} \tag{2.3.3}$$

where we write

$$h_i = 1 + \frac{x_i R_i \left\{ \dfrac{X p_i}{Q_i} - x_i - \dfrac{p_i}{Q_i} \left(\sum\limits_{j \neq i=1}^{n} \dfrac{Q_j}{p_j} x_j \right) \right\}}{\sum_n x_i^2 R_i}. \tag{2.3.4}$$

Next, we define

$$e_{ij}(1) = y_{ij} - \hat{\beta}_i x_{ij}, \quad e_{ij}(2) = y_{ij} - \hat{\beta} x_{ij} \text{ and } e_i = \hat{y}_i(2) - \hat{\beta}^* x_i \tag{2.3.5}$$

Then, the estimators of MSE's of g_{11}, g_{12} and g_{22} can be, respectively, expressed as

$$v_1 = B \Sigma_n \Sigma_n Q_i Q_{i'} \left(\frac{\hat{y}_i(1)}{p_i} - \frac{\hat{y}_{i'}(1)}{p_{i'}} \right)^2 + \Sigma_n \frac{Q_i}{p_i} B_i \Sigma_{mi} \Sigma_{mi} Q_{ij} Q_{ij'} \left(\frac{e_{ij}(1)}{p_{ij}} - \frac{e_{ij'}(1)}{p_{ij'}} \right)^2,$$

$$v_2 = B \Sigma_n \Sigma_n Q_i Q_{i'} \left(\frac{\hat{y}_i(2)}{p_i} - \frac{\hat{y}_{i'}(2)}{p_{i'}} \right)^2 + \Sigma_n \frac{Q_i}{p_i} B_i \Sigma_{mi} \Sigma_{mi} Q_{ij} Q_{ij'} \left(\frac{e_{ij}(2)}{p_{ij}} - \frac{e_{ij'}(2)}{p_{ij'}} \right)^2$$

and

$$v_3 = B \Sigma_n \Sigma_n Q_i Q_{i'} \left(\frac{e_i h_i}{p_i} - \frac{e_{i'} h_{i'}}{p_{i'}} \right)^2 + \Sigma_n \frac{Q_i}{p_i} B_i \Sigma_{mi} \Sigma_{mi} Q_{ij} Q_{ij'} \left(\frac{e_{ij}(2)}{p_{ij}} - \frac{e_{ij'}(2)}{p_{ij'}} \right)^2,$$

where B and B_i are as in (2.2.2) and other terms as in (2.3.1), (2.3.4) and (2.3.5).

If enough non-zero valued y_{ij}'s are not covered in a selected sample then improvements upon $\hat{Y}$ may not be effected by any of g_{11}, g_{12} or g_{22}. So, it is considered important, rather imperative, to enhance the "information-content" in a realized sample by extending from the initial sample to an adaptive sample and accordingly revise each of $\hat{Y}, g_{11}, g_{12}, g_{22}$ basing them each on an adaptive sample. We take up this study in the next section. For simplicity, here we revise only y_{ij}'s for the adaptive sample but not the x_{ij}'s.

2.4 Adaptive sampling as an application

Table 1 summarizes data from EC, 1990 for the district Birbhum and shows how the earners by 10 specific rural unregistered industries are distributed in the 21 blocks of the district over 1286 villages. The 10 industries in Birbhum district which we shall consider are the following, numbered and coded: 1.Handloom(H), 2.Bamboo(B), 3.Husking(HU), 4. Pottery(P), 5.Silk(S), 6.Stone-breaking(SB), 7. Bidi-manufacturing(BM), 8.Ironsmithy(IS), 9.Carpentry(C) and 10.Paddy-crushing(PC). This same data was used in Chaudhuri, Bose and Ghosh (2004).

Table 1

Showing the distribution of earners in Birbhum

Industry Number and Code	Number of earners by industry	Number of villages with earners	Number of blocks with earners	Ranges of earners industry-wise in blocks	
				Minimum	Maximum
1(H)	4582	199	21	6	1701
2(B)	3715	314	21	18	509
3(HU)	2352	648	21	10	210
4(P)	2012	146	21	1	227
5(S)	1543	19	6	6	1177
6(SB)	3886	36	6	1	1940
7(BM)	1539	154	21	1	309
8(IS)	1523	474	21	30	119
9(C)	1381	372	21	15	123
10(PC)	1139	75	15	2	351
Total	23672	1286	21		

The figures in Table 1 show that there is a wide variation in the distribution of the earners through different industries in Birbhum. While 1523 ironsmiths(8) are spread over 474 villages covering all the 21 blocks, the 1543 workers in the silk industry(5) are concentrated in only 19 villages and 6 blocks; 3715 bamboo industry(2) earners are found in 314 villages and all 21 blocks while the 3886 stone-breakers(6) are localised over only 36 villages covering only 6 of the 21 blocks; paddy-crushers(10) are found in only 75 villages in 15 blocks.

We need to choose a suitable sample in order to estimate the total numbers of these 23,672 earners by these 10 separate industries. However,

the localization of earners as evident from Table 1 indicates that it is difficult to employ a standard sampling design to catch these earners in sufficient numbers to throw up useful estimates.

Again, from Table 1 it seems that if an initial sample is chosen with selection probabilities that make no use of the distribution of the earners among the respective villages, for example if there is no or scant representation in the sample of the 19 villages with silk related or of the 75 villages accommodating the paddy-crushing earners, then appropriate estimation of the numbers of earners through these industries will be of dubious levels of accuracy.

So, in the present empirical situation we are motivated to implement an adaptive sample by studying the mutual relations of association among the 10 industries across the 1286 villages in our illustrated district by a reference to the Table 2 below, partially reproduced from Chaudhuri et al. (2004). The figures in the parentheses indicate percentages of the values in the rows for the respective columns in terms of the diagonals to which the respective columns correspond.

Table 2 indicates for example, that of the 19 villages in which silk-earners live, 63.16% have earners by Husking and 47.37% have Iron-smiths. So, if the sample contains some of the 12 villages with earners by Husking, or some of the 9 with Iron-smiths, then, through these sampled villages some of the silk-earners may be reached. Such associations among the industries through village-wise co-inhabitance may be utilized to get estimates of the number of silk-related earners. Similarly, this can be done for the other industries as well. In this way, Table 2 is exclusively used to effectively construct networks of the villages exploiting this association to enhance the information content in a suitably extended adaptive sample.

Table 2

Presenting the respective numbers of villages with earners industry-wise showing a specimen of association of the industries in the district

	1(H)	3 (HU)	4(P)	5(S)	8(IS)	10(PC)
1(H)	199	121	30	8	90	13
	(100)	(18.67)	(20.55)	(42.11)	(18.99)	(17.33)
3(HU)	121	648	76	12	27	33
	(60.80)	(100)	(52.05)	(63.16)	(5.70)	(44.00)
4(P)	30	76	146	3	63	13
	(15.08)	(11.73)	(100)	(15.79)	(13.29)	(17.33)
5(S)	8	12	3	19	9	4
	(4.02)	(1.85)	(2.05)	(100)	(1.90)	(5.33)
8(IS)	90	272	63	9	474	26
	(45.23)	(41.98)	(43.15)	(47.37)	(100)	(34.67)
10(PC)	13	33	13	4	26	75
	(6.53)	(5.09)	(8.90)	(21.05)	(5.49)	(100.00)

However, the above adaptive sample may result in sample-coverage beyond one's means. So, we now modify adaptive sampling into a "Size-constrained Adaptive Sampling". To implement this our proposal is that after ascertaining the sets $A(i)$, for $i \epsilon s$, one confines the determination of the values of y_j's only for $B(i), i \epsilon s$, where $B(i)$'s are suitable subsets of $A(i)$ to be drawn by simple random sampling without replacement. Writing l_i as the cardinality of $B(i)$, we suggest that $B(i)$'s are to be so chosen that $\sum\limits_{i \epsilon s} l_i$ may not exceed a predetermined limit, say, L, which may be fixed as a certain fraction of $\sum\limits_{i \epsilon s} m_i$. The size-constrained Adaptive sample corresponding to s is then $B(s)$, which is the union of $B(i)$ over i in s. Corresponding to t_i in (1.2.15), we now define

$$e_i = \frac{1}{l_i} \sum_{j \epsilon B(i)} y_j$$

and employ, instead of $\hat{T} = t(s, t_i | i \epsilon s)$ as in (1.2.16), an estimator $\hat{T} = t(s, e_i | i \epsilon s)$, which is also unbiased for Y.

Let E_2, V_2 denote expectation, variance operators with respect to SRSWOR of $B(i)$ from $A(i)$'s independently across i in s, E_1, V_1 the same over

the initial sampling using the design p and $E = E_1E_2,\ V = E_1V_2 + V_1E_2$ the overall expectation, variance operators as introduced in Section 1.2. Then, as in (1.2.11),

$$E_2(e_i) = t_i, \quad v_2(e_i) = \left(\frac{1}{l_i} - \frac{1}{m_i}\right) \frac{1}{(l_i - 1)} \sum_{j \epsilon B(i)} (y_j - e_i)^2$$

and

$$E_2(v_2(e_i)) = V_2(e_i) = \left(\frac{1}{l_i} - \frac{1}{m_i}\right) \left(\frac{1}{m_i - 1}\right) \sum_{j \epsilon A(i)} (y_j - t_i)^2.$$

So, we can derive an unbiased estimator for $V(t(s, e_i|i\epsilon s))$ if $t(s, y_i|i\epsilon s)$ is an unbiased estimator for Y which has an unbiased estimator for the variance of this estimator.

For example, if $t(s, y_i|i\epsilon s)$ is the HTE as shown in (1.2.4), with s containing a fixed number of distinct units for every s with $p(s) > 0$, then, provided $\pi_{ij} > 0$ for every i, j, on using(1.2.5) and following the arguments in Raj(1968), an unbiased estimator for $V(t(s, e_i|i\epsilon s))$ can be seen to be

$$v = \sum_{i \epsilon s} \frac{v_2(e_i)}{\pi_i} + \sum\sum_{i<j \epsilon s} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j}\right)^2.$$

We recall that the estimator $\hat{Y}$ and the greg estimators g_{11}, g_{12} and g_{22} for the initial stratified two stage sample with RHC scheme employed in both stages, were given in (2.2.1) and (2.3.3). Now, we need to derive corresponding estimators based on the size constrained adaptive sample obtained from this initial sample. This is illustrated below with reference to an unstratified single-stage sampling with the constraining of the size of an adaptive sample. The corresponding expressions for stratified two-stage sampling by RHC method in both the stages is easily obtained from this along usual lines, and hence is omitted here. The usual RHCE is as given by $\hat{Y}_{RHC} = \Sigma_n y_i \frac{Q_i}{p_i}$, as shown in (1.2.12). Correspondingly, we now define

$$f = \Sigma_n t_i \frac{Q_i}{p_i},\ g = \Sigma_n e_i \frac{Q_i}{p_i}.$$

Then, with $A = \frac{\Sigma_n N_i^2 - N}{N(N-1)}$ and B as in (2.2.2),

$$\begin{aligned} V_1(f) &= A(\Sigma_{i=1}^N \frac{t_i}{p_i} - T)^2, \\ v_1(f) &= B\Sigma_n Q_i(\frac{t_i}{p_i} - f)^2 = B(\Sigma_n t_i^2 \frac{Q_i}{p_i^2} - f^2) \end{aligned}$$

$$
\begin{aligned}
v_2(g) &= \Sigma_n (\frac{Q_i}{p_i}^2) v_2(e_i), \\
\text{and } f_i &= e_i^2 - v_2(e_i) \text{ satisfies } E_2(f_i) = t_i^2.
\end{aligned}
$$

So, if we write $v(g) = v_2(g) + B[\Sigma_n \frac{Q_i}{p_i^2} f_i - \{g^2 - \Sigma_n (\frac{Q_i}{p_i})^2 v_2(e_i)\}]$ then we see that

$$
\begin{aligned}
E_2\left[v(g)\right] &= V_2(g) + B\left[\Sigma_n \frac{Q_i}{p_i^2} t_i^2 - V_2(g) - (\Sigma_n \frac{Q_i}{p_i} t_i)^2 + \Sigma_n (\frac{Q_i}{p_i})^2 V_2(e_i)\right] \\
&= V_2(g) + B\left[\Sigma_n \frac{Q_i}{p_i^2} t_i^2 - f^2\right] \\
&= V_2(g) + v_1\left[\Sigma_n \frac{Q_i}{p_i} t_i\right].
\end{aligned}
$$

And so,

$$
\begin{aligned}
E_1 E_2\left[v(g)\right] &= E_1\left[V_2(g)\right] + E_1\left[v_1\left(\Sigma_n \frac{Q_i}{p_i} t_i\right)\right] \\
&= E_1\left[V_2(g)\right] + V_1\left[\Sigma_n \frac{Q_i}{p_i} t_i\right] \\
&= E_1\left[V_2(g)\right] + V_1 E_2\left[\Sigma_n \frac{Q_i}{p_i} e_i\right] \\
&= E_1\left[V_2(g)\right] + V_1\left[E_2(g)\right] = V(g).
\end{aligned}
$$

Hence, $v(g)$ is an unbiased estimator for $V(g)$. This $v(g)$ may be written in a compact form as

$$
v^*(g) = (1+B)\Sigma_n (\frac{Q_i}{p_i})^2 v_2(e_i) + B\left[\Sigma_n \frac{Q_i}{p_i^2} f_i - g^2\right].
$$

We apply this procedure to our data in the next section.

2.5 Simulation-based illustrations of findings.

Using the EC data indicated in Section 1 we first choose any one industry out of the 10 industries and consider the estimation of the total number of earners in the district through this chosen industry. The whole exercise is then repeated for each of the 10 industries.

First, the estimators $\hat{Y}, g_{11}, g_{12}, g_{22}$ as obtained in Section 2.2 are computed. These are based on 2-stage RHC-RHC sampling. Then, we obtain estimators based on adaptive sampling where the initial sample is a 2-stage

RHC-RHC sample. For each village in the Indian district illustrated here, we define its 'neighborhood' to consist of itself and all other villages with a common geographical boundary with it. We take advantage of the figures in Table 2 to form "networks" appropriately. To construct a network for a village, we take specific sets of industries and then check the condition C^* which we take as the existence of at least one earner through any one or more industries of this set. It is interesting to note that, the set of industries considered for defining the network, may or may not include the particular industry for which we are estimating the total. Then, we compute estimates based on adaptive sampling and also estimates based on 'Size-constrained' adaptive sampling, starting from the same initial 2-stage RHC-RHC sample.

We can evaluate the method of 'size-constrained' adaptive sampling for our example, since we have all necessary census data at hand. Towards this, we replicate the sampling, both the original and the adaptive, a total of 1000 times and for each sample, we construct the confidence Interval as in (1.1.3). Then we use the criteria ACP and ARL, as defined in Section 1.1, to compare the performance of the alternative estimators.

Table 3 below compares some alternative estimators for the total numbers of earners through different industries. The values for ACP and ARL are shown for the traditional estimator, greg estimators and also their adaptive-sample versions. Moreover, for a specific replicate, the number (c) of villages chosen in the original sample and the number (d) in the adaptive sample were also computed. But since it was found that d far exceeds c, we had recourse to 'size-restricted' adaptive sampling. We present the values of ACP and ARL, based on SRSWOR's from the networks in various percentages, namely, 8, 10, 15 and 25, rounded upwards to the nearest integer. In addition, we also compute for a specific replicate, the number (a) of villages containing a particular industry in the original sample and the corresponding number (b) in the adaptive sample.

We note that for a specific replicate, one single original sample and its corresponding adaptive sample gives all the estimates for the different industries and so the values of c and d remain constant for all estimates. The value of a depends on the industry for which the estimation is done and b depends on the choice of industry and also the network used. Finally, for estimation for the group of all 10 industries together, $a = c$ and $b = d$.

For ease of understanding, we explain the notation of Table 3 below:

- $\hat{Y}, g_{11}, g_{12}, g_{22}$ are as given by equations (2.2.1) and (2.3.3). These are based on 2-stage RHC-RHC sampling.
- $\hat{Y}^*, g_{11}^*, g_{12}^*, g_{22}^*$ denote modified estimators based on adaptive sampling. The corresponding network used is shown in parentheses. Networks 1, 2, 3 and 4 denote the network formed according to industry sets [1(H), 2(B) & 4(P)], [9(C)], [8(IS) & 9(C)] and [3(HU) & 8(IS)] respectively.
- $\tilde{Y}, \tilde{g}_{11}, \tilde{g}_{12}, \tilde{g}_{22}$ denote estimators based on adaptive sampling by network 4, on which size restriction is applied. The respective percentage(8, 10, 15 or 25) of the adaptive sample-sizes allowed is shown in parentheses.
- a, b, c, d as explained above.

Table 3

Numerical performances of alternative procedures: Few illustrations

Industry Type	Estimator	ACP	ARL	Estimator	ACP	ARL	Estimator	ACP	ARL
$1(H)$	$\hat{Y}$	77.9	1.71	$\hat{Y}^*(1)$	82.5	1.36	$\hat{Y}^*(4)$	85.0	1.28
	g_{11}	77.8	1.70	$g_{11}^*(1)$	83.0	1.37	$g_{11}^*(4)$	85.8	1.28
	g_{12}	77.9	1.71	$g_{12}^*(1)$	83.0	1.37	$g_{12}^*(4)$	85.6	1.28
	g_{22}	78.8	1.90	$g_{22}^*(1)$	85.4	1.60	$g_{22}^*(4)$	87.2	1.53
			$a = 24$			$b = 74$			$b = 64$
	$\tilde{Y}(10)$	77.5	1.98	$\tilde{Y}(15)$	74.7	1.94	$\tilde{Y}(25)$	76.5	1.67
	$\tilde{g}_{11}(10)$	77.4	1.99	$\tilde{g}_{11}(15)$	74.5	1.94	$\tilde{g}_{11}(25)$	75.7	1.67
	$\tilde{g}_{12}(10)$	77.5	1.98	$\tilde{g}_{12}(15)$	74.7	1.94	$\tilde{g}_{12}(25)$	76.1	1.67
	$\tilde{g}_{22}(10)$	77.6	2.10	$\tilde{g}_{22}(15)$	76.2	2.08	$\tilde{g}_{22}(25)$	77.1	1.80
			$b = 35$			$b = 39$			$b = 47$
$2(B)$	$\hat{Y}$	85.5	1.54	$\hat{Y}^*(1)$	88.7	1.29	$\hat{Y}^*(4)$	87.4	1.24
	g_{11}	85.9	1.54	$g_{11}^*(1)$	89.2	1.29	$g_{11}^*(4)$	87.5	1.24
	g_{12}	85.8	1.55	$g_{12}^*(1)$	88.9	1.29	$g_{12}^*(4)$	87.7	1.24
	g_{22}	86.4	1.46	$g_{22}^*(1)$	88.6	1.20	$g_{22}^*(4)$	88.5	1.16
			$a = 34$			$b = 115$			$b = 95$
	$\tilde{Y}(10)$	89.9	1.40	$\tilde{Y}(15)$	86.5	1.31	$\tilde{Y}(25)$	88.6	1.27
	$\tilde{g}_{11}(10)$	89.9	1.40	$\tilde{g}_{11}(15)$	87.0	1.31	$\tilde{g}_{11}(25)$	88.7	1.27
	$\tilde{g}_{12}(10)$	90.0	1.41	$\tilde{g}_{12}(15)$	86.9	1.31	$\tilde{g}_{12}(25)$	88.6	1.27
	$\tilde{g}_{22}(10)$	89.9	1.33	$\tilde{g}_{22}(15)$	86.7	1.25	$\tilde{g}_{22}(25)$	89.3	1.19
			$b = 52$			$b = 60$			$b = 72$
$3(HU)$	$\hat{Y}$	88.6	0.81	$\hat{Y}^*(3)$	89.2	0.82	$\hat{Y}^*(4)$	88.7	0.73
	g_{11}	89.3	0.82	$g_{11}^*(3)$	89.7	0.83	$g_{11}^*(4)$	89.4	0.73
	g_{12}	89.2	0.82	$g_{12}^*(3)$	89.8	0.83	$g_{12}^*(4)$	89.3	0.73
	g_{22}	88.6	0.66	$g_{22}^*(3)$	88.2	0.67	$g_{22}^*(4)$	92.4	0.54
			$a = 73$			$b = 115$			$b = 260$
	$\tilde{Y}(8)$	89.6	0.88	$\tilde{Y}(15)$	90.8	0.81	$\tilde{Y}(25)$	94.3	0.80
	$\tilde{g}_{11}(8)$	89.1	0.87	$\tilde{g}_{11}(15)$	90.8	0.81	$\tilde{g}_{11}(25)$	94.8	0.80
	$\tilde{g}_{12}(8)$	89.4	0.88	$\tilde{g}_{12}(15)$	90.8	0.81	$\tilde{g}_{12}(25)$	95.0	0.80
	$\tilde{g}_{22}(8)$	91.0	0.70	$\tilde{g}_{22}(15)$	91.1	0.67	$\tilde{g}_{22}(25)$	94.1	0.63
			$b = 114$			$b = 154$			$b = 190$

Table 3 (Continued)

Industry Type	Estimator	ACP	ARL	Estimator	ACP	ARL	Estimator	ACP	ARL
$4(P)$	$\hat{Y}$	83.8	1.83	$\tilde{Y}(10)$	87.6	1.81	$\tilde{Y}(25)$	88.0	1.53
	g_{11}	83.8	1.83	$\tilde{g}_{11}(10)$	87.5	1.81	$\tilde{g}_{11}(25)$	88.4	1.53
	g_{12}	83.9	1.83	$\tilde{g}_{12}(10)$	87.4	1.81	$\tilde{g}_{12}(25)$	87.8	1.53
	g_{22}	84.3	1.73	$\tilde{g}_{22}(10)$	87.7	1.70	$\tilde{g}_{22}(25)$	89.6	1.49
			$a = 16$			$b = 22$			$b = 30$
$5(S)$	$\hat{Y}$	49.3	3.13	$\hat{Y}^*(2)$	69.2	3.14	$\hat{Y}^*(4)$	69.8	2.61
	g_{11}	49.8	3.14	$g^*_{11}(2)$	68.8	3.18	$g^*_{11}(4)$	69.7	2.62
	g_{12}	49.6	3.14	$g^*_{12}(2)$	69.0	3.17	$g^*_{12}(4)$	69.8	2.61
	g_{22}	47.9	3.02	$g^*_{22}(2)$	68.5	3.07	$g^*_{22}(4)$	71.8	2.48
			$a = 4$			$b = 5$			$b = 9$
	$\tilde{Y}(8)$	58.1	3.41	$\tilde{Y}(15)$	60.5	3.34	$\tilde{Y}(25)$	63.3	3.21
	$\tilde{g}_{11}(8)$	57.7	3.44	$\tilde{g}_{11}(15)$	60.8	3.42	$\tilde{g}_{11}(25)$	63.9	3.21
	$\tilde{g}_{12}(8)$	57.8	3.42	$\tilde{g}_{12}(15)$	60.5	3.39	$\tilde{g}_{12}(25)$	63.3	3.21
	$\tilde{g}_{22}(8)$	59.4	3.35	$\tilde{g}_{22}(15)$	60.7	3.30	$\tilde{g}_{22}(25)$	63.3	3.10
			$b = 5$			$b = 6$			$b = 7$
$6(SB)$	$\hat{Y}$	65.9	2.62	$\hat{Y}^*(4)$	83.7	1.96	$\tilde{Y}(8)$	86.8	2.37
	g_{11}	65.9	2.60	$g^*_{11}(4)$	83.5	1.97	$\tilde{g}_{11}(8)$	86.9	2.38
	g_{12}	65.8	2.61	$g^*_{12}(4)$	83.5	1.97	$\tilde{g}_{12}(8)$	86.7	2.38
	g_{22}	71.3	2.90	$g^*_{22}(4)$	87.2	2.38	$\tilde{g}_{22}(8)$	89.6	2.71
			$a = 6$			$b = 16$			$b = 9$
$7(T)$	$\hat{Y}$	87.0	1.53	$\hat{Y}^*(4)$	89.5	1.02	$\tilde{Y}(25)$	89.3	1.30
	g_{11}	86.9	1.53	$g^*_{11}(4)$	89.6	1.02	$\tilde{g}_{11}(25)$	89.4	1.31
	g_{12}	86.8	1.53	$g^*_{12}(4)$	89.7	1.03	$\tilde{g}_{12}(25)$	89.2	1.31
	g_{22}	86.8	1.68	$g^*_{22}(4)$	91.7	1.22	$\tilde{g}_{22}(25)$	90.0	1.54
			$a = 21$			$b = 54$			$b = 44$
$8(IS)$	$\hat{Y}$	92.1	0.86	$\hat{Y}^*(2)$	92.7	0.84	$\hat{Y}^*(4)$	92.8	0.77
	g_{11}	92.5	0.86	$g^*_{11}(2)$	93.5	0.84	$g^*_{11}(4)$	93.1	0.77
	g_{12}	92.3	0.87	$g^*_{12}(2)$	93.2	0.85	$g^*_{12}(4)$	93.0	0.77
	g_{22}	93.6	0.75	$g^*_{22}(2)$	92.9	0.73	$g^*_{22}(4)$	93.2	0.64
			$a = 53$			$b = 77$			$b = 187$
	$\tilde{Y}(10)$	93.5	0.92	$\tilde{Y}(15)$	92.7	0.89	$\tilde{Y}(25)$	91.3	0.89
	$\tilde{g}_{11}(10)$	93.2	0.92	$\tilde{g}_{11}(15)$	93.1	0.90	$\tilde{g}_{11}(25)$	91.0	0.89
	$\tilde{g}_{12}(10)$	93.3	0.92	$\tilde{g}_{12}(15)$	93.2	0.90	$\tilde{g}_{12}(25)$	91.2	0.90
	$\tilde{g}_{22}(10)$	93.8	0.77	$\tilde{g}_{22}(15)$	93.0	0.80	$\tilde{g}_{22}(25)$	92.4	0.79
			$b = 94$			$b = 108$			$b = 133$

Table 3 (Continued)

Industry Type	Estimator	ACP	ARL	Estimator	ACP	ARL	Estimator	ACP	ARL
$9(C)$	$\hat{Y}$	91.7	1.04	$\hat{Y}^*(4)$	91.8	0.86	$\tilde{Y}(25)$	90.0	1.04
	g_{11}	91.6	1.04	$g_{11}^*(4)$	92.0	0.86	$\hat{g}_{11}(25)$	89.8	1.04
	g_{12}	91.9	1.05	$g_{12}^*(4)$	92.1	0.86	$\hat{g}_{12}(25)$	90.1	1.04
	g_{22}	91.4	0.93	$g_{22}^*(4)$	90.3	0.72	$\hat{g}_{22}(25)$	91.0	0.91
			$a = 41$			$b = 113$			$b = 82$
$10(PC)$	$\hat{Y}$	85.9	1.85	$\hat{Y}^*(1)$	88.6	1.60	$\hat{Y}^*(4)$	87.0	1.53
	g_{11}	86.0	1.85	$g_{11}^*(1)$	88.4	1.61	$g_{11}^*(4)$	87.8	1.54
	g_{12}	85.9	1.85	$g_{12}^*(1)$	88.1	1.61	$g_{12}^*(4)$	87.7	1.54
	g_{22}	86.3	1.83	$g_{22}^*(1)$	87.8	1.61	$g_{22}^*(4)$	87.9	1.52
			$a = 9$			$b = 20$			$b = 21$
	$\tilde{Y}(10)$	90.2	1.97	$\tilde{Y}(15)$	83.8	1.85	$\tilde{Y}(25)$	87.3	1.87
	$\tilde{g}_{11}(10)$	90.5	1.99	$\tilde{g}_{11}(15)$	84.3	1.86	$\tilde{g}_{11}(25)$	87.8	1.87
	$\tilde{g}_{12}(10)$	90.3	1.98	$\tilde{g}_{12}(15)$	84.2	1.86	$\tilde{g}_{12}(25)$	87.3	1.87
	$\tilde{g}_{22}(10)$	90.2	1.99	$\tilde{g}_{22}(15)$	84.2	1.84	$\tilde{g}_{22}(25)$	87.8	1.84
			$b = 13$			$b = 15$			$b = 17$
All 10	$\hat{Y}$	91.1	0.65	$\hat{Y}^*(1)$	89.5	0.65	$\hat{Y}^*(2)$	91.2	0.68
	g_{11}	91.1	0.65	$g_{11}^*(1)$	90.3	0.66	$g_{11}^*(2)$	91.8	0.68
	g_{12}	91.2	0.66	$g_{12}^*(1)$	90.6	0.66	$g_{12}^*(2)$	91.7	0.69
	g_{22}	93.5	0.78	$g_{22}^*(1)$	91.8	0.78	$g_{22}^*(2)$	91.6	0.81
			$a = 112$			$b = 241$			$b = 164$
			$c = 112$			$d = 241$			$d = 164$
	$\hat{Y}^*(3)$	90.2	0.58	$\hat{Y}^*(4)$	91.6	0.46	$\tilde{Y}(8)$	91.4	0.85
	$g_{12}^*(3)$	91.5	0.58	$g_{11}^*(4)$	91.8	0.46	$\tilde{g}_{11}(8)$	91.1	0.84
	$g_{12}^*(3)$	91.3	0.59	$g_{12}^*(4)$	91.7	0.46	$\tilde{g}_{12}(8)$	91.1	0.84
	$g_{22}^*(3)$	92.4	0.73	$g_{22}^*(4)$	93.8	0.64	$\tilde{g}_{22}(8)$	92.4	0.96
			$b = 273$			$b = 357$			$b = 167$
			$d = 273$			$d = 357$			$d = 167$
	$\tilde{Y}(10)$	94.9	0.87	$\tilde{Y}(15)$	91.8	0.74	$\tilde{Y}(25)$	92.1	0.73
	$\tilde{g}_{11}(10)$	94.9	0.88	$\tilde{g}_{11}(15)$	92.7	0.74	$\tilde{g}_{11}(25)$	92.3	0.73
	$\tilde{g}_{12}(10)$	95.1	0.88	$\tilde{g}_{12}(15)$	92.5	0.75	$\tilde{g}_{12}(25)$	92.7	0.73
	$\tilde{g}_{22}(10)$	95.9	1.01	$\tilde{g}_{22}(15)$	92.9	0.88	$\tilde{g}_{22}(25)$	93.7	0.88
			$b = 185$			$b = 220$			$b = 263$
			$d = 185$			$d = 220$			$d = 263$

2.6 Comments and Recommendations

For a good estimator, the value of ACP should be high and close to 95.0 and that of ARL should be as small as possible. Also, 'b' should be large compared to 'a' but 'd' should not exceed 'c' by too much. Applying these collective criteria, the 3 greg estimators do not seem to show appreciably better results than the original RHC-RHC estimator. This is possibly because the original estimator itself is based on an initial sample which has been chosen well enough through suitable stratification and use of appropriate size measures in both stages and so it does reasonably well. The single regressor used here in greg is incapable of yielding further accuracy in estimation by the regression technique. More improvement could have been achieved if the available auxiliary variable was better associated with the variable of interest. However, in large scale surveys, a search for such a highly associated variable is not always possible and one has to use what is readily available.

For the highly localized industries, namely, Handloom(1), Silk(5) and Stone-breaking(6), only 'adaptive' sampling achieves significant improvement. For the rest, the original sampling is good enough. Among the 4 types of industry sets used in our present work as the condition C^* for networking, the industry set Husking and Iron-smithy (3 & 8) seems to be the most suitable in improving the precision in estimation. But this entails increasing the over-all initial sample-size, say from 112 to 357 in the adaptive sample in one replicate. So, constraining the sample-size is important.

Our suggested procedures happen to bring down the sample size from 357 to 167, 185, 220 and 263 respectively by 8%, 10%, 15% and 25% subsampling, with upward rounding to integers. It is sometimes found that an estimator based on size-restricted adaptive sample has a lower ACP for a larger sample size than for a smaller sample. This is because the former estimator gives a smaller estimate of MSE and consequently a narrower CI which fails to cover the true value. Thus it may be noted that the former estimator may sometimes have smaller ACP but it performs better in terms of ARL having a smaller ARL and our findings have been appeared in Chaudhuri, Bose and Dihidar (2005).

So, our final recommendation is that first a good initial sampling scheme has to be employed utilizing available auxiliary data. This may achieve de-

sirable levels of efficiency for estimation of many of the characteristics. If it fails with respect to a few variables, as ascertained on computing the estimators for the coefficients of variation, then as alternatives greg estimators may be tried. If their estimated coefficients of variation also turn out to be large, then adaptive sampling may be tried. Since, field works for the adaptive sampling have to be implemented prior to the data analysis, we should undertake it for those variables for which one anticipates possible drops in efficiency level prior to the survey. However, if the resulting adaptive sample-size goes on spiraling up, a decision for sub-sampling has to be implemented at the field work stage. Whenever one considers going for adaptive sampling with or without size-constraining, prior data as in Tables 1 and 2 having the distribution of the variable of interest among the factors as well as the association table must be exploited for a proper guidance.

Chapter 3

A study on the feasibility of basing Horvitz & Thompson's estimator (HTE) on a sample by Rao, Hartley & Cochran's(RHC) scheme and several competitive variance estimators

Abstract

We derive formulae for the first and second order inclusion probabilities for Rao, Hartley and Cochran's (RHC, 1962) scheme of sampling. They enable us to evaluate, for a sample drawn according to the RHC scheme, the Horvitz and Thompson's (HT, 1952) estimator (HTE). We also derive expressions for several alternative unbiased variance estimators of the estimators given by Rao et al.(1962)(RHCE) and the HTE for estimating a population total when the sample is drawn by the RHC scheme. So, for a sample at hand drawn by RHC scheme, we may use either the RHCE or the HTE, on finding which one has the smaller coefficient of variation. Using the data used in Chapter 2, we compare the relative accuracies in estimation by dint of simulation, using the classical and the new variance

estimators. Our criteria for comparison are the actual coverage percentage and the average coefficient of variation. We demonstrate that as against the RHCE, the HTE is a viable contender in estimation from an RHC sample.

3.1 Introduction

We consider the problem of estimating the population total Y when some size measures for the population units are available. For this situation, the sampling scheme of Rao et al.(1962) discussed in Section 1.2, is useful and for which the estimator RHCE as given in (1.2.12) is available. However, we wanted to see if the popular estimator due to Horvitz and Thompson (1952), as given in (1.2.4), could also be used in this case. Towards this, in this Chapter, in Section 3.2, we derive the inclusion probabilities for a RHC scheme, which in turn allow us to obtain the HTE and its variance from this sample. In Section 3.3, we propose some alternative variance estimators for the RHCE and HTE, besides the classical ones given in (1.2.14), (1.2.4) and (1.2.5). Finally, in Section 3.4, we do a simulation study based on the data used in Chapter 2, to examine the accuracy in estimation by these alternative estimators. Our criteria for comparison are the ACV, ACP and ARL, as defined in Section 1.1.

In our numerical study we estimate six totals, and we see that for each of these six cases, the HTE with Yates and Grundy (1953)'s form of variance estimator has the smallest ACV among all other forms. In particular, with this form, the ACV for the HTE is even smaller than the ACV for the RHCE based on the usual variance estimator as given by Rao et al.(1962), even though the sample is drawn according to the RHC scheme in its optimal form of group formation. The ACP of the HTE is also larger than the ACP of the RHCE for all six while the ARL of HTE is smaller than that of the RHCE for all industries. So, for estimating the population total from an RHC sample, along with the usual RHCE, the HTE also emerges as a viable option.

3.2 Derivation of inclusion probabilities

We recall the inclusion probabilities defined in (1.2.1) which are required to compute the HTE given in (1.2.4). Suppose a sample of size n is drawn

by the RHC sampling scheme, described in Section 1.2. With notation as in Chapter 1, we study the following two cases separately.

Case 1. $N/n =$ **an integer.** $N/n = m$, (say).

Here $k = n$. For integers T and M, $T \geq M$, we use the following notation:

$$\alpha(T, M) = \binom{T}{M}.$$

Let G denote the total number of possible random n groups by RHC scheme. A group of m units containing unit i can be formed from the N units of U in $\alpha(N-1, m-1)$ ways. Let S_{m-1}^{N-1} denote summation over these $\alpha(N-1, m-1)$ m-tuples. Once a group containing unit i is formed, let G_1 denote the total number of possible random $(n-1)$ groups by RHC scheme. Then, clearly,

$$G = \frac{N!}{(m!)^n n!} \quad \text{and} \quad G_1 = \frac{(N-m)!}{(m!)^{n-1}(n-1)!} \tag{3.2.1}$$

Let p_{r_l} be the normed size measure of the l^{th} unit r_l in the i^{th} group consisting of m distinct units, $(r_l \neq i)$. Then, π_i can be written as

$$\pi_i = \frac{G_1}{G}[S_{m-1}^{N-1}(\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}})] \tag{3.2.2}$$

To derive formulae for π_{ij} we first note that, the number of m-tuples of distinct units of U with unit i included but unit j not included is equal to $\alpha(N-2, m-1)$. Again, the number of m-tuples $(j, s_1, .., s_{m-1})$ of distinct units of U such that $j \neq i$, $s_t \neq i \neq j \neq r_l$, $l = 1, .., m-1$, $t = 1, .., m-1$ is $\alpha(N-m-1, m-1)$. Let S_{m-1}^{N-2} and S_{m-1}^{N-m-1} respectively denote summations over these $\alpha(N-2, m-1)$ and $\alpha(N-m-1, m-1)$ m-tuples.

Let G_2 denote the number of possible ways of forming $(n-2)$ random groups of m units other than the units in (i) any random group containing unit i but excluding unit j and (ii) any random group containing unit j but not containing any unit in the random group formed as in (i). Then,

$$G_2 = \frac{(N-2m)!}{(m!)^{n-2}(n-2)!}. \tag{3.2.3}$$

Hence,

$$\pi_{ij} = \frac{G_2}{G}[S_{m-1}^{N-2}\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}}(S_{m-1}^{N-m-1}\frac{p_j}{p_j + \sum_{t=1}^{m-1} p_{s_t}})]. \tag{3.2.4}$$

We note that,

$$\frac{G_1}{G} = \frac{nm!(N-m)!}{N!}, \ \frac{G_2}{G} = \frac{(N-2m)!(m!)^2 n(n-1)}{N!}.$$

One possible check for the correctness of the above rather complicated formulae is to verify if the well-known consistency conditions namely

$$(I)\ \sum_{i=1}^{N} \pi_i = n \text{ and } (II) \sum_{j=1, j\neq i}^{N} \pi_{ij} = (n-1)\pi_i, \ \Rightarrow \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \pi_{ij} = n(n-1) \tag{3.2.5}$$

are satisfied by (3.2.2) and (3.2.4).

From (3.2.2), it follows that $\sum_{i=1}^{N} \pi_i$ contains $N \times \binom{N-1}{m-1}$ terms. We make groups of m suitable terms such that the summation of m terms in each group yields 1. The total number of such groups yielding 1 is $N \times \binom{N-1}{m-1}\frac{1}{m}$. Hence from (3.2.1) we may show that,

$$\begin{aligned} \sum_{i=1}^{N} \pi_i &= \frac{(N-m)!}{(m!)^{n-1}(n-1)!} \frac{(m!)^n n!}{N!} \binom{N-1}{m-1} \frac{N}{m} 1 \\ &= \frac{(N-m)!m!n}{N!} \frac{(N-1)!}{(m-1)!(N-m)!} \frac{N}{m} = n \end{aligned} \tag{3.2.6}$$

Thus condition (I) of (3.2.5) is satisfied.

Again, from (3.2.1), (3.2.3) and (3.2.4), using more combinatorial arguments we can show that,

$$\begin{aligned} &\sum_{j=1, j\neq i}^{N} \pi_{ij} \\ &= \frac{(N-2m)!(m!)^n n!}{(m!)^{n-2}(n-2)!N!} \left[\sum_{j=1 j\neq i}^{N} (S_{m-1}^{N-2} \frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}} (S_{m-1}^{N-m-1} \frac{p_j}{p_j + \sum_{t=1}^{m-1} p_{s_t}}))\right] \\ &= \frac{(N-2m)!(m!)^2 n(n-1)}{N!} \left[\binom{N-m-1}{m-1}(N-m)\frac{1}{m}(S_{m-1}^{N-1} \frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}})\right] \\ &= \frac{(n-1)n(m!)(N-m)!}{N!}(S_{m-1}^{N-1} \frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}}) = (n-1)\pi_i. \end{aligned} \tag{3.2.7}$$

Thus (3.2.6) and (3.2.7) together imply that $\sum \sum_{j\neq i} \pi_{ij} = n(n-1)$ and so condition (II) of (3.2.5) is satisfied.

Case 2. N/n **is not an integer.**

Here we recall that in each of the k groups, the number of units is m and the remaining $(n-k)$ groups have $(m+1)$ units in each. These two

types of sizes of groups in turn lead us to re-form here the values of G, G_1 and G_2 as defined in Case 1.

In this case, let G reduce to $G^{'}$; G_1 correspond to two terms namely $G_1^{'}$ and $G_1^{''}$ and G_2 correspond to three terms namely $G_2^{(1)}$, $G_2^{(2)}$ and $G_2^{(3)}$. Their values may be written as follows

$$G^{'} = \frac{N!}{(m!)^k k!(m+1)!^{n-k}(n-k)!} \tag{3.2.8}$$

$$G_1^{'} = \frac{(N-m)!}{(m!)^{k-1}(k-1)!(m+1)!^{n-k}(n-k)!}, \quad G_1^{''} = \frac{(N-m-1)!}{(m!)^k k!(m+1)!^{n-k-1}(n-k-1)!} \tag{3.2.9}$$

$$\begin{aligned} G_2^{(1)} &= \frac{(N-2m)!}{(m!)^{(k-2)}(k-2)!(m+1)!^{n-k}(n-k)!} \\ G_2^{(2)} &= \frac{(N-2m-1)!}{(m!)^{(k-1)}(k-1)!(m+1)!^{n-k-1}(n-k-1)!} \\ G_2^{(3)} &= \frac{(N-2m-2)!}{(m!)^k k!(m+1)!^{n-k-2}(n-k-2)!} \end{aligned} \tag{3.2.10}$$

Then, applying similar notations and arguments as in Case 1, it follows on simplification that

$$\pi_i = \frac{G_1^{'}}{G^{'}}[S_{m-1}^{N-1}(\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}})] + \frac{G_1^{''}}{G^{'}}[S_m^{N-1}(\frac{p_i}{p_i + \sum_{l=1}^{m} p_{r_l}})] \tag{3.2.11}$$

$$\pi_{ij} = A_1 + A_2 + A_3 + A_4 \tag{3.2.12}$$

where,

$$A_1 = \frac{G_2^{(1)}}{G^{'}}[S_{m-1}^{N-2}\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}}(S_{m-1}^{N-m-1}\frac{p_j}{p_j + \sum_{t=1}^{m-1} p_{s_t}})],$$

$$A_2 = \frac{G_2^{(2)}}{G^{'}}[S_{m-1}^{N-2}\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}}(S_{m}^{N-m-1}\frac{p_j}{p_j + \sum_{t=1}^{m} p_{s_t}})],$$

$$A_3 = \frac{G_2^{(2)}}{G^{'}}[S_{m}^{N-2}\frac{p_i}{p_i + \sum_{l=1}^{m} p_{r_l}}(S_{m-1}^{N-m-2}\frac{p_j}{p_j + \sum_{t=1}^{m-1} p_{s_t}})],$$

$$A_4 = \frac{G_2^{(3)}}{G^{'}}[S_{m}^{N-2}\frac{p_i}{p_i + \sum_{l=1}^{m} p_{r_l}}(S_{m}^{N-m-2}\frac{p_j}{p_j + \sum_{t=1}^{m} p_{s_t}})].$$

As in Case 1, we verify the correctness of the above expressions by checking conditions (I) and (II) of (3.2.5). From (3.2.8), (3.2.9) and (3.2.11) it follows that

$$\sum_{i=1}^{N} \pi_i = \frac{m!kN!}{N!m!} + \frac{(m+1)!(n-k)N!}{N!(m+1)!} = k + (n-k) = n$$

and so condition (I) of (3.2.5) is satisfied. Again, using arguments as in Case 1 to sum the first two terms of π_{ij} over $j = 1, .., N$ $(j \neq i)$ and then doing the same for next two terms, we get the following results:

$$\sum_{j=1, j\neq i}^{N} (A_1 + A_2) = (n-1)\frac{G_1^{'}}{G^{'}}[S_{m-1}^{N-1}(\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}})].$$

$$\sum_{j=1, j\neq i}^{N} (A_3 + A_4) = (n-1)\frac{G_1^{''}}{G^{'}}[S_{m}^{N-1}(\frac{p_i}{p_i + \sum_{l=1}^{m} p_{r_l}})].$$

Hence, $\sum_{j=1, j\neq i}^{N} \pi_{ij} = (n-1)\pi_i$ and so $\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \pi_{ij} = n(n-1)$, showing that condition (II) of (3.2.5) is satisfied.

REMARK. Since these inclusion probabilities are all positive we may employ the HTE based on the RHC scheme to unbiasedly estimate Y and any of the variance estimators of the HTE. The estimated coefficients of variation (CV) are

$$\frac{+\sqrt{\hat{V}(\hat{Y}_{RHC})}}{\left|\hat{Y}_{RHC}\right|} \text{ and } \frac{+\sqrt{\hat{V}(\hat{Y}_{HT})}}{\left|\hat{Y}_{HT}\right|}$$

.

3.3 Several alternative variance estimators of the RHC estimator and the HTE

In this section, we present some unbiased variance estimators for $\hat{Y}_{RHC}$ and $\hat{Y}_{HT}$ for estimating Y from an RHC sample. For simplicity, we derive expressions for one stage sampling which may be extended to multistage sampling.

3.3.1 Variance estimators for the RHC estimator from an RHC sample

Two alternative unbiased variance estimators for the RHCE are available in the literature, one is shown in (1.2.14) and the other is due to Ohlsson(1989) which we also mentioned in Section 1.2. We now derive some unbiased variance estimators for the RHCE using two approaches.

Approach 1: Following Rao(1979), from (1.2.12) we may write $\hat{Y}_{RHC}$ as

$$\hat{Y}_{RHC} = \sum_{i=1}^{N} y_i b_{si} I_{si}, \tag{3.3.1}$$

where

$$b_{si} = Q_i/p_i; \quad I_{si} = 1 \text{ for } i \in s \text{ and } = 0 \text{ otherwise.} \tag{3.3.2}$$

Let $y_i \propto W_i$, $i = 1, 2, \ldots, N$, where W_i's are some known non-zero constants such that $V(\sum_{i\in s} y_i \frac{Q_i}{p_i}) = 0$. A choice $W_i = p_i \ \forall \ i$ achieves this. From Rao(1979) we get the variance of the RHCE as

$$V(\hat{Y}_{RHC}) = -\sum_{i<}^{N}\sum_{j=1}^{N} d_{ij}(\frac{y_i}{p_i} - \frac{y_j}{p_j})^2 p_i p_j, \tag{3.3.3}$$

and a non-negative unbiased estimator of this variance is necessarily of the form

$$\hat{V}(\hat{Y}_{RHC}) = -\sum_{i<}^{N}\sum_{j=1}^{N} d_{sij} I_{sij}(\frac{y_i}{p_i} - \frac{y_j}{p_j})^2 p_i p_j, \tag{3.3.4}$$

where $d_{ij} = E_P(b_{si}I_{si} - 1)(b_{sj}I_{sj} - 1)$, d_{sij} is free of y_i's with $E_P(d_{sij}I_{sij}) = d_{ij}$, E_P denotes the expectation operator generically for a sampling design; $I_{sij} = I_{si}I_{sj}$.

Let $p(s)$ be the probability of getting a particular sample s. We shall try two choices for d_{sij} :

$$\text{Choice 1} : d_{sij} = d_{ij}/\pi_{ij}; \quad \text{Choice 2} : d_{sij} = d_{ij}/[p(s)\binom{N-2}{n-2}]. \tag{3.3.5}$$

The estimator based on choice 1 of (3.3.5)

For this, we need to obtain an expression for d_{ij}. Since $\hat{Y}_{RHC}$ is unbiased,

from (3.3.1) we have

$$V(\hat{Y}_{RHC}) = E_P\left[\left(\hat{Y}_{RHC} - Y\right)^2\right] = E_P\left[\left(\sum_{i=1}^{N} b_{si}I_{si}y_i - \sum_{i=1}^{N} y_i\right)^2\right]$$
$$= \sum_{i=1}^{N} y_i^2 E_P\left[(b_{si}I_{si} - 1)^2\right] + \sum_{i=1}^{N}\sum_{j(\neq i)=1}^{N} y_i y_j E_P\left[(b_{si}I_{si} - 1)\,(b_{sj}I_{sj} - 1)\right]$$
$$= \sum_{i=1}^{N} y_i^2 d_{ii} + \sum_{i=1}^{N}\sum_{j(\neq i)=1}^{N} y_i y_j d_{ij} \qquad (3.3.6)$$

Again, from (1.2.12), we may write

$$V(\hat{Y}_{RHC}) = \frac{\sum_n {N_i}^2 - N}{N(N-1)}\left[\sum_{i=1}^{N} y_i^2(\frac{1}{p_i} - 1) - \sum_{i=1}^{N}\sum_{j(\neq i)=1}^{N} y_i y_j\right]$$

and comparing this with (3.3.6), we get

$$\begin{aligned} d_{ii} &= [\textstyle\sum_n {N_i}^2 - N]/[N(N-1)](\frac{1}{p_i} - 1), \text{ and} \\ d_{ij} &= [N - \textstyle\sum_n {N_i}^2]/[N(N-1)]; \quad i \neq j \end{aligned}$$

Using this value of d_{ij} in choice 1 of (3.3.5), from (3.3.4), we get an unbiased estimator of $V(\hat{Y}_{RHC})$ as:

$$\hat{V}_{Approach1,\pi_{ij}}(\hat{Y}_{RHC}) = \frac{\sum_n {N_i}^2 - N}{N(N-1)} \sum_{i\in s}\sum_{j\in s(j>i)} \frac{p_i p_j}{\pi_{ij}}(\frac{y_i}{p_i} - \frac{y_j}{p_j})^2. \qquad (3.3.7)$$

The estimator based on choice 2 of (3.3.5)

Here, we need to obtain an expression for $p(s)$. The various groups in RHC sampling scheme may be formed in

$$\begin{aligned} G &= \tfrac{N!}{(m!)^n n!} \text{ ways if } \tfrac{N}{n} = m \text{ is an integer, and in} \\ G' &= \tfrac{N!}{(m'!)^k k!(m'+1)!^{n-k}(n-k)!} \text{ ways otherwise, with } [N/n] = m' \end{aligned}$$

Once a group is formed, the probability of getting a sample $s = (i_1, i_2, .., i_n)$ is $\frac{p_{i_1}}{Q_1}\frac{p_{i_2}}{Q_2}...\frac{p_{i_n}}{Q_n}$. Hence, for these two cases, the probability of obtaining the sample s will be given by:

$$\text{Case 1: } p(s) = \sum_G \frac{1}{G}\frac{p_{i_1}}{Q_1}\frac{p_{i_2}}{Q_2}...\frac{p_{i_n}}{Q_n}, \quad \text{Case 2: } p(s) = \sum_{G'} \frac{1}{G'}\frac{p_{i_1}}{Q_1}\frac{p_{i_2}}{Q_2}...\frac{p_{i_n}}{Q_n}.$$

where $\sum_G$ and $\sum_{G'}$ denote summation over all groups in the two cases respectively.

Using this $p(s)$ in choice 2 of (3.3.5), from (3.3.4), we get another unbiased estimator of $V(\hat{Y}_{RHC})$ as follows:

$$\hat{V}_{Approach1,p(s)}(\hat{Y}_{RHC}) = \frac{\sum_n {N_i}^2 - N}{N(N-1)} \sum_{i \in s} \sum_{j \in s(j>i)} \frac{p_i p_j (n-2)!(N-n)!}{p(s)(N-2)!} (\frac{y_i}{p_i} - \frac{y_j}{p_j})^2. \tag{3.3.8}$$

Approach 2: From (3.3.6) it follows that an unbiased estimator for $V(\hat{Y}_{RHC})$ will be of the form

$$\hat{V}(\hat{Y}_{RHC}) = \sum_{i=1}^{N} y_i^2 d_{sii} I_{sii} + \sum_{i(\neq j)=1}^{N} \sum_{j=1}^{N} y_i y_j d_{sij} I_{sij} \tag{3.3.9}$$

where d_{sii}, d_{sij} are free of y_i and $E_P(d_{sii} I_{sii}) = d_{ii}, E_P(d_{sij} I_{sij}) = d_{ij}$.

It is clear that the following two choices for d_{sii} may be used:

$$\text{Choice 3: } d_{sii} = d_{ii}/\pi_i; \quad \text{Choice 4: } d_{sii} = d_{ii}/[p(s) \sum_{i=1}^{N} I_{si}],$$

where $\sum_{i=1}^{N} I_{si} = \binom{N-1}{n-1}$; together with the two choices of d_{sij} as given in (3.3.5). Using these in (3.3.9), the following two variance estimators are obtained.

$$\hat{V}_{Approach2,\pi_{ij}}(\hat{Y}_{RHC}) = \frac{\sum_n {N_i}^2 - N}{N(N-1)} \left[\sum_{i=1}^{N} y_i^2 (\frac{1}{p_i} - 1) \frac{I_{si}}{\pi_i} - \sum_{i=1}^{N} \sum_{j(\neq i)=1}^{N} y_i y_j \frac{I_{sij}}{\pi_{ij}} \right], \tag{3.3.10}$$

$$\hat{V}_{Approach2,p(s)}(\hat{Y}_{RHC}) =$$

$$\frac{\sum_n {N_i}^2 - N}{N(N-1)} \left[\sum_{i=1}^{N} y_i^2 (\frac{1}{p_i} - 1) \frac{I_{si}(n-1)!(N-n)!}{p(s)(N-1)!} - \sum_{i=1}^{N} \sum_{j(\neq i)=1}^{N} y_i y_j \frac{I_{sij}(n-2)!(N-n)!}{p(s)(N-2)!} \right]. \tag{3.3.11}$$

3.3.2 Variance estimators for the HTE from an RHC sample

Using arguments as in the previous subsection, we can obtain alternative unbiased estimators for variance of the HTE as given in (1.2.4) and (1.2.5) by the pairs of estimators in (3.3.12), (3.3.13) and (3.3.14), (3.3.15) given below:

$$\hat{V}_{HT,\pi_{ij}}(\hat{Y}_{HT}) = \sum_{i=1}^{N} y_i^2 \frac{1-\pi_i}{\pi_i} \frac{I_{si}}{\pi_i} + \sum_{i=1}^{N} \sum_{(j\neq i)=1}^{N} y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I_{sij}}{\pi_{ij}}, \tag{3.3.12}$$

$$\hat{V}_{HT,p(s)}(\hat{Y}_{HT}) =$$

$$\sum_{i=1}^{N} y_i^2 \frac{1-\pi_i}{\pi_i} \frac{I_{si}(n-1)!(N-n)!}{p(s)(N-1)!} + \sum_{i=1}^{N} \sum_{j(\neq i)=1}^{N} y_i y_j \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} \frac{I_{sij}(n-2)!(N-n)!}{p(s)(N-2)!}. \tag{3.3.13}$$

$$\hat{V}_{YG,\pi_{ij}}(\hat{Y}_{HT}) = \sum_{i\in s} \sum_{j\in s(j>i)} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2, \tag{3.3.14}$$

$$\hat{V}_{YG,p(s)}(\hat{Y}_{HT}) = \sum_{i\in s} \sum_{j\in s(j>i)} \frac{(\pi_i\pi_j - \pi_{ij})(n-2)!(N-n)!}{p(s)(N-2)!} (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2. \tag{3.3.15}$$

3.4 Application

We use the data used in Section 2.5 to compare the variance estimators considered in Section 3.3, for estimating the total numbers of earners from 6 unorganized industries in Birbhum district in India. The industries we now consider are Bamboo(B), Husking(HU), Pottery(P), Tobacco(T), Ironsmithy(IS) and Carpentry(C). As in Section 2.5, we use a stratified two-stage sampling, but for simplicity, we now use RHC scheme at first stage and SRSWOR at second stage.

As before, we need to estimate y_i and using the expressions for SRSWOR as given in (1.2.11), we get

$$\hat{y}_i = \frac{M_i}{m_i} \sum_{m_i} y_{ij} \ , \ \bar{y}_i = \frac{1}{m_i} \sum_{m_i} y_{ij} \text{ and } \hat{V}_i = M_i^2 (\frac{1}{m_i} - \frac{1}{M_i}) (\frac{1}{m_i - 1}) \sum_{m_i} (y_{ij} - \bar{y}_i)^2,$$

where $\sum_{m_i}$ denotes summation over m_i sampled villages for the i^{th} sampled block. Then, using (1.2.12), and results from Chaudhuri, Adhikari and Dihidar (2000), unbiased estimator for Y and an unbiased estimator of its variance are given by:

$$\hat{Y}_{RHC,SRS} = \sum_{i\in s} \frac{Q_i}{p_i} (\frac{M_i}{m_i} \sum_{m_i} y_{ij}) \text{ and } \hat{V}(\hat{Y}_{RHC,SRS}) = v_{RHC}|_{y_i=\hat{y}_i} + \sum_{i\in s} \frac{Q_i}{p_i} \hat{V}_i \tag{3.4.1}$$

where $v_{RHC}|_{y_i=\hat{y}_i}$ is any one of the variance estimators for RHCE in case of unistage sampling as in (1.2.14), (3.3.7), (3.3.8), (3.3.10) and (3.3.11), computed with y_i's replaced by $\hat{y}_i$'s.

We may also compute the HTE for Y as

$$\hat{Y}_{HT,SRS} = \sum_{i \in s} \frac{1}{\pi_i}(\frac{M_i}{m_i}\sum_{m_i} y_{ij}) \text{ and } \hat{V}(\hat{Y}_{HT,SRS}) = v_{HT}|_{y_i=\hat{y}_i} + \sum_{i \in s} \frac{1}{\pi_i}\hat{V}_i \quad (3.4.2)$$

where $v_{HT}|_{y_i=\hat{y}_i}$ is any one of the variance estimators for HTE in case of unistage sampling as in (3.3.12)-(3.3.15), computed with y_i's replaced by $\hat{y}_i$'s.

Again, as in Section 2.5, we replicate the sampling and compare the efficacies of the alternative variance estimators. We use the criteria of ACP, ACV and ARL as defined in Section 1.1. In our illustration with this data, the cv's turn out large which is undesirable. But, we still endure with them because our motivation is to implement a comparative study of the RHCE with the other proposed estimators and this comparison is seemingly accomplished rather well.

Table 1 shows the ACV(%) of the RHCE using (3.4.1) where in columns 2-6, the first stage variance estimators are obtained from (1.2.14), (3.3.7), (3.3.8), (3.3.10) and (3.3.11), respectively. Table 2 shows the ACV(%) of the HTE using (3.4.2) where in columns 2-5, the first stage variance estimators are obtained from (3.3.12)–(3.3.15). Tables 1A and 2A give the ACP values corresponding to the variance estimators in Tables 1 and 2 respectively, with ARL of the confidence intervals given in parentheses. In our computations, the variance estimators came out positive for all the estimators except v_5 which was negative for 13 samples for (HU) and 16 samples for (IS) out of the 1000 samples drawn.

Table 1: $ACV(\%)$'s using alternative variance estimators based on RHCE for samples drawn by RHC scheme at first stage

Industry	RHC's variance estimator v_1	New variance estimator Approach 1		New variance estimator Approach 2	
		Choice 1 v_2	Choice 2 v_3	Choice 1 v_4	Choice 2 v_5
(B)	35.46	36.26	35.06	35.92	32.15
(HU)	23.54	23.93	23.45	23.80	21.11
(P)	46.02	46.69	45.99	46.53	43.77
(T)	45.88	46.58	45.66	46.41	43.71
(IS)	23.33	23.80	23.01	23.49	19.98
(C)	29.20	29.80	28.91	29.57	26.40

Table 2: $ACV(\%)$'s using alternative variance estimators based on HTE for samples drawn by RHC scheme at first stage

Industry	New estimator for HT's variance		New estimator for YG's variance	
	Choice 1 v_6	Choice 2 v_7	Choice 1 v_8	Choice 2 v_9
(B)	33.94	32.26	29.29	28.43
(HU)	25.78	24.83	19.57	19.22
(P)	43.01	42.06	40.05	39.51
(T)	44.65	43.54	42.07	41.47
(IS)	24.63	23.01	17.35	16.90
(C)	29.45	28.05	24.14	23.53

Table 1A: $ACP(\%)$ and (ARL) using alternative variance estimators based on RHCE for samples drawn by RHC scheme at first stage

Industry	RHC's variance estimator v_1	New variance estimator Approach 1		New variance estimator Approach 2	
		Choice 1 v_2	Choice 2 v_3	Choice 1 v_4	Choice 2 v_5
(B)	88.8	88.2	87.4	89.0	83.5
	(1.39)	(1.42)	(1.37)	(1.41)	(1.26)
(HU)	88.5	89.0	87.1	88.9	81.3
	(0.93)	(0.94)	(0.92)	(0.93)	(0.83)
(P)	85.3	85.7	85.7	85.4	84.1
	(1.80)	(1.83)	(1.80)	(1.82)	(1.72)
(T)	75.2	75.5	74.8	75.5	72.2
	(1.80)	(1.83)	(1.79)	(1.82)	(1.71)
(IS)	92.9	92.5	92.2	92.6	84.4
	(0.91)	(0.93)	(0.90)	(0.92)	(0.79)
(C)	87.0	87.6	86.5	87.6	82.0
	(1.14)	(1.17)	(1.13)	(1.16)	(1.03)

Table 2A: $ACP(\%)$ and (ARL) using alternative variance estimators based on HTE for samples drawn by RHC scheme at first stage

Industry	New estimator for HT's variance		New estimator for YG's variance	
	Choice 1 v_6	Choice 2 v_7	Choice 1 v_8	Choice 2 v_9
(B)	92.2	89.9	87.8	86.5
	(1.33)	(1.26)	(1.15)	(1.11)
(HU)	93.9	91.9	86.3	85.9
	(1.01)	(0.97)	(0.77)	(0.76)
(P)	86.3	85.5	84.2	83.9
	(1.69)	(1.65)	(1.57)	(1.55)
(T)	76.3	75.9	75.3	74.6
	(1.75)	(1.71)	(1.65)	(1.63)
(IS)	97.4	96.1	91.7	91.0
	(0.97)	(0.90)	(0.68)	(0.66)
(C)	91.8	88.3	86.6	85.3
	(1.15)	(1.10)	(0.95)	(0.92)

Even though the ACV values are rather large for this data, as remarked earlier, our main objective is to compare the RHCE with other estimators. Tables 1 and 2 show that among the alternative variance estimators, v_9 has the smallest ACV across all 6 industries and thus, based on the criterion of ACV, the HTE with the YG form of variance estimator performs better than the RHCE.

Tables 1A and 2A show that v_6 has larger ACP than others for all the 6 industries. In terms of ARL, v_9 is the best though it is not as impressive as v_6 in terms of the ACP criterion. Thus, using the ACP and ARL for comparison, the HTE again performs better than the RHCE for all the industries considered. So, even though estimation is done based on a sample drawn according to the RHC scheme, the HTE is a good competitor of the RHCE and may be put to practice, with a preference for it over RHCE if the estimated cv of the former turns out less than that for the latter. With the live data as studied here based on RHC samples, the HTE with one of its variance estimators seems to be a better combination than the RHCE with one of its variance estimators.

3.5 Concluding Remarks

When a sample is drawn by the RHC scheme, the only estimator which could be employed so far in practice was RHCE. But on deriving the requisite formulae we find the HTE as a feasible alternative. Our recommendation is to employ, for a realized sample, the RHC estimate or the HT estimate preferring the one with the smaller CV, provided the YG estimate turns out non-negative. With several numerical examples in uni-stage sampling varying N, n, p_i's we observed '$\pi_i\pi_j > \pi_{ij}$' in all our cases ensuring positivity of the YG estimate of the variance of the HTE.

Chapter 4

Model-cum-design based estimation of the prevalence rate of a disease in a locality

Abstract

In this chapter, we study the problem of estimating the prevalence rate of a disease in a geographical area, based on data collected from a sample of locations within this area. If there are several locations with zero incidence of the disease, the usual estimators are not suitable and so we develop a new estimator, together with an unbiased estimator of its variance, which may be appropriately used in such situations. An application of this estimator is illustrated with data from a large-scale survey which was carried out in the city of Kolkata, India, to estimate the prevalence rate of stroke.

We show that spatial modeling may be used to smoothen the observed data before applying our proposed estimator. Our computations show that this smoothing helps to reduce the coefficient of variation and such a model-cum-design based procedure is useful for estimating the prevalence rate. This method may of course be used in other similar situations.

4.1 Introduction

In this chapter, we consider the problem of estimating the prevalence rate of a certain disease in a specified geographical area. We first consider this

problem in its general form and then focus on the survey which was conducted in Kolkata Municipal area to estimate the prevalence rate of stroke.

We consider an area which is divided into a number of subareas or locations and the prevalence rate of a certain disease for the entire area is to be estimated using a sample survey. Suppose a two-stage sample is believed appropriate where a sample of individuals is chosen from a first stage sample of locations within this area. Relevant data are collected for the sampled individuals. However, it is often found that several sampled locations have no individuals with the disease and in such cases the standard estimators are not appropriate for estimating the over-all prevalence rate for the entire area from the location-specific estimated rates. So, we develop a new estimator based on the Hartley-Ross estimator for use in such situations. We also obtain the unbiased variance estimator for the proposed estimator so that the cv of the estimator may be computed in practice. We illustrate our method with data collected in a survey in Kolkata Municipal area.

We next tried to investigate if our earlier estimators could be improved through the use of suitable auxiliary variables. However, in our study we found that in the micro level within the Municipal area under study, (i.e., the Municipal blocks in our context) there were no recorded data on auxiliary variables. This hurdle made it difficult to use the usual model based approach where information on suitable covariates is effectively exploited to obtain the estimate, for example, as in generalized regression (greg) estimators in small area estimation (cf. Särndal, Swenson and Wretman (1992)), which could be utilized with our data studied in Chapter 2. So, in this study, we tried instead other ways of first smoothing the region-based data. We found that spatial smoothing was effective and useful as it required no additional data other than a regional map. We also noted that similar spatial smoothing has been used in the literature for estimating disease rates, for instance, a well known example being the study on sudden infant death syndrome studied by Cressie and Chan (1989). So, following their approach, we first do a spatial analysis of the data from the different localities and then, if spatial dependence is found in the data, a spatial model is fitted to the location-specific prevalence rates. For this, variance stabilizing transformations are used followed by suitable modeling of these location-specific rates. Various spatial models for the location-specific rates are tried out and a suitable one is chosen. In addition, if available, other well correlated

covariates may be incorporated in the model. In the next step, using these smoothed rates and keeping in mind the two-stage sampling design used, a new estimator and an unbiased estimator for its variance, are derived for estimating the overall prevalence rate for the entire area. The estimator is similar to the Hartley-Ross estimator but with some modifications. Using these, the prevalence rate for the entire area and its standard error can be computed. Thus, our final estimator for the prevalence rate is a model-cum-design based estimator.

In Section 4.2, the sampling method is described and the new estimator is derived. An unbiased estimator of the variance of the new estimator is given in Section 4.3. We illustrate our proposed method by applying it to a survey which was undertaken in Kolkata, India to estimate the prevalence rate of stroke. Data from this survey is first smoothed by spatial modeling and then this was followed by the use of the newly developed estimator. This application and resulting computations are described in Section 4.4. Some concluding remarks are given in Section 4.5.

4.2 Sampling Methodology

4.2.1 Preliminaries and Sampling method

We consider a geographical area subdivided into N locations , e.g., wards in a county, labeled $1, 2, \ldots, N$. Let $R_i (= Y_i/X_i)$ denote the prevalence rate of the disease in location i, where Y_i is the (unknown) number of individuals with the disease and X_i is the population in location i, $i = 1, 2 \ldots, N$. Let R be the prevalence rate for the entire area, expressed as the ratio of the unknown number of individuals with disease in the entire area per thousand of the area population, i.e.,

$$R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} \times 1000 = \frac{Y}{X} \times 1000 \ , \ \text{ say.}$$

X_i values are readily available from Census figures. Our objective is to estimate R.

Suppose a two-stage sample is deemed appropriate from this area and simple random sampling without replacement is to be employed in both stages with locations as the first stage units and individuals in a location being the second stage units. A random sample (say, s) of n locations is

first drawn from the N locations. Then, for the i^{th} sampled location, a random sample of x_i' individuals is selected and among them, those with the specified disease are identified. Let this number be y_i'. Then the disease rate per thousand for i^{th} location is estimated by

$$1000 \times \frac{y_i'}{x_i'} = 1000 \times r_i, \quad \text{say,} \quad \text{with} \quad r_i = \frac{y_i'}{x_i'}.$$

In surveys, it may often be found that several of the y_i' values are zero, even though the corresponding x_i' values are not all equal, e.g., in our study data we find that nine wards have $y_i' = 0$, but their x_i' values ranged from 250 to as high as 779. It seems inappropriate to view the 2 cases of zero incidence of disease among 250 individuals and also among 779 individuals as equivalent. So, we follow Cressie (1993) who in a study for sudden infant death syndrome proposed a transformed rate with 1 added to the numerator. Thus, it seems meaningful to estimate R_i by the transformed disease rate per thousand for location i as :

$$1000 \times r_i^* = 1000 \times \frac{y_i' + 1}{x_i'}, \text{ where } r_i^* = \frac{y_i' + 1}{x_i'}, \quad i \in s \tag{4.2.1}$$

instead of by $1000 \times r_i = 1000 \times (y_i'/x_i')$ in order to facilitate discrimination between the locations with zero diseases but among different numbers of individuals.

However, our situation is different from that of Cressie (1993) as our objective is to finally estimate R and we are dealing with sample data and this extra $+1$ in the numerator creates complications while combining these r_i^*'s to estimate R.

If one wants to estimate R from these r_i^*'s, then none of the traditional estimators can be used to get an appropriate estimator of R. So, to estimate R we need to develop a new unbiased ratio-type estimator of Y for a two stage sampling design. We develop such an estimator in the spirit of the well known exactly unbiased Hartley-Ross estimator. For details on the Hartley-Ross estimator for one stage design we refer to Hartley and Ross (1954), Cochran (1977) and Chaudhuri and Stenger (2005). Our proposed estimator is also shown to be exactly unbiased and it may be used in other survey situations where the objective is to estimate a population total or a population ratio based on a two-stage sample.

4.2.2 A new unbiased estimator

Let E_j and V_j be the expectation and variance operators for the sampling at the jth stage, $j = 1, 2$ and E and V denote the overall expectation and variance operators. Let

$$Y = \sum_{i=1}^{N} Y_i,\ \bar{Y} = \frac{Y}{N},\quad X = \sum_{i=1}^{N} X_i,\ \bar{X} = \frac{X}{N},\ R_i = \frac{Y_i}{X_i},\ \bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i.$$

For a sample s, we recall that $r_i = y_i'/x_i'$, $r_i^* = \frac{y_i'+1}{x_i'}$, $i \in s$ and

$$\bar{y} = \frac{1}{n}\sum_{i\in s} Y_i,\quad \bar{y'} = \frac{1}{n}\sum_{i\in s} y_i',\quad \bar{x} = \frac{1}{n}\sum_{i\in s} X_i,\quad \bar{x'} = \frac{1}{n}\sum_{i\in s} x_i',\quad \bar{r^*} = \frac{1}{n}\sum_{i\in s} r_i^*.$$

For our sampling method described above, we may assume that y_i' has a Hypergeometric distribution with mean $x_i'R_i$. Clearly,

$$E_2(r_i) = R_i \text{ and } V_2(r_i) = \frac{R_i(1-R_i)}{x_i'}\left(\frac{X_i - x_i'}{X_i - 1}\right)$$

and so, from (4.2.1), it follows that

$$E_2(r_i^*) = R_i + \frac{1}{x_i'} = \frac{Y_i}{X_i} + \frac{1}{x_i'},\ i \in s;\quad E(\bar{r^*}) = E_1E_2(\bar{r^*}) = \bar{R} + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{x_i'}. \tag{4.2.2}$$

The following theorem gives an unbiased estimator for Y based on the r_i^* values.

Theorem 4.2.1 An unbiased estimator for Y is given by

$$\hat{Y}_{new} = X\bar{r^*} - \frac{N}{n}\sum_{i\in s}\frac{X_i}{x_i'} + \frac{N-1}{n-1}\left(\sum_{i\in s} r_i^* X_i - n\bar{r^*}\bar{x}\right).$$

Proof Consider $\hat{\bar{Y}} = \bar{r^*}\bar{X}$ as an estimator for $\bar{Y}$. This is a biased estimator and using (4.2.2), its bias can be shown to be:

$$E(\hat{\bar{Y}}) - \bar{Y} = E(\bar{r^*}\bar{X}) - \bar{Y} = \frac{\bar{X}}{N}\sum_{i=1}^{N}\frac{1}{x_i'} + \bar{X}\bar{R} - \bar{Y}. \tag{4.2.3}$$

Let us define

$$c = \frac{N-1}{N}\frac{1}{n-1}\sum_{i\in s}(r_i^* - \bar{r^*})(X_i - \bar{x}). \tag{4.2.4}$$

On using (4.2.2), it follows that

$$E_2(c)=\frac{N-1}{N}\frac{n}{n-1}\left\{\frac{1}{n}\sum_{i\in s}X_iE_2(r_i^*)-\bar{x}E_2(\bar{r^*})\right\}=\frac{N-1}{N}\frac{n}{n-1}\left\{\bar{y}+\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}-\bar{x}\bar{u}\right\} \tag{4.2.5}$$

on writing $\bar{u}=\sum_{i\in s}U_i/n$ with $U_i=Y_i/X_i+1/x_i'$. Let $\bar{U}=\sum_{i=1}^{N}U_i/N$. Since $U_iX_i=Y_i+X_i/x_i'$ and $\bar{U}=\bar{R}+\frac{1}{N}\sum_{i=1}^{N}1/x_i'$, it follows from (4.2.5) and Cochran (1977, p. 25) that

$$\begin{aligned}
E(c) &= E_1E_2(c)=\tfrac{N-1}{N}\tfrac{n}{n-1}\left\{\bar{Y}+\tfrac{1}{N}\sum_{i=1}^{N}\frac{X_i}{x_i'}-E_1(\bar{x}\bar{u})\right\}\\
&= \tfrac{N-1}{N}\tfrac{n}{n-1}\left\{\bar{Y}+\tfrac{1}{N}\sum_{i=1}^{N}\frac{X_i}{x_i'}-\bar{X}\bar{U}-\frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})(U_i-\bar{U})\right\}\\
&= \tfrac{N-1}{N}\tfrac{n}{n-1}\left\{\bar{Y}+\tfrac{1}{N}\sum_{i=1}^{N}\frac{X_i}{x_i'}-\bar{X}\bar{U}-\frac{N-n}{n}\frac{1}{N-1}(\bar{Y}+\frac{1}{N}\sum_{i=1}^{N}\frac{X_i}{x_i'}-\bar{X}\bar{U})\right\}\\
&= \tfrac{N-1}{N}\tfrac{n}{n-1}\left\{\left(1-\tfrac{N-n}{n}\tfrac{1}{N-1}\right)\left(\bar{Y}+\tfrac{1}{N}\sum_{i=1}^{N}\frac{X_i}{x_i'}-\bar{X}\bar{U}\right)\right\}\\
&= \bar{Y}+\tfrac{1}{N}\sum_{i=1}^{N}\frac{X_i}{x_i'}-\bar{X}\bar{U}\\
&= E(\tfrac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'})-\left\{E(\bar{r^*}\bar{X})-\bar{Y}\right\},
\end{aligned}$$

on using (4.2.3). This implies that

$$E(c-\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}+\bar{r^*}\bar{X})=\bar{Y}$$

Hence, an unbiased estimator for $\bar{Y}$ is given by

$$\hat{\bar{Y}}_{new}=\bar{X}\bar{r^*}-\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}+\frac{N-1}{N}\frac{1}{n-1}\sum_{i\in s}(r_i^*-\bar{r^*})(X_i-\bar{x}).$$

So, an estimator of Y will be

$$\hat{Y}_{new}=N\hat{\bar{Y}}_{new}=N\left[\bar{X}\bar{r^*}-\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}+\frac{N-1}{N}\frac{1}{n-1}\sum_{i\in s}(r_i^*-\bar{r^*})(X_i-\bar{x})\right].$$

Clearly this estimator is also exactly unbiased for Y and thus the theorem follows.

The form of $V(\hat{Y}_{new})$, the variance of $\hat{Y}_{new}$, is given in (4.3.1). An expression for $\hat{V}(\hat{Y}_{new})$, an unbiased estimator of $V(\hat{Y}_{new})$, is derived in the Section 4.3. Using these expressions for $\hat{Y}_{new}$ and $\hat{V}(\hat{Y}_{new})$, the following is immediate.

Corollary 4.2.1 An estimator for R (per thousand) and an unbiased estimator for its variance, are given by

$$\hat{R}_{new} = \frac{\hat{Y}_{new}}{X} \times 1000, \quad \hat{V}(\hat{R}_{new}) = (\hat{V}(\hat{Y}_{new})/X^2) \times 1000^2.$$

4.3 Unbiased variance estimator of the new estimator

Let $t_i' = (y_i' + 1)(X_i)/x_i'$. Then, on simplification using (4.2.1), $\hat{Y}_{new}$ as given in Theorem 4.2.1 may be written as $\hat{Y}_{new} = \sum_{i\in s} t_i' b_{si} - \frac{N}{n}\sum_{i\in s}\frac{X_i}{x_i'}$ where $b_{si} = X/nX_i + (N-1)/(n-1) - \bar{x}(N-1)/X_i(n-1)$. Hence,

$$V(\hat{Y}_{new}) = V(\sum_{i\in s} t_i' b_{si}) + V(\frac{N}{n}\sum_{i\in s}\frac{X_i}{x_i'}) - 2\mathrm{Cov}[\sum_{i\in s} t_i' b_{si}, \frac{N}{n}\sum_{i\in s}\frac{X_i}{x_i'}] = F + G - 2H, \text{ say.} \tag{4.3.1}$$

To obtain an unbiased estimator of this variance, we obtain unbiased estimators of F, G and H separately as shown below:

4.3.1 Estimating G :

It is easy to see that G may be estimated by

$$\hat{G} = N^2(\frac{1}{n} - \frac{1}{N})\frac{1}{n-1}\sum_{i\in s}(\frac{X_i}{x_i'} - \frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'})^2.$$

4.3.2 Estimating H :

On simplification, H reduces to

$$\begin{aligned} H &= XN\ \mathrm{Cov}\left\{\tfrac{1}{n}\sum_{i\in s}\frac{t_i'}{X_i}, \frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}\right\} + \tfrac{N-1}{n-1}Nn\ \mathrm{Cov}\left\{\tfrac{1}{n}\sum_{i\in s} t_i', \frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}\right\} \\ &- \tfrac{N-1}{n-1}Nn\ \mathrm{Cov}\left\{\bar{x}\tfrac{1}{n}\sum_{i\in s}\frac{t_i'}{X_i}, \frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}\right\}. \end{aligned}$$

By separately estimating each of the covariance terms above and writing $T_i = \sum_{j\neq i, j\in s}\frac{t_j'}{X_j}$, it can be shown that an estimator for H is given by

$$
\begin{aligned}
\hat{H} &= XN(\tfrac{1}{n}-\tfrac{1}{N})\tfrac{1}{n-1}\sum_{i\in s}\left(\frac{t_i'}{X_i}-\frac{1}{n}\sum_{i\in s}\frac{t_i'}{X_i}\right)\left(\frac{X_i}{x_i'}-\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}\right)\\
&\quad+\tfrac{N-1}{n-1}Nn(\tfrac{1}{n}-\tfrac{1}{N})\tfrac{1}{n-1}\sum_{i\in s}\left(t_i'-\frac{1}{n}\sum_{i\in s}t_i'\right)\left(\frac{X_i}{x_i'}-\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}\right)\\
&\quad-\tfrac{N-1}{n-1}N(\tfrac{1}{n}-\tfrac{1}{N})\tfrac{1}{n-1}\left[\sum_{i\in s}\left(t_i'-\frac{1}{n}\sum_{i\in s}t_i'\right)\left(\frac{X_i}{x_i'}-\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}\right)\right.\\
&\quad\left.+\sum_{i\in s}\left(X_iT_i-\frac{1}{n}\sum_{i\in s}X_iT_i\right)\left(\frac{X_i}{x_i'}-\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}\right)\right]\\
&= XN(\tfrac{1}{n}-\tfrac{1}{N})\tfrac{1}{n-1}\sum_{i\in s}\left(\frac{t_i'}{X_i}-\frac{1}{n}\sum_{i\in s}\frac{t_i'}{X_i}\right)\left(\frac{X_i}{x_i'}-\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}\right)\\
&\quad+(N-1)N(\tfrac{1}{n}-\tfrac{1}{N})\tfrac{1}{n-1}\sum_{i\in s}\left(t_i'-\frac{1}{n}\sum_{i\in s}t_i'\right)\left(\frac{X_i}{x_i'}-\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}\right)\\
&\quad-\tfrac{N-1}{n-1}N(\tfrac{1}{n}-\tfrac{1}{N})\tfrac{1}{n-1}\sum_{i\in s}\left(X_iT_i-\frac{1}{n}\sum_{i\in s}X_iT_i\right)\left(\frac{X_i}{x_i'}-\frac{1}{n}\sum_{i\in s}\frac{X_i}{x_i'}\right).
\end{aligned}
$$

4.3.3 Estimating F :

We note that

$$\hat{F}=\hat{V}(\sum_{i\in s}t_i'b_{si}) \tag{4.3.2}$$

where $t_i'=(y_i'+1)(X_i)/x_i'$ and $\sum_{i\in s}t_i'b_{si}$ is the usual Hartley-Ross estimator written as in homogeneous linear unbiased estimator form with Y_i's replaced by t_i''s.

Using the variance estimation method as in Raj (1968) for multistage designs, we have

$$\hat{V}(\sum_{i\in s}t_i'b_{si})=v_1(\sum_{i\in s}Y_ib_{si})|_{Y_i=t_i'}+\sum_{i\in s}b_{si}v_2(t_i') \tag{4.3.3}$$

where the first term on R.H.S of (4.3.3) denotes the unbiased estimator of $V_1(\sum_{i\in s}Y_ib_{si})$ evaluated at $Y_i=t_i'$ and $v_2(t_i')$ is the unbiased estimator of $V_2(t_i')$.

To derive $v_1(\sum_{i\in s}Y_ib_{si})$ we proceed as in Rao (1979). Towards this, we write $Y_i=aW_i,\ i=1,2,\ldots N$, where a is an arbitrary constant and W_i's are some known non-zero constants such that $MSE(\sum_{i\in s}Y_ib_{si})=0$. Since $\sum_{i\in s}Y_ib_{si}=\sum_{i\in s}aW_ib_{si}$, it is clear that on choosing $W_i=X_i\ \forall i$,

$$\sum_{i\in s}Y_ib_{si}=a(X+\frac{N-1}{n-1}n\bar{x}-\frac{N-1}{n-1}n\bar{x})=aX=Y,$$

and consequently, $MSE(\sum_{i\in s} Y_i b_{si}) = 0$.

Noting that $\sum_{i\in s} Y_i b_{si}$ is unbiased, on simplification after applying Theorem 1 of Rao (1979), it follows that (a) the variance of the H-R estimator will be given by

$$V_1(\sum_{i\in s} Y_i b_{si}) = -\sum_{i(<j)=1}^{N}\sum_{j=1}^{N} d_{ij}(\frac{Y_i}{X_i} - \frac{Y_j}{X_j})^2 X_i X_j,$$

where $d_{ij} = E_P(b_{si}-1)(b_{sj}-1)$, E_P denoting the expectation operator with respect to the sampling design; and (b) a non-negative quadratic unbiased estimator of this variance is necessarily of the form

$$v_1(\sum_{i\in s} Y_i b_{si}) = -\sum_{i(<j)=1}^{N}\sum_{j=1}^{N} \frac{d_{ij} I_{si} I_{sj}}{\pi_{ij}}(\frac{Y_i}{X_i} - \frac{Y_j}{X_j})^2 X_i X_j$$

$$= -\frac{N(N-1)}{n(n-1)}\sum_{i(<j)=1}^{N}\sum_{j=1}^{N} d_{ij} I_{si} I_{sj}(\frac{Y_i}{X_i} - \frac{Y_j}{X_j})^2 X_i X_j, \tag{4.3.4}$$

where $I_{si} = 1$, if unit $i \in s$, else 0; $\pi_{ij} = n(n-1)/N(N-1)$ being the inclusion probability for both units i and j in sample s for SRSWOR.

As $\sum_{i\in s} y_i b_{si}$ is unbiased for Y, we have $\sum_{s\ni i} p(s) b_{si} = 1$ for all i, where $p(s)$ is the probability of sample s being drawn. Hence, $d_{ij} = \sum_{s\ni i,j} b_{si} b_{sj} p(s) - 1$, which simplifies to $d_{ij} = (n!(N-n)!/N!)\sum_{s\ni i,j} b_{si} b_{sj} - 1$ since $p(s) = n!(N-n)!/N!$ for SRSWOR.

Now, using a little algebra, d_{ij} can be written as

$$d_{ij} = a_{ij} - b_{ij} + c_{ij}, \text{ where} \tag{4.3.5}$$

$$a_{ij} = \frac{n-1}{N(N-1)}\frac{X^2}{nX_iX_j} + \frac{X}{NX_i} + \frac{X}{NX_j} + \frac{n(N-1)}{N(n-1)} - 1, \tag{4.3.6}$$

$$b_{ij} = \frac{n!(N-n)!}{N!}(\frac{N-1}{n-1})\left\{\frac{2X}{nX_iX_j} + \frac{N-1}{n-1}\frac{1}{X_i} + \frac{N-1}{n-1}\frac{1}{X_j}\right\}\sum_{s\ni i,j}\bar{x}, \tag{4.3.7}$$

$$c_{ij} = \frac{n!(N-n)!}{N!}\frac{1}{X_iX_j}(\frac{N-1}{n-1})^2\sum_{s\ni i,j}\bar{x}^2. \tag{4.3.8}$$

From (4.3.6), it is clear that, a_{ij} can be computed easily for any sample containing both the ith and jth units. But to simplify further the terms b_{ij} and c_{ij} we need to calculate $\sum_{s\ni i,j}\bar{x}$ and $\sum_{s\ni i,j}\bar{x}^2$ which can be computed only if all the X_i values are available.

Let $\bar{x}_{n-2}(i,j)$ and $\bar{X}_{N-2}(i,j)$ be the sample and population averages of all the X_i values except those for units i and j. We recall that for any sample $s \ni i,j$, $\bar{x} = \{X_i + X_j + (n-2)\bar{x}_{n-2}(i,j)\}/n$ and for such an s, $p(s) = (n-2)!(N-n)!/(N-2)!$. So, $E_P(\bar{x}_{n-2}(i,j)) = \sum_{s\ni i,j}(n-2)!(N-n)!\bar{x}_{n-2}(i,j)/(N-2)! = \bar{X}_{N-2}(i,j)$. Utilizing this fact, from (4.3.7) after simplification we have,

$$b_{ij} = \left\{\frac{2X}{nX_iX_j} + (\frac{N-1}{n-1})\frac{1}{X_i} + (\frac{N-1}{n-1})\frac{1}{X_j}\right\}\left\{\frac{(X_i+X_j)(N-n)+X(n-2)}{N(N-2)}\right\}.$$

To simplify the c_{ij} as in (4.3.8), we use the fact that

$$\sum_{s\ni i,j}\frac{(n-2)!(N-n)!}{(N-2)!}\bar{x}^2_{n-2}(i,j) = E_P(\bar{x}^2_{n-2}(i,j)) = V_P(\bar{x}_{n-2}(i,j)) + (E_P(\bar{x}_{n-2}(i,j)))^2.$$

On simplification, we get

$$c_{ij} = \frac{N-1}{n-1}\frac{1}{X_iX_j}\frac{1}{Nn}\left\{(X_i+X_j)^2 + 2(n-2)(X_i+X_j)\bar{X}_{N-2}(i,j)\right\}$$

$$+\frac{(N-1)(n-2)^2}{(n-1)Nn(X_iX_j)}\left\{\frac{N-n}{(n-2)(N-2)(N-3)}\sum_{k(\neq i,j)=1}^{N}(X_k-\bar{X}_{N-2}(i,j))^2 + \bar{X}^2_{N-2}(i,j)\right\}.$$

Substituting these values in (4.3.5) to get d_{ij}, we get the final expression for $v_1(\sum_{i\in s}Y_ib_{si})$ from (4.3.4).

Next, to obtain the expression for $v_2(t_i')$ in (4.3.2), we note from the result obtained in Subsection (4.2.2) that

$$V_2(t_i') = V_2((y_i'+1)(X_i)/x_i') = X_i^2V_2(\frac{y_i'}{x_i'}) = X_i^2V_2(r_i) = \frac{X_i^2}{x_i'}(R_i - R_i^2)\left(\frac{X_i - x_i'}{X_i - 1}\right). \tag{4.3.9}$$

Since $E_2(r_i^*) = R_i + \frac{1}{x_i'}$, it follows that

$$\hat{R}_i = r_i^* - \frac{1}{x_i'}.$$

Also, from $V_2(r_i^*) = \frac{1}{x_i'}(R_i - R_i^2)\left(\frac{X_i-x_i'}{X_i-1}\right)$, we have

$$\begin{aligned}
E_2((r_i^*)^2) &= \tfrac{1}{x_i'}(R_i - R_i^2)\left(\tfrac{X_i-x_i'}{X_i-1}\right) + (E_2(r_i^*))^2 \\
&= \tfrac{1}{x_i'}(R_i - R_i^2)\left(\tfrac{X_i-x_i'}{X_i-1}\right) + (R_i + \tfrac{1}{x_i'})^2 \\
&= \tfrac{1}{x_i'}(R_i - R_i^2)\left(\tfrac{X_i-x_i'}{X_i-1}\right) + R_i^2 + (\tfrac{1}{x_i'})^2 + \tfrac{2R_i}{x_i'} \\
&= R_i^2\left[1 - \tfrac{1}{x_i'}(\tfrac{X_i-x_i'}{X_i-1})\right] + \tfrac{R_i}{x_i'}\left[2 + (\tfrac{X_i-x_i'}{X_i-1})\right] + \tfrac{1}{x_i'^2} \\
&= R_i^2\left[1 - \tfrac{1}{x_i'}(\tfrac{X_i-x_i'}{X_i-1})\right] + \tfrac{1}{x_i'}\left[2 + (\tfrac{X_i-x_i'}{X_i-1})\right]E_2\left(r_i^* - \tfrac{1}{x_i'}\right) + \tfrac{1}{x_i'^2}.
\end{aligned}$$

This, after little simplification yields that

$$\widehat{R_i^2} = \frac{r_i^{*2} x_i'^2 (X_i - 1) - r_i^* x_i' (3X_i - 2 - x_i') + (2X_i - 1 - x_i')}{(x_i' - 1) x_i' X_i}.$$

Using these values in (4.3.9), we obtain $v_2(t_i') = \frac{X_i^2}{x_i'}(\hat{R}_i - \widehat{R_i^2}) \left(\frac{X_i - x_i'}{X_i - 1}\right)$. Finally, from (4.3.2) and (4.3.3), we obtain $\hat{F}$.

Using the estimates of F, G and H as shown above, one may unbiasedly estimate $V(\hat{Y}_{new})$ by $\hat{V}(\hat{Y}_{new}) = \hat{F} + \hat{G} - 2\hat{H}$.

4.4 Application

4.4.1 Survey Description

Kolkata is one of the largest metropolitan cities in India and it was chosen as the site for a study on the prevalence of four major neurological disorders including stroke. A large-scale sample survey was carried out to collect data on these diseases. The Kolkata Municipal Corporation (KMC) area consists of 141 wards and from these wards, 89 wards were first chosen by simple random sampling without replacement. From each sampled ward, a number of individuals were again chosen by simple random sampling without replacement. A total of 52,377 individuals were studied from these 89 wards and using some clearly specified definitions and methods of evaluation, the persons with stroke and three other neurological diseases among them were identified. We illustrate our proposed method using the data on stroke from this survey.

Cressie (1993) discusses in detail a spatial analysis of data on sudden infant death syndrome (SIDS) and the first part of our analysis is similar to that in some respects. The major difference of our study from that in Cressie (1993) is that for us, observations are available from a two-stage sample of the locations and individuals whereas, in Cressie (1993) complete data was available and so sampling was not needed. Moreover, our objective is different from theirs as our goal is to estimate the overall prevalence rate for the entire area.

4.4.2 Spatial smoothing

In the notation of Section 4.2, for our survey, $N = 141$ and $n = 89$. It was found that there are 9 wards with $y_i^{'} = 0$ with the $x_i^{'}$ values for these wards ranging from 250 to 779. So, using the rates r_i^* as in (4.2.1) seems justified. These $1000 \times r_i^*$'s values ranged from 1.3 to 20.0 and were found to be highly skewed with the variability in the r_i^* being less for locations with larger $x_i^{'}$ values. Hence, some transformation of r_i^* values was needed to remove possible dependence of the variance on the x_i''s and to achieve stability of variance. For this, following Cressie (1993), we tried the square-root transformation and some other transformations as suggested therein: e.g.

1. $U_i = \sqrt{(1000 \times r_i^*)}$, 2. $V_i = \sqrt{(1000 \times r_i^*)} + \sqrt{(1000 \times r_i)}$,
3. $W_i = \sin^{-1}(1000 \times r_i^*)$, 4. $Y_i = \sinh^{-1}(1000 \times r_i^*)$,
5. $Z_i = \log(1000 \times r_i^*)$.

The values of x_i' in the study varied from 175 to 1920. In order to compare the stability of the variance associated with the above exploratory transformations, it was necessary to eliminate the effect of the x_i''s from the variance formula of r_i^*. This was accomplished by following Cressie and Read (1989) and first partitioning the set of 89 wards into 5 similarly sized subsets with roughly equal $(\frac{1}{x_i'})$ values. For each group, medians and inter-quartile ranges (IQ) of different transformations were available. Plots of $(\text{IQ})^2/\text{ave}(1/x_i')$ versus median for the different transformations indicated that by using the transformation $Z_i = \log(1000 \times r_i^*)$, the variance was showing to be no longer a function of mean. The unequal number of x_i''s made it impossible to assume homoscedasticity. So, again as in Cressie (1989), for better result, we tried the transformation $(x_i')^{1/k}\log(1000 \times r_i^*)$ for several values of $k = 1, \ldots, 4$ and it was observed that, $(x_i')^{1/3}\log(1000 \times r_i^*)$ did have roughly equal variances. Hence, the modified transformed data namely,

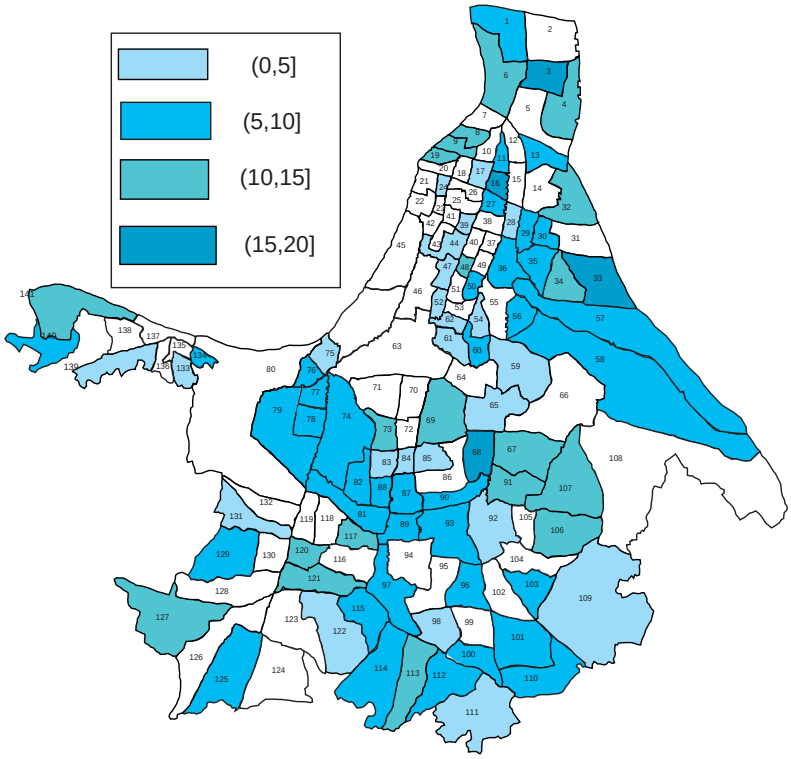
$$(x_i')^{\frac{1}{3}} Z_i = (x_i')^{\frac{1}{3}} \log(1000 \times r_i^*) \tag{4.4.1}$$

seemed useful in this respect as this transformation approximately stabilized the variance. Moreover, whereas the r_i^* values were skewed, the $(x_i')^{\frac{1}{3}} Z_i$'s were approximately Gaussian.

Next, we checked if there was any dependence of the data on any associated variable. In this survey, as in many other similar surveys, specially

where one has to work under a stringent budget, it was not possible to collect data on any covariates. So, spatial dependence was studied and for the sampled wards, the modified stroke rates per thousand namely, $1000 \times r_i^*$ were used to draw the choropleth map, as shown in Figure 1. In this figure, the wards which are not in the sample are left blank. The figure shows that there are clusters of colours, indicating a possible spatial clustering. This could be due to some unknown surrogate variable(s) on which data were not available.

Figure 1: Map showing $1000 \times r_i^*$ for sampled wards in KMC area



For the purpose of spatial smoothing, we assume that, on a map of the area, each ward is identified with a central position, called the centroid. With an arbitrary point in this map as the origin, the position of each centroid is fixed by a set of coordinates (e_i, s_i), say, e_i giving the 'longitude' and s_i the 'latitude' of centroid i. Thus, we form a spatial lattice $D = \{(e_i, s_i) : \ i = 1, 2, \ldots, 141\}$ for the 141 wards. Then we may suppose that

$$Z_i = Z(e_i, s_i) \text{ and } (Z_1, \ldots, Z_N) = (\mu_1, \ldots, \mu_N) + \delta,$$

where $N = 141$ is the number of all wards; μ_i is the mean at ward i;

$\delta = (\delta_{\mathbf{1}}, \ldots, \delta_{\mathbf{N}})$ is a vector with N elements and $\delta \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the covariance matrix of random variables at all wards. We recall that our observations are available from a two-stage sample of locations and individuals. So, in this case, the large scale variation in the mean vector for our sampled wards has been modeled as a linear model with the location parameters $\{(e_i, s_i), i \in s\}$ along with other covariates, if available, though not present here in our study. The small-scale variation is modeled by fitting a conditional autoregressive (CAR) covariance model to $\mathbf{\Sigma}'$: the submatrix of $\mathbf{\Sigma}$ for our sampled wards :

$$\mathbf{\Sigma}' = (\mathbf{I} - \rho\mathbf{W})^{-\mathbf{1}}\mathbf{\Delta}\sigma^{\mathbf{2}},$$

where ρ and σ are scalar parameters to be estimated by spatial regression, $\mathbf{I}$ is the identity matrix, $\mathbf{W}$ is the weighted neighbor matrix and $\mathbf{\Delta}$ is a diagonal matrix used to account for nonhomogeneous variance of the marginal distributions. The use of this matrix $\mathbf{\Delta}$ is to symmetrize the dispersion matrix.

We define the weighted neighbor matrix $\mathbf{W}$ following Cressie (1993, p. 557). In KMC area, the minimum and maximum distances between the sampled wards are respectively 0.25 km and 13.40 km. Any ward which is within 1 km of the ith ward is called a 'neighbor' of ward i and the set of neighbors of ward i is labelled as N_i, $i \in s$. To determine the neighbor weights we recall that $((x_i')^{\frac{1}{3}}Z_i \mid (x_j')^{\frac{1}{3}}Z_j : j \in N_i)$ has approximately equal variance, say σ^2. So, to obtain the CAR covariance matrix of $\{Z_i : i \in s\}$ as well as to symmetrize it, we choose the following distance decay weights $w_{i,j}$:

$$w_{i,j} = \frac{\min\{d_{ij} : i, j \in s, i \neq j\}}{d_{ij}} \left[\frac{x_j'}{x_i'}\right]^{\frac{1}{3}}, \quad \text{if } j \text{ is a neighbour of } i$$

$$w_{i,j} = 0, \text{ otherwise,}$$

where d_{ij} is the distance between i and j, and the diagonal matrix $\mathbf{\Delta}$ with ith diagonal entry as $\Delta_i = 1/(x_i')^{2/3}$. This choice ensures that $w_{i,j}(x_i')^{2/3} = w_{j,i}(x_j')^{2/3}$ and thus the matrix $(\mathbf{\Delta^{-1}W})$ is symmetric and therefore the CAR covariance matrix $\mathbf{\Sigma}'$ is symmetric.

Two commonly used measures of spatial correlation are Moran's autoregression coefficient I_M and Geary's contiguity ratio C. For expressions of

I_M and C, tests on them and other details, we refer to Moran (1950), Geary (1954) and Cliff & Ord (1973, 1981). Values of $I_M > \frac{-1}{n-1}$ or $0 < C < 1$ indicate positive spatial autocorrelation. For our data, on computation, we find that $I_M = 0.2486$ and $C = 0.7551$, and so, both measures indicate positive spatial dependence. Again, using the fact that I_M and C are asymptotically normally distributed with $E(I_M) = -\frac{1}{n-1}$, $E(C) = 1$ and variances depending on the spatial weights, we carried out tests for statistical significance of the observed I_M and C. Both tests showed statistically significant positive autocorrelation.

Hence, there appeared to be a spatial component to stroke in KMC area and this was used to smoothen the data. Under a conditional autoregressive (CAR) Gaussian error model, we used the S+SPATIALSTATS function slm to fit a model for Z_i. As a simple linear model did not give a satisfactory fit, a second order trend surface model with data locations (e_i, s_i) was fitted. The fitted model for Z_i retaining only statistically signficant spatial coefficients was then obtained as:

$$z_i = 0.3248 s_i + 0.0195 e_i^2 - 0.0336 e_i s_i. \tag{4.4.2}$$

The χ^2 test of goodness of fit and standard residual diagnostic techniques showed that model (4.4.2) fits the data well. Figure 2 illustrates below the raw and smoothed values of r_i^* for the sampled wards.

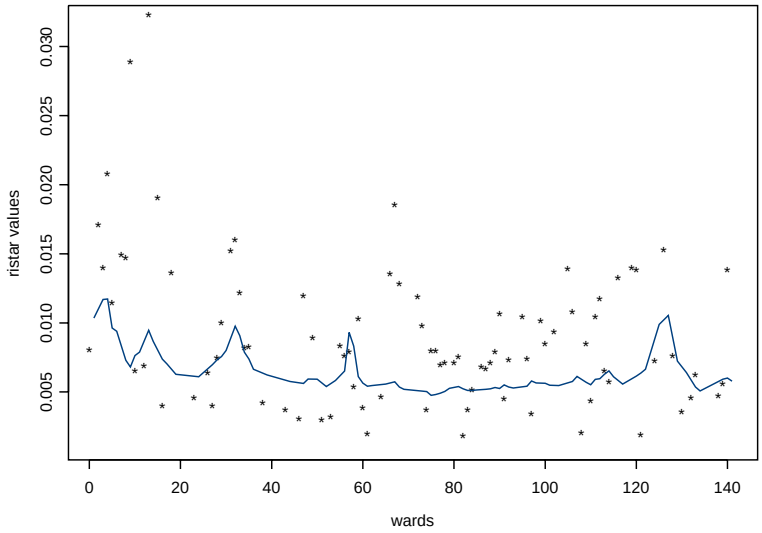
4.4.3 Use of our proposed estimator

Model (4.4.2) was used to obtain values for Z_i for the 89 sampled wards and from these, using (4.4.1), smoothed values for r_i^* were computed for $i \in s$. Now, using Corollary 4.2.1, we obtained an estimate of R together with an estimate of its variance. Let se denote the standard error of an estimate. For the sake of comparison, the coefficient of variation (cv) defined by cv $= \sqrt{\hat{V}(\hat{R})}/\hat{R} \times 100$ was also computed. In the following, the estimates of R and their corresponding standard errors are given per thousand. From our data, we get the model-cum-design based estimate as

$$\hat{R}_{\text{model-cum-design based}} \text{ (with } r_i^*) = 5.18, \text{ with se } = 0.32; \text{ cv } = 6.24\%. \tag{4.4.3}$$

Hence, our estimate of prevalence rate of stroke in KMC area is about 52 per $10,000$ and this estimate has a cv of about 6%. It is worth noting

Figure 2. Raw and smoothed r_i^* values



that this method of estimation using spatial smoothing and our proposed estimator gives an estimate with quite small cv.

Alternatively, instead of using the spatial modeling, if we compute the estimator directly from the observed ward-wise rates r_i^*, $i \in s$, then our design-based estimates come out as:

$$\hat{R}_{\text{design based}} \text{ (with } r_i^*) = 6.19 \text{ with se} = 0.52;\ \text{cv } = 8.41\%. \qquad (4.4.4)$$

Thus, the solely design based estimator in (4.4.4) has larger standard error and larger cv than the model-cum-design based one in (4.4.3) and this shows that spatial smoothing is useful in this context.

For this survey, if one had used the traditional Hartley & Ross (1954) estimator based on r_i values, then the corresponding estimates would be:

$$\hat{R}_{\text{design based}} \text{ (with } r_i) = 6.35 \text{ with se} = 0.54;\ \text{cv } = 8.44\%. \qquad (4.4.5)$$

Comparing the figures in (4.4.3), (4.4.4) and (4.4.5), we see that $\hat{R}$ in (4.4.3) has smaller se and also smaller cv than those in both (4.4.4) and (4.4.5). The reduction in cv is small if we use r_i^* instead of r_i but this reduction is much larger when the smoothing is also used as in (4.4.3). It may be noted that $\hat{R}$ in (4.4.3) is smaller than those in (4.4.4) and (4.4.5) and we can expect this to possibly inflate the cv in case (4.4.3). However,

we find that in spite of this, the cv in (4.4.3) is smaller than those in (4.4.4) and (4.4.5) by more than 2%, which is quite appreciable.

To further strengthen our arguments on the gain in spatial smoothing a small simulation exercise was also carried out. For this we considered the set of 89 wards to be the population since for these wards the y_i', x_i' values are known. From this population, we took samples of varying sizes and for these samples the above exercise was repeated, i.e. we obtained $\hat{R}$ based on the raw r_i and raw r_i^* (i.e. the design based estimates) and also $\hat{R}$ based on the spatially smoothed r_i^* (i.e. model-cum-design based estimates). The se and cv of these estimates were also calculated. The figures are summarized in Table 1 below.

Table 1: Simulation results taking random samples of size n from 89 wards

n	Moran's I_M	Geary's C	$\hat{R}$ based on model-cum-design with r_i^*		$\hat{R}$ based on only design with r_i^*		$\hat{R}$ based on only design with r_i	
			se	cv(%)	se	cv(%)	se	cv(%)
48	0.397	0.4584	0.41	8.54	0.72	10.39	0.78	11.33
45	0.2451	0.5019	0.44	8.81	0.79	11.08	0.85	11.90
40	0.6480	0.5977	0.45	8.87	0.90	13.36	0.97	14.22
35	0.7152	0.3382	0.45	10.01	1.05	15.59	1.11	16.42
30	0.8681	0.2604	0.68	12.36	1.16	17.54	1.22	18.15
25	0.4065	0.2812	0.86	19.30	1.43	21.44	1.49	21.81

Table 1 shows that the model-cum-design based estimates have consistently smaller se and cv values compared to the design-based estimates, thus indicating the gain due to spatial smoothing.

4.5 Conclusion

When estimating the prevalence rate of a disease in a geographical area from survey data based on samples taken from different locations within this area, the new estimator proposed in this article is useful for getting efficient estimates. We recommend that the sampling method should be kept simple so that it can be implemented easily in large scale surveys. Useful estimates may still be obtained by using a model-cum-design approach

by first using spatial modeling to smoothen the different location-specific observed prevalence rates for the sampled locations. Spatial smoothing is particularly useful in surveys where data on suitable covariates are not available. Our new estimator is unbiased for the population total.

Our method, when applied to real data from a survey on stroke, shows that the prevalence rate for the entire area may be estimated with a small coefficient of variation. We illustrate that a purely design based approach may be improved by taking advantage of spatial modeling. Thus, it seems that this model-cum-design based approach is worth recommending for similar studies.

Chapter 5

Estimating sensitive proportions using multiple randomized responses from distinct persons sampled

Abstract

Warner (1965) pioneered randomized response techniques to estimate the proportion of people bearing a sensitive characteristic. We present results for the situation where the distinct persons chosen in an SRSWR are identified but each one independently gives a randomized response, repeated as many times as he/she is selected. Two estimators are proposed for the sensitive proportion and compared against relevant competitors. We first study this problem where RR's are generated by Warner's device, and then extend it to the devices of Christofides (2003) and Kuk (1990).

5.1 Introduction

In some socio-ecomomic surveys, the object is to gather data regarding some sensitive or stigmatizing variable. This variable could be a qualitative one, for example, habitual tax evasion, reckless driving, indiscriminate gambling, alcoholism, and our objective is to estimate the proportion of people in a given community bearing this characteristic. Alternatively, the variable could also be a quantitative one, for example, the survey may be for

investigating the number of female foeticides among the females in a region, or the quantity of bribe accepted by a given group of people over a given period. In such situations, rather than trying to elicit a direct response (DR) from the individuals, an attractive course is to apply randomized response device techniques addressed to each person suitably sampled. These techniques have the advantage of protecting the privacy of the respondent. This encourages them to respond truthfully to the questions which are appropriately designed so as to generate enough data for suitable inference about the problem at hand.

In this chapter we focus on qualitative sensitive characteristics and begin with Warner's (1965) celebrated randomized response technique. Traditionally, a simple random sample with replacement (SRSWR) with a predetermined number of draws is taken and the over-all sample mean of a linear transform of the gathered randomized responses (RR's) is used to unbiasedly estimate the required proportion.

In direct response surveys it is known, from Basu(1958), Raj and Khamis (1958), Pathak (1962), Korwar and Serfling (1970) and others, that if one uses the responses from the distinct units sampled by SRSWR method, alternative unbiased estimators performing better than the classical estimator, namely the sample mean, are available. In this chapter we investigate the counter-part involving the randomized responses instead of direct responses.

We may note here that Warner(1965) recommended the use of the mean of *all* the RR's. The estimators proposed by us, using the same data as Warner's, is based on first averaging the RR's generated by the same respondent. It may be noted here that the proposed estimators are preferable when multiple randomized responses from distinct persons sampled are available. When RR's are gathered by Warner's device from an SRSWR, it is on record which respondent among each of the sampled ones gives the RR more than once. Our proposed estimators use exactly the same data as are used to obtain Warner's classical estimator. Thus, our proposals do not run any additional risk for jeopardizing the respondent-privacy as is inherent in the Warner's original one. So, we think that only consideration should be relative efficiencies and we have presented situations when our procedures may happen to fare better than the Warner's.

First, we consider estimation using independently repeated RR's gath-

ered by applying Warner's device to respondents chosen by an SRSWR, a scheme which still enjoys public attention. For this, in Section 5.2 we give some preliminaries, and in Section 5.3, we propose and study two new alternative estimators. The efficiencies of these estimators are also compared against certain known competitors in this section and we also provide unbiased estimators of the variances of the proposed estimators. In Sections 5.4 and 5.5 respectively, we apply the randomization devices due to Kuk (1990) and Christofides (2003) and propose alternative estimators based on distinct respondents.

5.2 Some Preliminaries

As in Section 1.2, let θ denote the proportion of persons bearing the sensitive characteristic. We recall from Section 1.2 that while applying Warner's RRD to respondents selected by an SRSWR, an independent RR is generated by each respondent every time he/she is selected. Then Warner's unbiased estimator for θ is based on all these RR's and is shown in (1.2.20); its variance being as in (1.2.22).

We remark here that we are estimating a proportion and so, ideally $\hat{\theta}$ should lie in the interval $[0, 1]$ always. But, in Warner's estimator, and also in all the other estimators we consider here, there is no in-built mechanism which will guarantee this, and there may be some pathological examples where $\hat{\theta}$ falls outside the desired range. If this happens, one takes $\hat{\theta} = 0$ or 1, according as whether the estimated value of θ is smaller than zero or exceeds unity. In such cases, our estimator $\hat{\theta}$ becomes biased and one uses the MSE to study such estimators. However, in this chapter and also in Chapter 6 where we only give possible ranges of θ, this problem is avoided.

Let $\nu(1 \leq \nu \leq n)$ be the number of distinct persons among those chosen in $n(2 \leq n < N)$ draws from U.

Mangat et al. (1995) studied the case where in the SRSWR chosen in n draws, the $\nu(1 \leq \nu < n)$ distinct persons found are requested to perform Warner's RR trial only once each. They proposed an unbiased estimator for θ based on these ν RR's and gave its variance. We denote this estimator

by $\hat{\theta}_{W1}$ and show it below, together with its variance.

$$\hat{\theta}_{W1} = \frac{(\nu'/\nu) - (1-p)}{(2p-1)}, \quad V(\hat{\theta}_{W1}) = \Phi_W E_P(\frac{1}{\nu}) + \left[NE_P(\frac{1}{\nu}) - 1\right] \frac{\theta(1-\theta)}{N-1}, \tag{5.2.1}$$

where ν' is the number of persons out of the ν distinct persons, who find match with attribute A/A^c in Warner's RRD, and all other notation are as used in (1.2.17)-(1.2.22).

Mangat et al. (1995) noted that $\hat{\theta}_{W1}$ outperforms $\hat{\theta}_W$ ($\hat{\theta}_{W1} \succ \hat{\theta}_W$), in the sense that $V(\hat{\theta}_{W1}) < V(\hat{\theta}_W)$ if N, n, p and θ happen to be such that

$$\theta(1-\theta) > \frac{n(N-1)(6N+n-1)}{N\{6Nn - 12N - n(n-1)\}} \frac{p(1-p)}{(2p-1)^2}.$$

In particular, they remarked that when $N = 100$, $n = 10$ and $p = 0.9$, $\hat{\theta}_{W1} \succ \hat{\theta}_W$ for $0.236 \leq \theta \leq 0.764$ and $\hat{\theta}_W \succ \hat{\theta}_{W1}$ otherwise. It is to be observed that a value of p acceptable to a respondent should be away from 0 and 1 and most preferably near 0.5 on either side, say, $0.45 \leq p < 0.5$ or $0.5 < p \leq 0.55$. Singh et al. (2001) also examined performances based on RR's from SRSWR using all the units and also separately only the distinct units, but their RR's are not gathered by Warner's RRD and so will not be pursued here.

Arnab (1999) permits each person in course of selection by SRSWR to perform Warner's RR trial independently each time on his/her re-appearance. The resulting estimator for θ considered by him is shown to outperform $\hat{\theta}_{W1}$ but nothing conclusive has been established about his estimator for θ versus $\hat{\theta}_W$. Chaudhuri and Pal (2008) restrict to only one RR by Warner's (1965) device from each distinct person chosen by SRSWR. They use $\hat{\theta}_W$, $\hat{\theta}_{W1}$ above and also Horvitz and Thompson's (1952) estimator. They show efficiency comparisons among the estimators with numerical illustrations and derive variance estimators.

In Section 5.3 we consider the case where an independent RR by Warner's device is obtained from each distinct person every time the person is sampled and these are utilized in constructing the estimators. We consider several estimators for using RR's elicited from distinct sampled persons in an SRSWR with n draws. The study of these estimators is facilitated by the use of results from the following works which are all in the context of estimating θ using *direct* rather than *randomized* responses through SRSWR. Basu (1958) showed that when an SRSWR based on n draws has

$\nu(1 \leq \nu \leq n)$ distinct units, the mean of these ν distinct units has a variance smaller than that of the mean of all the n observations. Raj and Khamis (1958), Pathak (1962) and Korwar and Serfling (1970) derived useful results on moments of ν and $\frac{1}{\nu}$ to help in proving the relevant results introduced by Basu (1958), who observed that the distinct observations in an SRSWR provide the sufficient statistic, given the detailed sample observations.

5.3 Estimators based on independently repeated RR's by Warner's (1965) device in SRSWR with n draws

For an SRSWR sample s in n draws, let f_{is} denote the the number of times unit i appears in the sample s. Then

$$f_{is} \leq n, \quad f_{is} \geq 1, \quad \forall \quad i \in s. \tag{5.3.1}$$

Analogous to (1.2.17), we define

$$\begin{aligned} I_{ij} &= 1, \text{ if the } i^{th} \text{ person in his/her } j^{th} \text{ appearance gets a match for} \\ &\quad \text{the card-type drawn and the true attribute } A \text{ or } A^c \\ &= 0, \quad \text{otherwise;} \quad j = 1, \ldots, f_{is}. \end{aligned} \tag{5.3.2}$$

Let u denote the *set* of distinct persons in s. Then $\nu = |u|$; the cardinality of u. We define

$$m_i = \frac{1}{f_{is}} \sum_{j=1}^{f_{is}} I_{ij} \text{ and } g_i = \frac{m_i - (1-p)}{(2p-1)}, \tag{5.3.3}$$

where as before, p is the proportion of cards marked A in Warner's RRD; $p \neq 1/2$. Then, with E_R and V_R being the RR based expectations and variances as used in (1.2.21), from (5.3.2) and (5.3.3) we get

$$\begin{aligned} E_R(I_{ij}) &= py_i + (1-p)(1-y_i) = (1-p) + (2p-1)y_i = E_R(m_i) \\ V_R(I_{ij}) &= E_R(I_{ij})(1 - E_R(I_{ij})) = p(1-p) \\ V_R(m_i) &= \frac{p(1-p)}{f_{is}} \\ E_R(g_i) &= y_i, \quad V_R(g_i) = \frac{\Phi_W}{f_{is}}. \end{aligned} \tag{5.3.4}$$

We now propose two alternative estimators for θ based on g_i for $i \in u$.

5.3.1 Two new estimators

(i) We first consider the following estimator for θ, namely

$$\hat{\theta}_{W2} = \frac{1}{\nu}\sum_{i\in u} g_i = \bar{g}(\nu), \text{ say.} \tag{5.3.5}$$

Similarly, let us define

$$\bar{y}(\nu) = \frac{1}{\nu}\sum_{i\in u} y_i. \tag{5.3.6}$$

Using (5.3.4) it then follows that

$$E(\hat{\theta}_{W2}) = E_P E_R\left(\hat{\theta}_{W2}\right) = E_P\left(\frac{1}{\nu}\sum_{i\in u} y_i\right) = E_P(\bar{y}(\nu)) = \theta$$

and hence $\hat{\theta}_{W2}$ is unbiased for θ.

Pathak (1962) proved some results in the context of direct response surveys which we can apply in our context. We state his results below. Writing $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2 = \frac{N\theta(1-\theta)}{N-1}$, he showed that

$$\begin{aligned}
V_P\left(\bar{y}(\nu)\right) &= \left[E_P(\tfrac{1}{\nu}) - \tfrac{1}{N}\right] S^2 = \left[E_P(\tfrac{1}{\nu}) - \tfrac{1}{N}\right] \tfrac{N\theta(1-\theta)}{N-1}, \\
E_P(\tfrac{1}{\nu}) &= \frac{1}{N^n}\sum_{l=1}^{N} l^{n-1} \\
V_P\left[\nu\bar{y}(\nu)\right] &= N\theta(1-\theta)\left[(1-\tfrac{1}{N})^n - (1-\tfrac{2}{N})^n\right] \\
&\quad +\theta^2 N\left[(1-\tfrac{1}{N})^n - N(1-\tfrac{1}{N})^{2n} + (N-1)(1-\tfrac{2}{N})^n\right].
\end{aligned} \tag{5.3.7}$$

Using these results and (5.3.4), the variance of $\hat{\theta}_{W2}$ is

$$\begin{aligned}
V(\hat{\theta}_{W2}) &= V_P E_R(\hat{\theta}_{W2}) + E_P V_R(\hat{\theta}_{W2}) \\
&= V_P[\tfrac{1}{\nu}\textstyle\sum_{i\in u} y_i] + E_P[\tfrac{1}{\nu^2}\sum_{i\in u}\tfrac{\Phi_W}{f_{is}}] \\
&= [E_P(\tfrac{1}{\nu}) - \tfrac{1}{N}]S^2 + \Phi_W E_P\left[\tfrac{1}{\nu^2}\textstyle\sum_{i\in u}\tfrac{1}{f_{is}}\right] \\
&= \left[\tfrac{1}{N^{n-1}(N-1)}\textstyle\sum_{l=1}^{N-1} l^{n-1}\right](\theta - \theta^2) + \Phi_W E_P(\tfrac{1}{\nu^2}\textstyle\sum_{i\in u}\tfrac{1}{f_{is}}).
\end{aligned} \tag{5.3.8}$$

(ii) Next, we consider another estimator for θ, namely, the Horvitz Thompson estimator (HTE) in this case as

$$\hat{\theta}_{W3} = \frac{1}{N}\sum_{i\in u}\frac{g_i}{\pi_i}. \tag{5.3.9}$$

For SRSWR it is clear that the inclusion probabilities are given by: $\pi_i = 1 - (1-\frac{1}{N})^n$ for every i and $\pi_{ij} = 1 - 2(1-\frac{1}{N})^n + (1-\frac{2}{N})^n$ for every

$i, j(i \neq j)$. Thus,

$$E_P(\nu) = \sum_{i=1}^{N} \pi_i = N\left[1 - (1 - \frac{1}{N})^n\right]. \tag{5.3.10}$$

Now using (5.3.4) and (5.3.7) and noting that π_i is a constant independent of i, it follows that

$$\begin{aligned} V(\hat{\theta}_{W3}) &= \tfrac{\Phi_W}{N^2\pi_i^2} E_P\left[\textstyle\sum_{i\in u} \frac{1}{f_{is}}\right] + \tfrac{1}{N^2\pi_i^2} V_P\left[\nu\bar{y}(\nu)\right] \qquad (5.3.11) \\ &= \tfrac{\Phi_W}{N^2\pi_i^2} E_P\left[\textstyle\sum_{i\in u} \frac{1}{f_{is}}\right] \\ &\quad + \tfrac{\theta(1-\theta)}{N(N-1)\pi_i^2}\left[(N-1)\left\{(1-\tfrac{1}{N})^n - (1-\tfrac{2}{N})^n\right\}\right] \\ &\quad + \tfrac{\theta^2}{N^2\pi_i^2}\left[N(1-\tfrac{1}{N})^n - N^2(1-\tfrac{1}{N})^{2n} + N(N-1)(1-\tfrac{2}{N})^n\right]. \qquad (5.3.12) \end{aligned}$$

5.3.2 Efficiency comparisons among $\hat{\theta}_W, \hat{\theta}_{W1}, \hat{\theta}_{W2}, \hat{\theta}_{W3}$.

In this section we compare the hitherto known estimators $\hat{\theta}_W$ and $\hat{\theta}_{W1}$ given in (1.2.20) and (5.2.1), with our proposed estimators $\hat{\theta}_{W2}$ and $\hat{\theta}_{W3}$ given in (5.3.5) and (5.3.9). From (5.3.1), we have

$$\sum_{i\in u} \frac{1}{f_{is}} \leq \nu, \quad \sum_{i\in u} \frac{1}{f_{is}} \geq \frac{\nu}{n}. \tag{5.3.13}$$

(i) Comparing $\hat{\theta}_{W2}$ with $\hat{\theta}_{W1}$: On comparing (5.2.1) and (5.3.8), it follows from the first inequality in (5.3.13) that $V(\hat{\theta}_{W2}) \leq V(\hat{\theta}_{W1})$. Hence our proposed estimator $\hat{\theta}_{W2}$ uniformly outperforms $\hat{\theta}_{W1}$, i.e.

$$\hat{\theta}_{W2} \succ \hat{\theta}_{W1} \text{ uniformly.}$$

(ii) Comparing $\hat{\theta}_{W3}$ with $\hat{\theta}_{W1}$: Here, an inequality due to Korwar and Serfling (1970) will be useful, and we first state it below. With $Q = \frac{1}{n} + \frac{1}{2N} + \frac{n-1}{12N^2}$ and for $n \geq 3$,

$$Q - \frac{1}{720N} < E_P(\frac{1}{\nu}) \leq Q. \tag{5.3.14}$$

Now, from (5.2.1) and (5.3.12), we may write on simplification

$$V(\hat{\theta}_{W1}) - V(\hat{\theta}_{W3}) = \Phi_W\left[E_P\left(\frac{1}{\nu}\right) - \frac{1}{N^2\pi_i^2} E_P(\sum_{i\in u}\frac{1}{f_{is}})\right] + \frac{N\theta(1-\theta)}{N-1} A_1 - \theta^2 A_2,$$

where

$$A_1 = \frac{1}{N^n}\sum_{l=1}^{N-1} l^{n-1} - \frac{N-1}{N^2\pi_i^2}\left\{(1-\frac{1}{N})^n - (1-\frac{2}{N})^n\right\}$$

and

$$A_2 = \frac{1}{N\pi_i^2}\left[(1-\frac{1}{N})^n - N(1-\frac{1}{N})^{2n} + (N-1)(1-\frac{2}{N})^n\right].$$

Hence using the second inequality in (5.3.13), (5.3.10), and the upper bound of $E_P(\frac{1}{\nu})$ from (5.3.14), it can be seen that

$$\begin{aligned} V(\hat{\theta}_{W1}) - V(\hat{\theta}_{W3}) &\leq \Phi_W\left(Q - \tfrac{E_P(\nu)}{nN^2\pi_i^2}\right) + \tfrac{N\theta(1-\theta)}{N-1}A_1 - \theta^2 A_2 \\ &\leq \Phi_W\left[Q - \tfrac{1}{nN\pi_i}\right] + \tfrac{N\theta(1-\theta)}{N-1}A_1 - \theta^2 A_2. \end{aligned}$$

Thus

$$V(\hat{\theta}_{W1}) \leq V(\hat{\theta}_{W3}) \text{ if } \frac{N\theta(1-\theta)}{N-1}A_1 - \theta^2 A_2 \leq \Phi_W(\frac{1}{nN\pi_i} - Q) \quad (5.3.15)$$

Similarly, using the first inequality in (5.3.13), (5.3.10), and the lower bound of $E_P(\frac{1}{\nu})$ from (5.3.14), it can be seen on simplification that

$$V(\hat{\theta}_{W1}) - V(\hat{\theta}_{W3}) > \Phi_W\left[Q - \tfrac{1}{720N} - \tfrac{1}{N\pi_i}\right] + \tfrac{N\theta(1-\theta)}{N-1}A_1 - \theta^2 A_2.$$

So,

$$V(\hat{\theta}_{W3}) < V(\hat{\theta}_{W1}) \text{ if } \frac{N\theta(1-\theta)}{N-1}A_1 - \theta^2 A_2 \geq \Phi_W\left[\frac{1}{N\pi_i} - Q + \frac{1}{720N}\right] \quad (5.3.16)$$

In practice, since we will usually have $\frac{1}{720N} \approx 0$, we may conclude that

$$V(\hat{\theta}_{W3}) < V(\hat{\theta}_{W1}) \text{ if } \frac{N\theta(1-\theta)}{N-1}A_1 - \theta^2 A_2 \geq \Phi_W\left[\frac{1}{N\pi_i} - Q\right]. \quad (5.3.17)$$

Thus, $\hat{\theta}_{W3} \succ \hat{\theta}_{W1}$ if (5.3.17) holds and $\hat{\theta}_{W1} \succ \hat{\theta}_{W3}$ if (5.3.15) holds.

Following Pathak (1962), for large N we may approximate A_1 and A_2 by $\tilde{A}_1$ and $\tilde{A}_2$, respectively, with

$$\tilde{A}_1 = \frac{1}{2nN} + \frac{5(n-1)}{12nN^2} \text{ and } \tilde{A}_2 = \frac{n-1}{2nN} - \frac{(n-1)(n-2)}{3nN^2}$$

Now, from (5.3.17), for given N, n and p, we can compute the range of θ for which $\hat{\theta}_{W3} \succ \hat{\theta}_{W1}$. We give some illustrations in Table 1.

In situations where RRD will be applied, we may expect θ to be small. So, the illustrations in Table 1 show that for most θ values we may encounter in practice, $\hat{\theta}_{W3}$ will outperform $\hat{\theta}_{W1}$.

(iii) Comparing $\hat{\theta}_{W2}$ with $\hat{\theta}_W$: From (1.2.22) and (5.3.8), on using the first inequality in (5.3.13) and the upper bound of $E_P(\frac{1}{\nu})$ from (5.3.14), we have

$$\begin{aligned} V(\hat{\theta}_{W2}) - V(\hat{\theta}_W) &= \Phi_W \left[E_P \left(\tfrac{1}{\nu^2} \textstyle\sum_{i \in u} \frac{1}{f_{is}} \right) - \frac{1}{n} \right] + \theta(1-\theta) \left[\frac{NE_P(\frac{1}{\nu}) - 1}{N-1} - \frac{1}{n} \right] \\ &\leq \Phi_W \left[Q - \tfrac{1}{n} \right] + \theta(1-\theta) \left[\tfrac{NQ-1}{N-1} - \tfrac{1}{n} \right]. \end{aligned}$$

So,

$$V(\hat{\theta}_{W2}) \leq V(\hat{\theta}_W), \text{ if } \theta(1-\theta) \geq \Phi_W (N-1) \left(\frac{nQ-1}{N+n-1-NnQ} \right). \tag{5.3.18}$$

Similarly, from (1.2.22) and (5.3.8), using (5.3.13) and (5.3.14), one gets

$$V(\hat{\theta}_{W2}) - V(\hat{\theta}_W) > \frac{\Phi_W}{n} \left[Q - \frac{1}{720N} - 1 \right] + \theta(1-\theta) \left[\frac{NQ - \frac{721}{720}}{N-1} - \frac{1}{n} \right]$$

which implies that

$$V(\hat{\theta}_{W2}) > V(\hat{\theta}_W) \text{ if } \theta(1-\theta) \leq \Phi_W (N-1) \left(\frac{Q - \frac{1}{720N} - 1}{N + \frac{721n}{720} - 1 - NnQ} \right). \tag{5.3.19}$$

Using (5.3.18), we can ascertain when $\hat{\theta}_{W2} \succ \hat{\theta}_W$. We give some numerical illustrations in Table 1. However, it is hard to find feasible values of θ satisfying (5.3.19) for appropriate values of N, n, p, thereby preventing $\hat{\theta}_W$ from outperforming $\hat{\theta}_{W2}$.

(iv) Comparing $\hat{\theta}_{W3}$ with $\hat{\theta}_W$: From (1.2.21) and (5.3.11) and applying (5.3.13) and (5.3.10) as in the earlier cases, it follows that

$$V(\hat{\theta}_{W3}) - V(\hat{\theta}_W) = \Phi_W \left[\frac{1}{N^2 \pi_i^2} E_P \left(\sum_{i \in u} \frac{1}{f_{is}} \right) - \frac{1}{n} \right] + V_P \left[\frac{\nu}{N\pi_i} \bar{y}(\nu) \right] - V_P(\bar{y}(n))$$

$$\leq \Phi_W \left[\frac{1}{N\pi_i} - \frac{1}{n} \right] + V_P \left[\frac{\nu}{N\pi_i} \bar{y}(\nu) \right] - V_P(\bar{y}(n))$$

where $\bar{y}(n)$ and $\bar{y}(\nu)$ are as used in (1.2.21) and defined in (5.3.6) respectively. Hence we have

$$V(\hat{\theta}_{W3}) \leq V(\hat{\theta}_W) \text{ if } V_P \left[\frac{\nu}{N\pi_i} \bar{y}(\nu) \right] - V_P(\bar{y}(n)) + \Phi_W \left[\frac{n - N\pi_i}{nN\pi_i} \right] \leq 0.$$

Now on using (5.3.7) and after some algebra it can be shown that

$$V(\hat{\theta}_{W3}) \leq V(\hat{\theta}_W) \text{ if}$$

$$\theta^2\left[A_2+\frac{NA_1}{N-1}-\frac{1}{N-1}\sum_{l=1}^{N-1}\left(\frac{l}{N}\right)^{n-1}+\frac{1}{n}\right]+\theta\left[\frac{1}{N-1}\sum_{l=1}^{N-1}\left(\frac{l}{N}\right)^{n-1}-\frac{NA_1}{N-1}-\frac{1}{n}\right]$$
$$+\Phi_W\left[\frac{n-N\pi_i}{nN\pi_i}\right]\leq 0. \qquad (5.3.20)$$

Similarly, from (1.2.21) and (5.3.11), on applying the other inequality in (5.3.13) and (5.3.10), we can show that

$$V(\hat{\theta}_W)\leq V(\hat{\theta}_{W3}) \text{ if}$$

$$\theta^2\left[A_2+\frac{NA_1}{N-1}-\frac{1}{N-1}\sum_{l=1}^{N-1}\left(\frac{l}{N}\right)^{n-1}+\frac{1}{n}\right]+\theta\left[\frac{1}{N-1}\sum_{l=1}^{N-1}\left(\frac{l}{N}\right)^{n-1}-\frac{NA_1}{N-1}-\frac{1}{n}\right]$$
$$+\frac{\Phi_W}{n}\left[\frac{1-N\pi_i}{N\pi_i}\right]\geq 0. \qquad (5.3.21)$$

For some illustrative values of N, n and p, Table 1 shows the ranges of θ for which $\hat{\theta}_{W3}\succ\hat{\theta}_W$.

(v) Comparing $\hat{\theta}_{W2}$ with $\hat{\theta}_{W3}$: From (5.3.8) and (5.3.12),

$$V(\hat{\theta}_{W2})-V(\hat{\theta}_{W3})=\Phi_W\left[E_P(\frac{1}{\nu^2}\sum_{i\in u}\frac{1}{f_{is}})-\frac{1}{N^2\pi_i^2}E_P(\sum_{i\in u}\frac{1}{f_{is}})\right]+\frac{N\theta(1-\theta)}{N-1}A_1-\theta^2A_2.$$

Then, as earlier, it can be shown that

$$V(\hat{\theta}_{W2})\leq V(\hat{\theta}_{W3}) \text{ if (5.3.15) holds.}$$

Again, if we assume that $\frac{1}{720Nn}$ is negligible, then,

$$V(\hat{\theta}_{W2})>V(\hat{\theta}_{W3}) \text{ if } \frac{N\theta(1-\theta)A_1}{N-1}-\theta^2A_2\geq\Phi_W\left[\frac{1}{N\pi_i}-\frac{Q}{n}\right].$$

For some N, n and p, Table 1 shows the ranges of θ for which $\hat{\theta}_{W2}\succ\hat{\theta}_{W3}$.

Table 1: Some illustrations showing efficiency comparisons of the alternative estimators under Warner's model.

Comparison	N	n	p ($\neq 0.5$)	θ
$\hat{\theta}_{W2} \succ \hat{\theta}_{W1}$	any	any	any	any
$\hat{\theta}_{W3} \succ \hat{\theta}_{W1}$	100	10	0.40, 0.60	≤ 0.843
	200	50	0.45, 0.55	≤ 0.762
	200	40	0.45, 0.55	≤ 0.840
	200	30	0.45, 0.55	≤ 0.959
	250	50	0.45, 0.55	≤ 0.752
$\hat{\theta}_{W2} \succ \hat{\theta}_{W}$	100	10	0.9, 0.1	$0.235 \leq \theta \leq 0.765$
	200	30	0.9, 0.1	$0.196 \leq \theta \leq 0.804$
$\hat{\theta}_{W3} \succ \hat{\theta}_{W}$	100	10	0.92, 0.08	$0.150 \leq \theta \leq 0.365$
	100	10	0.95, 0.05	$0.069 \leq \theta \leq 0.445$
	200	30	0.92, 0.08	$0.155 \leq \theta \leq 0.366$
$\hat{\theta}_{W2} \succ \hat{\theta}_{W3}$	158	49	0.93, 0.07	$\theta \geq 0.909$
	158	49	0.94, 0.06	$\theta \geq 0.828$
	200	49	0.94, 0.06	$\theta \geq 0.893$
	200	49	0.95, 0.05	$\theta \geq 0.803$

5.3.3 Unbiased variance estimators

Now we give the variance estimators for the four estimators considered so far in this chapter. Among them, the variance estimator for $\hat{\theta}_W$ was given by Warner and this is stated first; for the other three estimators we obtain the variance estimators as shown below.

1. For $\hat{\theta}_W$ as in (1.2.20), it is clear that an unbiased estimator for $V(\hat{\theta}_W)$ is

$$\hat{V}(\hat{\theta}_W) = \frac{1}{n(n-1)} \sum_{k=1}^{n} (r_k - \bar{r}(n))^2$$

where as in (1.2.20) $\bar{r}(n) = \frac{1}{n}\sum_{k=1}^{n} r_k = \hat{\theta}_W$. With $\hat{\lambda}$ as used in (1.2.20), and as shown by Warner(1965), this reduces to

$$\hat{V}(\hat{\theta}_W) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)(2p-1)^2}.$$

2. For $\hat{\theta}_{W1}$ as in (5.2.1), in order to provide unbiased estimators of $V(\hat{\theta}_{W1})$, we recall that in this case, we generate only a single RR for every

distinct person sampled and then consider the r_k as used in (1.2.20) for $k \in u$. Let

$$C_1 = (N-1)\left[(1-\frac{1}{N})^n - (1-\frac{2}{N})^n\right]$$

$$C_2 = N(1-\frac{1}{N})^n - N^2(1-\frac{1}{N})^{2n} + N(N-1)(1-\frac{2}{N})^n. \qquad (5.3.22)$$

Then, since r_i, r_j's $\forall i \neq j$ are independent, using results from Pathak (1962) we may derive the following:

$$E_P E_R\left[\sum\sum_{i\neq j, i,j\in u} r_i r_j\right] = E_P\left[\sum\sum_{i\neq j, i,j\in u} y_i y_j\right]$$

$$= \theta\left[C_1\frac{N}{N-1} - N\pi_i\right] + \theta^2\left[C_2 + N^2\pi_i^2 - C_1\frac{N}{N-1}\right].$$

Hence an unbiased estimator for θ^2 is

$$(\widehat{\theta^2})_W = \frac{1}{C_2 + N^2\pi_i^2 - C_1\frac{N}{N-1}}\left[\sum\sum_{i\neq j, i,j\in u} r_i r_j - \hat{\theta}_{W1}\left(C_1\frac{N}{N-1} - N\pi_i\right)\right].$$

So two alternative unbiased estimators for $V(\hat{\theta}_{W1})$ may be obtained as

$$v_1(\hat{\theta}_{W1}) = \Phi_W E_P(\frac{1}{\nu}) + \left[\frac{NE_P(\frac{1}{\nu}) - 1}{N-1}\right]\left(\hat{\theta}_{W1} - (\widehat{\theta^2})_W\right),$$

and also as

$$v_2(\hat{\theta}_{W1}) = \Phi_W\frac{1}{\nu} + \left[\frac{NE_P(\frac{1}{\nu}) - 1}{N-1}\right]\left(\hat{\theta}_{W1} - (\widehat{\theta^2})_W\right)$$

3. For $\hat{\theta}_{W2}$ as in (5.3.5), we may provide unbiased estimators of $V(\hat{\theta}_{W2})$ as follows: Using (5.3.22) and results from Pathak (1962) we derive the following:

$$E_P E_R\left[\sum\sum_{i\neq j, i,j\in u} g_i g_j\right] = \theta\left[C_1\frac{N}{N-1} - N\pi_i\right] + \theta^2\left[C_2 + N^2\pi_i^2 - C_1\frac{N}{N-1}\right].$$

So, an unbiased estimator for θ^2 is

$$(\widetilde{\theta^2})_W = \frac{1}{C_2 + N^2\pi_i^2 - C_1\frac{N}{N-1}}\left[\sum\sum_{i\neq j, i,j\in u} g_i g_j - \hat{\theta}_{W2}(C_1\frac{N}{N-1} - N\pi_i)\right].$$

Thus, an unbiased estimator for

$$\left[E_P(\frac{1}{\nu}) - \frac{1}{N}\right]\frac{N\theta(1-\theta)}{N-1} \text{ is } \left[\frac{1}{N^{n-1}(N-1)}\sum_{l=1}^{N-1} l^{n-1}\right](\hat{\theta}_{W2} - (\widetilde{\theta^2})_W)$$

and consequently, from (5.3.8) an unbiased estimator for $V(\hat{\theta}_{W2})$ is given by

$$v_1(\hat{\theta}_{W2}) = \left[\frac{1}{N^{n-1}(N-1)}\sum_{l=1}^{N-1} l^{n-1}\right](\hat{\theta}_{W2} - (\widetilde{\theta^2})_W) + \Phi_W(\frac{1}{\nu^2}\sum_{i\in u}\frac{1}{f_{is}})$$

To derive a few more unbiased estimators of $V(\hat{\theta}_{W2})$, we define the following terms. Here $\bar{g}(\nu)(=\hat{\theta}_{W2})$ is defined in (5.3.5).

$$f_1(\nu) = \frac{1}{\nu-1}\sum_{i\in u}(g_i - \bar{g}(\nu))^2, \quad f_2(\nu) = \frac{1}{\nu-1}\sum_{i\in u}(y_i - \bar{y}(\nu))^2,$$

$$C(n) = \sum_{l=0}^{\nu-1}(-1)^l\binom{\nu}{l}(\nu-l)^n$$

$$C_3 = \frac{1}{N^n}\left(\sum_{l=1}^{N-1} l^{n-1}\right)\frac{N}{N-1}\frac{C(n)-C(n-1)}{C(n)},$$

$$C_4 = \frac{C(n-1)}{C(n)},\ C_5 = \left[(\frac{1}{\nu}-\frac{1}{N})+(\frac{N-1}{N^n-N})\right],\ C_6 = \left[(\frac{1}{\nu}-\frac{1}{N})+N^{1-n}(1-\frac{1}{\nu})\right].$$

Then, using $E_P[C_i f_2(\nu)] = [E_P(\frac{1}{\nu}) - \frac{1}{N}]S^2, i = 3,\ldots,6$ from Pathak (1962), on simplification, it follows that

$$\begin{aligned}
E_P E_R[C_3 f_1(\nu)] &= C_3 E_P E_R\left[(\tfrac{1}{\nu-1})\left\{\textstyle\sum_{i\in u}(g_i - \frac{\sum_{i\in u} g_i}{\nu})^2\right\}\right] \\
&= C_3 E_P E_R\left[(\tfrac{1}{\nu-1})\left\{\textstyle\sum_{i\in u} g_i^2 - \nu\{\bar{g}(\nu)\}^2\right\}\right] \\
&= C_3 E_P\left[(\tfrac{1}{\nu-1})\textstyle\sum_{i\in u}\{V_R(g_i) + (E_R(g_i))^2\}\right. \\
&\quad \left. - (\tfrac{\nu}{\nu-1})\{V_R(\bar{g}(\nu)) + (E_R(\bar{g}(\nu)))^2\}\right] \\
&= C_3 E_P\left[\tfrac{1}{\nu-1}\textstyle\sum_{i\in u}\left(\frac{\Phi_W}{f_{is}}+y_i^2\right) - (\tfrac{\nu}{\nu-1})\left(\frac{1}{\nu^2}\sum_{i\in u}\frac{\Phi_W}{f_{is}}+\{\bar{y}(\nu)\}^2\right)\right] \\
&= C_3 E_P\left[\tfrac{\Phi_W}{\nu}\textstyle\sum_{i\in u}\frac{1}{f_{is}} + \frac{1}{\nu-1}\{\sum_{i\in u} y_i^2 - \nu\{\bar{y}(\nu)\}^2\}\right] \\
&= C_3 E_P\left[\tfrac{\Phi_W}{\nu}\textstyle\sum_{i\in u}\frac{1}{f_{is}}\right] + E_P[C_3 f_2(\nu)] \\
&= E_P E_R\left[C_3\tfrac{\Phi_W}{\nu}\textstyle\sum_{i\in u}\frac{1}{f_{is}}\right] + \left[E_P(\tfrac{1}{\nu}) - \tfrac{1}{N}\right]S^2.
\end{aligned}$$

So, $C_3 f_1(\nu) - C_3\frac{\Phi_W}{\nu}\sum_{i\in u}\frac{1}{f_{is}}$ is an unbiased estimator of $V_P(\bar{y}(\nu)) = \left[E_P(\frac{1}{\nu}) - \frac{1}{N}\right]S^2$. So another unbiased estimator of $V(\hat{\theta}_{W2})$ is

$$\begin{aligned}
v_2(\hat{\theta}_{W2}) &= \tfrac{\Phi_W}{\nu^2}\textstyle\sum_{i\in u}\frac{1}{f_{is}} + C_3 f_1(\nu) - C_3\frac{\Phi_W}{\nu}\sum_{i\in u}\frac{1}{f_{is}} \\
&= \tfrac{\Phi_W}{\nu}(\tfrac{1}{\nu} - C_3)\textstyle\sum_{i\in u}\frac{1}{f_{is}} + C_3 f_1(\nu).
\end{aligned}$$

Similarly, we get three more unbiased estimators of $V(\hat{\theta}_{W2})$, namely

$$v_i(\hat{\theta}_{W2}) = \frac{\Phi_W}{\nu}\left(\frac{1}{\nu} - C_{i+1}\right)\sum_{i\in u}\frac{1}{f_{is}} + C_{i+1} f_1(\nu), \quad i = 3,4,5$$

4. For $\hat{\theta}_{W3}$ as in (5.3.9), we may provide unbiased estimators of $V(\hat{\theta}_{W3})$ as follows: Let

$$f_3(\nu) = \frac{1}{N^2}\left[\sum_{i\in u} g_i^2\left(\frac{1-\pi_i}{\pi_i^2}\right) + \sum\sum_{i\neq j, i,j\in u} g_i g_j\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\right)\right].$$

Then

$$\begin{aligned}
E(f_3(\nu)) &= E_P E_R(f_3(\nu)) \\
&= \tfrac{1}{N^2}\left[\textstyle\sum_{i=1}^N y_i^2\left(\frac{1-\pi_i}{\pi_i}\right) + \Phi_W E_P\left(\sum_{i\in u}\frac{1}{f_{is}}\left(\frac{1-\pi_i}{\pi_i^2}\right)\right) + \sum\sum_{i\neq j}^N y_i y_j\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\right] \\
&= V_P\left(\tfrac{1}{N}\textstyle\sum_{i\in u}\frac{y_i}{\pi_i}\right) + \frac{\Phi_W}{N^2}\frac{1-\pi_i}{\pi_i^2}E_P\left(\sum_{i\in u}\frac{1}{f_{is}}\right) \\
&= V_P\left[E_R(\hat{\theta}_{W3})\right] + \frac{\Phi_W}{N^2}\frac{1-\pi_i}{\pi_i^2}E_P\left(\sum_{i\in u}\frac{1}{f_{is}}\right) \\
&= V(\hat{\theta}_{W3}) - E_P\left[V_R(\hat{\theta}_{W3})\right] + \frac{\Phi_W}{N^2}\frac{1-\pi_i}{\pi_i^2}E_P\left(\sum_{i\in u}\frac{1}{f_{is}}\right).
\end{aligned}$$

Hence

$$V(\hat{\theta}_{W3}) = E_P E_R(f_3(\nu)) + E_P\left[\frac{\Phi_W}{N^2\pi_i^2}\sum_{i\in u}\frac{1}{f_{is}}\right] - \frac{\Phi_W}{N^2}\frac{1-\pi_i}{\pi_i^2}E_P\left[\sum_{i\in u}\frac{1}{f_{is}}\right].$$

Thus an unbiased estimator for $V(\hat{\theta}_{W3})$ is given by

$$v(\hat{\theta}_{W3}) = f_3(\nu) + \frac{\Phi_W}{\pi_i N^2}\sum_{i\in u}\frac{1}{f_{is}}.$$

5.4 Estimators based on independently repeated RR's by Kuk's (1990) device in SRSWR with n draws

Kuk (1990) proposed a scheme for generating randomized responses and gave an unbiased estimator for θ. In this section, we study some alternative unbiased estimators under this scheme. In Kuk's (1990) device, each of two boxes is filled with cards of two types, say, red and blue; with their mixing proportions being $p_1 : (1-p_1); (0 < p_1 < 1)$ in one box and $p_2 : (1-p_2); (0 < p_2 < 1)$ in the other; $p_1 \neq p_2$ and $p_1 + p_2 \neq 1$. Each selected person in an SRSWR is given these two boxes and requested to draw cards L times independently, either from the first box or from the second, according as whether this person bears characteristic A or not. The person selected on the k^{th} draw $(k = 1, \ldots, n)$ from U, is requested to give the randomized response as the number of red cards out of the L cards drawn. Thus each

time a person is selected, the person draws from only one box and gives a single response. The interviewer does not see which box the card is drawn from and thus the respondent's privacy is protected.

Let the randomized response given by the respondent selected at k_{th} draw be denoted by h_k. Then clearly,

$$E_R(h_k) = L[p_1 y_k + p_2(1-y_k)] \text{ and } V_R(h_k) = L\left[p_1(1-p_1)y_k + p_2(1-p_2)(1-y_k)\right].$$

Accordingly, we construct the transformed variable from h_k as

$$\rho_k = \left(\frac{h_k}{L} - p_2\right)/(p_1 - p_2)$$

such that

$$\begin{aligned} E_R\left[\rho_k\right] &= y_k, \\ V_R(\rho_k) &= \frac{(p_1-p_2)(1-p_1-p_2)y_k + p_2(1-p_2)}{L(p_1-p_2)^2} \\ &= a y_k + b \end{aligned} \tag{5.4.1}$$

say, with $a = \frac{1-p_1-p_2}{L(p_1-p_2)}$ and $b = \frac{p_2(1-p_2)}{L(p_1-p_2)^2}$.

It is now easy to verify that the unbiased estimator for θ as given by Kuk (1990) can be expressed in the form

$$\hat{\theta}_K = \frac{1}{n}\sum_{k=1}^{n} \rho_k, \tag{5.4.2}$$

and from (5.4.1), it is easy to see that

$$E(\hat{\theta}_K) = \theta, \quad V(\hat{\theta}_K) = \frac{\theta}{n}(1 + a - \theta) + \frac{b}{n} \tag{5.4.3}$$

with a and b as in (5.4.1).

5.4.1 Some alternative estimators

Now, we look for other possible alternative estimators for θ. For this, one simple option is to consider the case where each distinct respondent is asked to give a randomized response by Kuk's device only once. Thus, instead of the n responses used in θ_K as in (5.4.2), we now have only ν responses, say, $h_i, i \in u$. Based on each such h_i, we may construct a transformed response ρ_i for $i \in u$, as in (5.4.1). Then, corresponding to $\hat{\theta}_{W1}$ in (5.2.1), we may construct an unbiased estimator, say $\hat{\theta}_{K1}$. The Horvitz Thompson estimator

based on these responses, say $\hat{\theta}_{KHT}$, can also be constructed. These are as follows:

$$\hat{\theta}_{K1} = \frac{1}{\nu}\sum_{i\in u}\rho_i, \quad \hat{\theta}_{KHT} = \frac{1}{N}\sum_{i\in u}\frac{\rho_i}{\pi_i}, \tag{5.4.4}$$

where π_i is the inclusion probability of i, as usual. We note here that the estimator corresponding to $\hat{\theta}_{KHT}$ for Warner's model was studied by Chaudhuri and Pal (2008). Using (5.4.1), (5.3.7) and (5.3.22), we can obtain their variances as:

$$V(\hat{\theta}_{K1}) = (a\theta + b)E_P\left(\frac{1}{\nu}\right) + \left[NE_P\left(\frac{1}{\nu}\right) - 1\right]\frac{\theta(1-\theta)}{N-1}, \tag{5.4.5}$$

$$V(\hat{\theta}_{KHT}) = \theta\left[\frac{a}{N\pi_i} + \frac{C_1}{N\pi_i^2(N-1)}\right] + \theta^2\left[\frac{C_2}{N^2\pi_i^2} - \frac{C_1}{N\pi_i^2(N-1)}\right] + \frac{b}{N\pi_i}, \tag{5.4.6}$$

with C_1 and C_2 as in (5.3.22) and a, b as in (5.4.1).

Now, we consider the case where respondent i is asked to give a randomized response by Kuk's device whenever he is drawn in the sample s, as was the case for $\hat{\theta}_K$; but this time as in Section 5.3, we first average the f_{is} responses thus received from each distinct respondent. Towards this, suppose h_{ij} is the response (i.e., the number of red cards out of L) of the i^{th} respondent at his j^{th} appearance in the sample; $j = 1, \ldots, f_{is}$. Then, clearly,

$$E_R(h_{ij}) = L\left[p_1 y_i + p_2(1-y_i)\right] = L\left[p_2 + (p_1 - p_2)y_i\right].$$

Similar to the quantities in (5.3.3), we now write

$$\bar{h}_i = \frac{1}{f_{is}}\sum_{j=1}^{f_{is}} h_{ij}, \quad \gamma_i = \frac{(\frac{\bar{h}_i}{L} - p_2)}{(p_1 - p_2)}. \tag{5.4.7}$$

It can be now seen that

$$\begin{array}{lcl} E_R(\bar{h}_i) & = & L\left[p_2 + (p_1 - p_2)y_i\right] \\ V_R(\bar{h}_i) & = & \frac{L}{f_{is}}\left[p_2(1-p_2) + (p_1 - p_2)(1 - p_1 - p_2)y_i\right] \\ E_R(\gamma_i) & = & y_i \\ V_R(\gamma_i) & = & \frac{ay_i + b}{f_{is}} \end{array} \tag{5.4.8}$$

with a, b as in (5.4.1).

Now, as in Section 5.3, with γ_i as in (5.4.7) we may construct two new estimators for θ as follows:

$$\hat{\theta}_{K2} = \frac{1}{\nu}\sum_{i\in u}\gamma_i, \quad \hat{\theta}_{K3} = \frac{1}{N}\sum_{i\in u}\frac{\gamma_i}{\pi_i}. \tag{5.4.9}$$

From (5.4.8), (5.4.9), (5.3.7) and (5.3.22) it follows that

$$\begin{aligned} E(\hat{\theta}_{K2}) &= \theta, \\ E(\hat{\theta}_{K3}) &= \theta, \\ V(\hat{\theta}_{K2}) &= E_P\left[\frac{1}{\nu^2}\sum_{i\in u}\frac{ay_i+b}{f_{is}}\right] + \frac{N\theta(1-\theta)}{N-1}\left[E_P(\frac{1}{\nu}) - \frac{1}{N}\right], \\ V(\hat{\theta}_{K3}) &= \frac{1}{N^2\pi_i^2}E_P\left[\sum_{i\in u}\frac{ay_i+b}{f_{is}}\right] + \frac{1}{N^2\pi_i^2}\left[\theta^2\left(C_2 - \frac{C_1N}{N-1}\right) + \frac{\theta NC_1}{N-1}\right]. \end{aligned} \tag{5.4.10}$$

5.4.2 Efficiency comparisons among $\hat{\theta}_K, \hat{\theta}_{K1}, \hat{\theta}_{KHT}, \hat{\theta}_{K2}, \hat{\theta}_{K3}$

(i) Comparing $\hat{\theta}_{K2}$ with $\hat{\theta}_{K1}$ and $\hat{\theta}_{K3}$ with $\hat{\theta}_{KHT}$: Following the arguments in (5.3.1) we note that for this model

$$(a\theta+b)E_P\left(\frac{1}{\nu}\right) \geq E_P\left[\sum_{i\in u}\frac{ay_i+b}{f_{is}}\frac{1}{\nu^2}\right] \geq \frac{a\theta+b}{n}E_P\left(\frac{1}{\nu}\right) \tag{5.4.11}$$

and

$$N\pi_i(a\theta+b) \geq E_P\left[\sum_{i\in u}\frac{ay_i+b}{f_{is}}\right] \geq \frac{N\pi_i(a\theta+b)}{n}. \tag{5.4.12}$$

Consequently, as for Warner's model, here we have

$$\hat{\theta}_{K2} \succ \hat{\theta}_{K1} \text{ and } \hat{\theta}_{K3} \succ \hat{\theta}_{KHT} \text{ uniformly.}$$

So, we now compare $\hat{\theta}_K, \hat{\theta}_{K2}$ and $\hat{\theta}_{K3}$.

(ii) Comparing $\hat{\theta}_K$ with $\hat{\theta}_{K2}$: Using (5.4.11) and (5.3.14) we can show that

$\hat{\theta}_{K2} \succ \hat{\theta}_K$ if

$$\theta^2[N-1-NnQ+n]-\theta[N-1-NnQ+n-a(nQ-1)(N-1)]+b(nQ-1)(N-1)\leq 0$$

and $\hat{\theta}_K \succ \hat{\theta}_{K2}$ if

$$\theta^2\left[N-1-NnQ+n\frac{721}{720}\right]+\theta\left[(N-1)\left\{a\left(Q-\frac{1}{720N}-1\right)-1\right\}+NnQ-\frac{721n}{720}\right]$$
$$+b(N-1)(Q-1-\frac{1}{720N}) \geq 0.$$

(iii) Comparing $\hat{\theta}_K$ with $\hat{\theta}_{K3}$: Again using (5.4.12) we have that $\hat{\theta}_{K3} \succ \hat{\theta}_K$ if

$$\theta^2\left[\left(C_2 - \frac{C_1N}{N-1}\right)n + N^2\pi_i^2\right]$$

$$-\theta\left[N^2\pi_i^2(1+a)-n\left(N\pi_i a+\frac{C_1N}{N-1}\right)\right]+bN\pi_i(n-N\pi_i)\leq 0$$

and $\hat{\theta}_K \succ \hat{\theta}_{K3}$ if

$$\theta^2\left[\left(C_2-\frac{C_1N}{N-1}\right)n+N^2\pi_i^2\right]+\theta\left[\frac{C_1Nn}{N-1}-N^2\pi_i^2(1+a)+N\pi_i a\right]$$
$$-bN\pi_i(N\pi_i-1)\geq 0.$$

(iv) Comparing $\hat{\theta}_{K2}$ with $\hat{\theta}_{K3}$: Applying (5.4.11), (5.4.12) and (5.3.14), we may deduce that

$\hat{\theta}_{K2} \succ \hat{\theta}_{K3}$ if

$$\theta^2\left[\frac{C_2-\frac{C_1N}{N-1}}{N^2\pi_i^2}+\frac{NQ-1}{N-1}\right]$$
$$+\theta\left[\frac{a}{N\pi_i n}+\frac{C_1}{N\pi_i^2(N-1)}-aQ-\frac{NQ-1}{N-1}\right]+b\left(\frac{1}{N\pi_i n}-Q\right)\geq 0$$

and $\hat{\theta}_{K3} \succ \hat{\theta}_{K2}$ if

$$\theta^2\left[\frac{C_2-\frac{C_1N}{N-1}}{N^2\pi_i^2}+\frac{NQ-\frac{721}{720}}{N-1}\right]$$
$$+\theta\left[\frac{a}{N\pi_i}+\frac{C_1}{N\pi_i^2(N-1)}-\frac{a}{n}\left(Q-\frac{1}{720N}\right)-\frac{NQ-\frac{721}{720}}{N-1}\right]$$
$$+b\left(\frac{1}{N\pi_i}-\frac{Q-\frac{1}{720N}}{n}\right)\leq 0.$$

We now present the ranges of θ for some illustrative values of N, n, p_1, p_2 and L showing the gain in efficiencies among the alternative estimators considered in this section in Table 2.

Table 2: Some illustrations showing efficiency comparisons of the alternative estimators under Kuk's model

Comparison	N	n	p_1	p_2	L	θ
$\hat{\theta}_{K2} \succ \hat{\theta}_{K1}$	any	any	any	any	any	any
$\hat{\theta}_{K3} \succ \hat{\theta}_{KHT}$	any	any	any	any	any	any
$\hat{\theta}_{K2} \succ \hat{\theta}_{K}$	100	10	0.45	0.2	20	$0.247 \leq \theta \leq 0.663$
	100	20	0.45	0.2	20	$0.214 \leq \theta \leq 0.704$
	200	30	0.45	0.2	20	$0.198 \leq \theta \leq 0.723$
	100	10	0.75	0.2	20	$0.035 \leq \theta \leq 0.959$
$\hat{\theta}_{K3} \succ \hat{\theta}_{K}$	100	10	0.75	0.2	20	$0.029 \leq \theta \leq 0.483$
	200	30	0.75	0.2	20	$0.029 \leq \theta \leq 0.488$
	200	30	0.85	0.1	20	$0.008 \leq \theta \leq 0.509$
$\hat{\theta}_{K2} \succ \hat{\theta}_{K3}$	100	10	0.75	0.2	20	$\theta \geq 0.830$
	200	30	0.75	0.2	20	$\theta \geq 0.690$
	100	20	0.75	0.2	20	$\theta \geq 0.618$
	200	25	0.75	0.2	20	$\theta \geq 0.747$

5.4.3 Unbiased variance estimators

The following unbiased variance estimators under Kuk's model can be obtained using arguments similar to those used in Section 5.3.3 under Warner's model. We give the expressions below.

1. An unbiased variance estimator of $\hat{\theta}_K$ is

$$v(\hat{\theta}_K) = \frac{1}{n(n-1)} \sum_{k=1}^{n} (\rho_k - \bar{\rho}(n))^2, \text{ where } \bar{\rho}(n) = \frac{1}{n} \sum_{k=1}^{n} \rho_k$$

2. $V(\hat{\theta}_{K1})$ has the following two unbiased estimators

$$v_1(\hat{\theta}_{K1}) = (a\hat{\theta}_{K1} + b) E_P\left(\frac{1}{\nu}\right) + \left[N E_P\left(\frac{1}{\nu}\right) - 1\right] \frac{\hat{\theta}_{K1} - (\widehat{\theta^2})_K}{N-1}$$

and

$$v_2(\hat{\theta}_{K1}) = a\hat{\theta}_{K1} E_P\left(\frac{1}{\nu}\right) + b(\frac{1}{\nu}) + \left[N E_P\left(\frac{1}{\nu}\right) - 1\right] \frac{\hat{\theta}_{K1} - (\widehat{\theta^2})_K}{N-1},$$

where

$$(\widehat{\theta^2})_K = \frac{1}{C_2 + N^2\pi_i^2 - C_1 \frac{N}{N-1}} \left[\sum \sum_{i \neq j, i,j \in u} \rho_i \rho_j - \hat{\theta}_{K1} (C_1 \frac{N}{N-1} - N\pi_i) \right]$$

with C_1 and C_2 as defined in (5.3.22) and a, b are as used in (5.4.1).

3. An unbiased estimator of $V(\hat{\theta}_{KHT})$ is

$$v(\hat{\theta}_{KHT}) = \hat{\theta}_{KHT}\left[\frac{a}{N\pi_i}+\frac{C_1}{N\pi_i^2(N-1)}\right]+(\widetilde{\theta^2})_K\left[\frac{C_2}{N^2\pi_i^2}-\frac{C_1}{N\pi_i^2(N-1)}\right]+\frac{b}{N\pi_i}$$

where

$$(\widetilde{\theta^2})_K = \frac{1}{C_2 + N^2\pi_i^2 - C_1\frac{N}{N-1}}\left[\sum\sum_{i\neq j, i,j\in u}\rho_i\rho_j - \hat{\theta}_{KHT}(C_1\frac{N}{N-1} - N\pi_i)\right].$$

4. An unbiased estimator of $V(\hat{\theta}_{K2})$ is

$$v(\hat{\theta}_{K2}) = \left[\frac{1}{N^{n-1}(N-1)}\sum_{l=1}^{N-1} l^{n-1}\right](\hat{\theta}_{K2} - (\widetilde{\widetilde{\theta^2}})_K) + \frac{1}{\nu^2}\sum_{i\in u}\frac{a\gamma_i + b}{f_{is}},$$

where

$$(\widetilde{\widetilde{\theta^2}})_K = \frac{1}{C_2 + N^2\pi_i^2 - C_1\frac{N}{N-1}}\left[\sum\sum_{i\neq j, i,j\in u}\gamma_i\gamma_j - \hat{\theta}_{K2}(C_1\frac{N}{N-1} - N\pi_i)\right].$$

5. An unbiased estimator of $V(\hat{\theta}_{K3})$ is

$$v(\hat{\theta}_{K3}) = f_4(\nu) + \frac{1}{\pi_i N^2}\sum_{i\in u}\frac{a\gamma_i + b}{f_{is}},$$

where

$$f_4(\nu) = \frac{1}{N^2}\left[\sum_{i\in u}\gamma_i^2(\frac{1-\pi_i}{\pi_i^2}) + \sum\sum_{i\neq j, i,j\in u}\gamma_i\gamma_j(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}\pi_i\pi_j})\right].$$

5.5 Estimators based on independently repeated RR's by Christofides's (2003) device in SRSWR with n draws

Christofides (2003) proposed a randomization device and gave an estimator for θ based on RR's generated from this device. In this section we study some possible alternative unbiased estimators for θ using these RR's.

Christofides's (2003) device consists of a box filled with a number of cards marked $1, 2, \ldots, M$, in proportions $p_1, p_2, \ldots, p_M (0 < p_j < 1, \sum_{j=1}^{M} p_j = 1)$. The respondent selected at the k_{th} draw of an SRSWR is instructed to

draw a card from the box. Suppose the number on the card drawn is α_k, $\alpha_k = 1, \ldots, M$. Then, the respondent is instructed to report the number α_k if he does not bear the sensitive attribute A and to report the number $(M + 1 - \alpha_k)$ otherwise. The interviewer does not witness the card number nor the process involved in translating this number into the number reported; he only records the final number reported. Thus, the privacy of the respondent is protected.

Let y_k be the y-value (=1 or 0 according as bearing or non-bearing A) of the respondent selected at k_{th} draw. Let z_k be the final number reported by him, i.e., the RR generated by the device. Then,

$$z_k = (M + 1 - \alpha_k)y_k + \alpha_k(1 - y_k),\ k = 1, \ldots, n.$$

Therefore, writing $\mu = \sum_{\alpha_k=1}^{M} \alpha_k p_{\alpha_k}$, it is clear that

$$\begin{aligned}
E_R(z_k) &= E_R[(M + 1 - \alpha_k)y_k + \alpha_k(1 - y_k)] \\
&= (M + 1)y_k + (1 - 2y_k) \textstyle\sum_{\alpha_k=1}^{M} \alpha_k p_{\alpha_k} \\
&= (M + 1)y_k + (1 - 2y_k)\mu \\
&= (M + 1 - 2\mu)y_k + \mu, \text{ and} \\
V_R(z_k) &= V_R[(1 - 2y_k)\alpha_k] \\
&= (1 - 2y_k)^2[\textstyle\sum_{\alpha_k=1}^{M} \alpha_k^2 p_{\alpha_k} - \mu^2] \\
&= \textstyle\sum_{\alpha_k=1}^{M} \alpha_k^2 p_{\alpha_k} - \mu^2,
\end{aligned}$$

since $y_k^2 = y_k$. Let

$$c_k = \frac{z_k - \mu}{M + 1 - 2\mu}.$$

Then it follows that

$$E_R(c_k) = y_k \text{ and } V_R(c_k) = \frac{\sum_{\alpha_k=1}^{M} \alpha_k^2 p_{\alpha_k} - \mu^2}{(M + 1 - 2\mu)^2} = \Phi_C, \text{ say.}$$

The unbiased estimator for θ given by Christofides (2003) may be expressed in terms of the c_k values as

$$\hat{\theta}_C = \frac{1}{n} \sum_{k=1}^{n} c_k \quad \text{with} \quad V(\hat{\theta}_C) = \frac{\theta(1 - \theta)}{n} + \frac{\Phi_C}{n}.$$

5.5.1 Some alternative estimators

Now, analogous to the estimators proposed in Sections 5.3 and 5.4, we propose several alternative unbiased estimators for θ using the RR's obtained from Christofides's (2003) RRD.

Suppose only a single RR is obtained from every distinct respondent selected in an SRSWR. Let the number drawn from the box by respondent i be α_i, $\alpha_i = 1, \ldots, M$. Then the RR z_i given by respondent i can be written as

$$z_i = (M + 1 - \alpha_i)y_i + \alpha_i(1 - y_i), \ \forall i \in u.$$

with

$$E_R(z_i) = (M + 1 - 2\mu)y_i + \mu \quad \text{and} \quad V_R(z_i) = \sum_{\alpha_i=1}^{M} \alpha_i^2 p_{\alpha_i} - \mu^2.$$

Hence we have

$$c_i = \frac{z_i - \mu}{M + 1 - 2\mu} \text{ with } E_R(c_i) = y_i \text{ and } V_R(c_i) = \frac{\sum_{\alpha_i=1}^{M} \alpha_i^2 p_{\alpha_i} - \mu^2}{(M + 1 - 2\mu)^2} = \Phi_C.$$

This leads to the following two unbiased estimators for θ based on these c_i's.

$$\text{(i)} \quad \hat{\theta}_{C1} = \frac{1}{\nu} \sum_{i \in u} c_i \quad \text{and} \quad \text{(ii)} \quad \hat{\theta}_{CHT} = \frac{1}{N} \sum_{i \in u} \frac{c_i}{\pi_i}.$$

Using (5.3.7), the variances of the above estimators can be obtained as

$$\begin{aligned}
V(\hat{\theta}_{C1}) &= \Phi_C E_P\left(\tfrac{1}{\nu}\right) + \left[N E_P\left(\tfrac{1}{\nu}\right) - 1\right] \tfrac{\theta(1-\theta)}{N-1}, \text{ and} \\
V(\hat{\theta}_{CHT}) &= \tfrac{\Phi_C}{N\pi_i} + \tfrac{1}{N^2\pi_i^2} V_P(\nu \bar{y}(\nu)) \\
&= \tfrac{\Phi_C}{N\pi_i} \\
&\quad + \tfrac{\theta(1-\theta)}{N(N-1)\pi_i^2}\left[(N-1)\left\{(1-\tfrac{1}{N})^n - (1-\tfrac{2}{N})^n\right\}\right] \\
&\quad + \tfrac{\theta^2}{N^2\pi_i^2}\left[N(1-\tfrac{1}{N})^n - N^2(1-\tfrac{1}{N})^{2n} + N(N-1)(1-\tfrac{2}{N})^n\right].
\end{aligned}$$

We now construct the estimators based on independently repeated RR's in an SRSWR with n draws, by using the frequencies f_{is} as before. Let α_{ij} denote the card number drawn by the respondent i at his j_{th} appearance. $\alpha_{ij} = 1, \ldots, M$; and let z_{ij} be the corresponding RR reported. Then we may write

$$z_{ij} = (M + 1 - \alpha_{ij})y_i + \alpha_{ij}(1 - y_i), \ \forall i \in u,$$

with

$$E_R(z_{ij}) = (M + 1 - 2\mu)y_i + \mu \quad \text{and} \quad V_R(z_{ij}) = \sum_{\alpha_{ij}=1}^{M} \alpha_{ij}^2 p_{\alpha_{ij}} - \mu^2.$$

Now, let

$$\bar{z}_i = \frac{1}{f_{is}} \sum_{j=1}^{f_{is}} z_{ij} \quad \text{and} \quad d_i = \frac{\bar{z}_i - \mu}{M + 1 - 2\mu}.$$

Then,

$$E_R(d_i) = y_i \quad \text{and} \quad V_R(d_i) = \frac{V_R(z_{ij})}{f_{is}(M+1-2\mu)^2} = \frac{\Phi_C}{f_{is}}.$$

These d_i's lead to the following two new unbiased estimators for θ.

$$\text{(i)} \quad \hat{\theta}_{C2} = \frac{1}{\nu}\sum_{i\in u} d_i \quad \text{and} \quad \text{(ii)} \quad \hat{\theta}_{C3} = \frac{1}{N}\sum_{i\in u}\frac{d_i}{\pi_i}.$$

It is interesting to see that for this device, the variance of the transformed RR (i.e., c_k) is a constant, like in the case for Warner (see (1.2.19)). This simplifies the algebra for variance computations considerably and the variances $V(\hat{\theta}_{C2})$ and $V(\hat{\theta}_{C3})$ can be obtained mimicking the steps for obtaining $V(\hat{\theta}_{W2})$ and $V(\hat{\theta}_{W3})$ shown in (5.3.8) and (5.3.12), respectively. These variances are shown below.

$$\begin{aligned}
V(\hat{\theta}_{C2}) &= \left[\tfrac{1}{N^{n-1}(N-1)}\sum_{l=1}^{N-1} l^{n-1}\right](\theta-\theta^2) + \Phi_C E_P\left(\tfrac{1}{\nu^2}\sum_{i\in u}\frac{1}{f_{is}}\right) \\
V(\hat{\theta}_{C3}) &= \tfrac{\Phi_C}{N^2\pi_i^2} E_P\left[\sum_{i\in u}\frac{1}{f_{is}}\right] \\
&\quad + \tfrac{\theta(1-\theta)}{N(N-1)\pi_i^2}\left[(N-1)\left\{\left(1-\tfrac{1}{N}\right)^n - \left(1-\tfrac{2}{N}\right)^n\right\}\right] \\
&\quad + \tfrac{\theta^2}{N^2\pi_i^2}\left[N\left(1-\tfrac{1}{N}\right)^n - N^2\left(1-\tfrac{1}{N}\right)^{2n} + N(N-1)\left(1-\tfrac{2}{N}\right)^n\right]
\end{aligned}$$

5.5.2 Efficiency comparisons among $\hat{\theta}_C$, $\hat{\theta}_{C1}$, $\hat{\theta}_{CHT}$, $\hat{\theta}_{C2}$, $\hat{\theta}_{C3}$

(i) Comparing $\hat{\theta}_C$ with $\hat{\theta}_{C1}$: Following (5.3.1), (5.3.13) and (5.3.14) we may study the conditions under which one outperforms the other. We conclude that

$$\hat{\theta}_{C1} \succ \hat{\theta}_C \text{ if } \theta(1-\theta) > \frac{n(N-1)(6N+n-1)}{N\{6Nn-12N-n(n-1)\}}\Phi_C$$

and

$$\hat{\theta}_C \succ \hat{\theta}_{C1} \text{ if } \theta(1-\theta) \le \frac{(N-1)\left(NnQ-\frac{n}{720}-N\right)}{N\left(N-1-NnQ+\frac{721n}{720}\right)}\Phi_C.$$

(ii) Comparing $\hat{\theta}_{C1}$ with $\hat{\theta}_{C2}$: From (5.3.1) and (5.3.13) it is clear that $V\left(\hat{\theta}_{C2}\right) \le V\left(\hat{\theta}_{C1}\right)$ and so

$$\hat{\theta}_{C2} \succ \hat{\theta}_{C1} \text{ uniformly.}$$

(iii) Comparing $\hat{\theta}_{CHT}$ and $\hat{\theta}_{C3}$: From (5.3.1) and (5.3.13) it is again clear that $V\left(\hat{\theta}_{C3}\right) \leq V\left(\hat{\theta}_{CHT}\right)$ and so

$$\hat{\theta}_{C3} \succ \hat{\theta}_{CHT} \text{ uniformly.}$$

In view of (ii) and (iii) above, we now compare $\hat{\theta}_C$, $\hat{\theta}_{C2}$ and $\hat{\theta}_{C3}$.

(iv) Comparing $\hat{\theta}_C$ with $\hat{\theta}_{C2}$: From (5.3.1), (5.3.13) and (5.3.14) we may conclude that

$$\hat{\theta}_{C2} \succ \hat{\theta}_C \text{ if } \theta(1-\theta) \geq \Phi_C(N-1)\left[\frac{nQ-1}{N+n-1-NnQ}\right]$$

and

$$\hat{\theta}_C \succ \hat{\theta}_{C2} \text{ if } \theta(1-\theta) \leq \Phi_C(N-1)\left[\frac{Q-\frac{1}{720N}-1}{N+\frac{721n}{720}-1-NnQ}\right].$$

(v) Comparing $\hat{\theta}_C$ with $\hat{\theta}_{C3}$: From (5.3.1), (5.3.13) and (5.3.10) it follows that

$$\hat{\theta}_{C3} \succ \hat{\theta}_C \text{ if}$$

$$\theta^2\left[A_2+\frac{NA_1}{N-1}-\frac{1}{N-1}\sum_{l=1}^{N-1}\left(\frac{l}{N}\right)^{n-1}+\frac{1}{n}\right]$$

$$+\theta\left[\frac{1}{N-1}\sum_{l=1}^{N-1}\left(\frac{l}{N}\right)^{n-1}-\frac{NA_1}{N-1}-\frac{1}{n}\right]+\Phi_C\left[\frac{n-N\pi_i}{nN\pi_i}\right]\leq 0,$$

and

$$\hat{\theta}_C \succ \hat{\theta}_{C3} \text{ if}$$

$$\theta^2\left[A_2+\frac{NA_1}{N-1}-\frac{1}{N-1}\sum_{l=1}^{N-1}\left(\frac{l}{N}\right)^{n-1}+\frac{1}{n}\right]$$

$$+\theta\left[\frac{1}{N-1}\sum_{l=1}^{N-1}\left(\frac{l}{N}\right)^{n-1}-\frac{NA_1}{N-1}-\frac{1}{n}\right]+\frac{\Phi_C}{n}\left[\frac{1-N\pi_i}{N\pi_i}\right]\geq 0.$$

(vi) Comparing $\hat{\theta}_{C2}$ with $\hat{\theta}_{C3}$: From (5.3.13), (5.3.14) and (5.3.10) it can be seen that

$$\hat{\theta}_{C2} \succ \hat{\theta}_{C3} \text{ if } \frac{N\theta(1-\theta)A_1}{N-1}-\theta^2A_2 \leq \Phi_C\left(\frac{1}{nN\pi_i}-Q\right)$$

and

$$\hat{\theta}_{C3} \succ \hat{\theta}_{C2} \text{ if } \frac{N\theta(1-\theta)A_1}{N-1}-\theta^2A_2 \geq \Phi_C\left(\frac{1}{N\pi_i}-\frac{Q}{n}\right).$$

5.5.3 Unbiased variance estimators

1. An unbiased estimator for $V(\hat{\theta}_C)$ can be obtained proceeding as in the case of Warner's model. This is equal to

$$\hat{V}(\hat{\theta}_C) = v(\hat{\theta}_C) = \frac{1}{n(n-1)} \sum_{k=1}^{n} (c_k - \hat{\theta}_C)^2.$$

2. The unbiased estimator for $V(\hat{\theta}_{C1})$ can be obtained easily by replacing (i) c_i, c_j's in place of r_i, r_j's, (ii) replacing $\hat{\theta}_{C1}$ in place of $\hat{\theta}_{W1}$ and (iii) replacing Φ_C in place of Φ_W in the expressions for $v_1(\hat{\theta}_{W1})$ and $v_2(\hat{\theta}_{W1})$ in Warner's model section.

3. Unbiased estimators of $V(\hat{\theta}_{CHT})$ can be obtained as follows: Let

$$f_5(\nu) = \frac{1}{N^2} \left[\sum_{i \in u} c_i^2 \left(\frac{1-\pi_i}{\pi_i^2} \right) + \sum \sum_{i \neq i', i, i' \in u} c_i c_{i'} \left(\frac{\pi_{ii'} - \pi_i \pi_{i'}}{\pi_{ii'} \pi_i \pi_{i'}} \right) \right].$$

Then

$$\begin{aligned}
E(f_5(\nu)) &= E_P E_R(f_5(\nu)) \\
&= \tfrac{1}{N^2} \left[\textstyle\sum_{i=1}^{N} y_i^2 (\frac{1-\pi_i}{\pi_i}) + \Phi_C E_P \left(\sum_{i \in u} (\frac{1-\pi_i}{\pi_i^2}) \right) + \sum \sum_{i \neq i'}^{N} y_i y_{i'} (\frac{\pi_{ii'} - \pi_i \pi_{i'}}{\pi_i \pi_{i'}}) \right] \\
&= V_P (\tfrac{1}{N} \textstyle\sum_{i \in u} \frac{y_i}{\pi_i}) + \frac{\Phi_C}{N^2} E_P \left(\sum_{i \in u} (\frac{1-\pi_i}{\pi_i^2}) \right) \\
&= V_P \left[E_R(\hat{\theta}_{CHT}) \right] + \frac{\Phi_C}{N^2} \left(\frac{1-\pi_i}{\pi_i^2} \right) E_P(\nu) \\
&= V_P \left[E_R(\hat{\theta}_{CHT}) \right] + \frac{\Phi_C}{N^2} \left(\frac{1-\pi_i}{\pi_i^2} \right) N \pi_i \\
&= V(\hat{\theta}_{CHT}) - E_P V_R(\hat{\theta}_{CHT}) + \frac{\Phi_C}{N} \left(\frac{1-\pi_i}{\pi_i} \right).
\end{aligned}$$

This implies that

$$\begin{aligned}
V(\hat{\theta}_{CHT}) &= E_P E_R(f_5(\nu)) - \frac{\Phi_C}{N} \left(\frac{1-\pi_i}{\pi_i} \right) + E_P V_R(\hat{\theta}_{CHT}) \\
&= E_P E_R(f_5(\nu)) - \frac{\Phi_C}{N} \left(\frac{1-\pi_i}{\pi_i} \right) + E_P \left[\frac{1}{N^2} \sum_{i \in u} \frac{\Phi_C}{\pi_i^2} \right] \\
&= E_P E_R(f_5(\nu)) - \frac{\Phi_C}{N} \left(\frac{1-\pi_i}{\pi_i} \right) + \frac{\Phi_C}{N^2 \pi_i^2} E_P(\nu) \\
&= E_P E_R(f_5(\nu)) - \frac{\Phi_C}{N} \left(\frac{1-\pi_i}{\pi_i} \right) + \frac{\Phi_C}{N \pi_i} \\
&= E_P E_R \left(f_5(\nu) + \frac{\Phi_C}{N} \right)
\end{aligned}$$

Hence an unbiased estimator for $V(\hat{\theta}_{CHT})$ is given by

$$\hat{V}(\hat{\theta}_{CHT}) = v(\hat{\theta}_{CHT}) = f_5(\nu) + \frac{\Phi_C}{N}.$$

The unbiased estimators of $V(\hat{\theta}_{C2})$ and $V(\hat{\theta}_{C3})$ can be obtained as in $v(\hat{\theta}_{W2})$ and $v(\hat{\theta}_{W3})$ under Warner's model (shown in Section 5.3.3), by writing Φ_C in place of Φ_W, d_i's in place of g_i's, $\hat{\theta}_{C2}$ in place of $\hat{\theta}_{W2}$ and $\hat{\theta}_{C3}$ in place of $\hat{\theta}_{W3}$, respectively. These variance estimators are shown below.

4. Let

$$(\widetilde{\theta^2})_C = \frac{1}{C_2 + N^2\pi_i^2 - C_1\frac{N}{N-1}}\left[\sum\sum_{i\neq j, i,j\in u} d_i d_j - \hat{\theta}_{C2}(C_1\frac{N}{N-1} - N\pi_i)\right]$$

$$\hat{V}(\hat{\theta}_{C2}) = \left[\frac{1}{N^{n-1}(N-1)}\sum_{l=1}^{N-1} l^{n-1}\right](\hat{\theta}_{C2} - (\widetilde{\theta^2})_C) + \Phi_C(\frac{1}{\nu^2}\sum_{i\in u}\frac{1}{f_{is}})$$

Other alternative variance estimators can be easily obtained by following the same procedure as in the variance estimators $v_i(\hat{\theta}_{W2}), i = 2, 3, 4, 5$ for Warner's model section.

5. Let

$$f_6(\nu) = \frac{1}{N^2}\left[\sum_{i\in u} d_i^2(\frac{1-\pi_i}{\pi_i^2}) + \sum\sum_{i\neq i', i,i'\in u} d_i d_{i'}(\frac{\pi_{ii'} - \pi_i\pi_{i'}}{\pi_{ii'}\pi_i\pi_{i'}})\right].$$

An unbiased estimator for $V(\hat{\theta}_{C3})$ is given by

$$\hat{V}(\hat{\theta}_{C3}) = f_6(\nu) + \frac{\Phi_C}{\pi_i N^2}\sum_{i\in u}\frac{1}{f_{is}}.$$

Chapter 6

Estimation of a sensitive proportion using randomized response data obtained through inverse sampling

Abstract

In Chapter 5 we proposed some estimators for estimating the proportion of people bearing a sensitive characteristic in a population, using multiple randomized responses from distinct persons sampled. The sampling of respondents was done by SRSWR with a pre-fixed number of draws. In this chapter, we undertake a similar study, but this time, for sampling of respondents, we employ inverse sampling with equal probabilities with replacement. We propose some estimators and show certain advantages in estimation using randomized response data by Warner's (1965) device gathered through such a simple inverse sampling scheme. We also study this inverse sampling problem for Kuk's (1990) and Christofides's (2003) device.

6.1 Introduction

We continue with our study of randomized response techniques as in Chapter 5. Most of the results in this area are based on a sample with a preassigned sample size, say, n and traditionally, a simple random sample with replacement (SRSWR) is taken. In this chapter, unlike the usual practice

of taking an SRSWR sample of size n, we adopt an alternative inverse sampling plan. Here we pre-fix a number, say ν, of distinct persons we intend to cover but permit a random number of $n(\geq \nu)$ draws with equal probabilities with replacement that we may need to realize this.

In the context of direct response surveys, it is well known from Basu (1958), Chikkagoudar (1966) and Lanke (1975) that for estimating a population mean, if one uses the responses obtained from the distinct units sampled by *such an inverse sampling method*, then estimators performing better than the usual sample mean are available. It seems natural to investigate if such a result also holds for randomized response surveys. So, in this chapter, we study this problem and propose some alternative estimators for the population total based on randomized responses collected from each person selected by this inverse sampling. We may recall that a similar study with a sample of pre-fixed size, was considered in Chapter 5.

We first consider randomized responses (RR's) gathered by using Warner's (1965) randomized response device in Section 6.2. In this section we present our proposed estimators based on these RR data along with their variances. We compare the efficiencies of our proposed estimators together with some numerical illustrations. In this section, we also derive the unbiased estimators for the variances of the estimators considered in this chapter. Finally, in Sections 6.3 and 6.4, we extend this study to Kuk's (1990) and Christofides's (2003) randomization devices.

6.2 Estimation using Warner's (1965) RR device in Simple Inverse Sampling with replacement

We consider the situation where RR's are generated by Warner's (1965) technique from every person each time found to be selected when samples are chosen with equal probabilities with replacement, drawings being independently continued till a pre-assigned number $\nu(> 1)$ of distinct persons are selected.

In a sample s drawn as described above, suppose the 1^{st}, $2^{nd}, \ldots, (\nu-1)^{th}$ distinct person appears f_{1s}, $f_{2s}, \ldots, f_{(\nu-1)s}$ times, respectively, the $f_i's$ being random. Clearly, the ν^{th} distinct person appears only once. Then, denoting

the total sample size as n, we have

$$\sum_{i=1}^{\nu-1} f_{is} = (n-1) \tag{6.2.1}$$

and n here is a random variable.

Let $\sum'$ denote the sum over all possible choices of $f_{is}(i = 1, 2, \ldots, (\nu - 1))$ subject to (6.2.1). From Raj and Khamis (1958) we know that the probability that one needs $n(\geq \nu)$ draws to obtain $\nu(> 1)$ distinct persons is

$$P(n) = \frac{N\binom{N-1}{\nu-1}}{N^n} \sum{}' \frac{(n-1)!}{f_{1s}! \ldots f_{(\nu-1)s}!} = \frac{\binom{N-1}{\nu-1}}{N^{n-1}} \left[\Delta^{\nu-1} x^{n-1}|_{x=0}\right] \tag{6.2.2}$$

Here Δ is the difference operator such that $\Delta f(x) = f(x+1) - f(x)$ and we write $\Delta^n = \Delta \circ \Delta \circ \ldots \circ \Delta$(n times). Using this we get

$$\sum{}' \frac{(n-1)!}{f_{1s}! \ldots f_{(\nu-1)s}!} = (\nu-1)^{n-1} - \binom{\nu-1}{1}(\nu-2)^{n-1} + \ldots + (-1)^{\nu-2}\binom{\nu-1}{\nu-2}$$

$$= \left[\Delta^{\nu-1} x^{n-1}|_{x=0}\right].$$

Writing $E_{P|n}$, $V_{P|n}$ as the conditional expectation and variance operators subject to a given value of n and E_n, V_n as the corresponding operators over the probability distribution of n with respect to the above inverse sampling, we have

$$E_P = E_n E_{P|n}, \quad V_P = E_n V_{P|n} + V_n E_{P|n}.$$

6.2.1 Some alternative estimators

(i) Consider the classical estimator

$$e_W = \frac{1}{n} \sum_{k=1}^{n} r_k = \bar{r}(n), \tag{6.2.3}$$

where r_k is as in (1.2.19). This e_W is analogous to Warner's estimator $\hat{\theta}_W$ given in (1.2.20). It follows that:

$$E(e_W) = E_n E_{P|n}\left[E_R(e_W)\right] = \frac{1}{N}\sum_{i=1}^{N} y_i = \theta,$$

$$V(e_W) = E_n\left[V_{P|n}E_R(e_W) + E_{P|n}V_R(e_W)\right] = E_n\left[V_{P|n}(\bar{y}(n))\right] + \Phi_W E_n(\frac{1}{n}). \tag{6.2.4}$$

From (6.2.2) and Chikkagoudar (1966) we know that

$$E_n\left(\frac{1}{n}\right) = \sum_{n=\nu}^{\infty} \frac{1}{n} P(n) = \binom{N-1}{\nu-1} \sum_{n=\nu}^{\infty} \frac{N^{1-n}}{n} \left[\Delta^{\nu-1} x^{n-1}|_{x=0}\right], \qquad (6.2.5)$$

$$V_{P|n}(\bar{y}(n)) = \left[\left(\frac{N-n}{Nn}\right) + \frac{(n-1)(n-2)}{n^2} \frac{[\Delta^{\nu-1} x^{n-2}|_{x=0}]}{[\Delta^{\nu-1} x^{n-1}|_{x=0}]}\right] S^2, \qquad (6.2.6)$$

$$E_n V_{P|n}(\bar{y}(n)) =$$
$$S^2 \binom{N-1}{\nu-1} \left[\Delta^{\nu-1} \left\{\frac{1}{x} + \frac{N(x-3)}{x^2} log(\frac{N}{N-x}) + \frac{2}{x} \sum_{n=\nu}^{\infty} \frac{1}{n^2} \left(\frac{x}{N}\right)^{n-1}\right\} |_{x=0}\right]. \qquad (6.2.7)$$

Lanke (1975), however has shown that if we write $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2 = \theta(1-\theta)$, then

$$E_n V_{P|n}(\bar{y}(n)) = \frac{\sigma^2}{N-1} \left[N E_n\left(\frac{1}{n}\right) - E_n\left(\frac{3n+1}{(n+1)^2}\right)\right]. \qquad (6.2.8)$$

Using (6.2.4)–(6.2.8), we may deduce two variance formulae $V_1(e_W)$ and $V_2(e_W)$, say, for e_W as shown below.

$$V_1(e_W) = \frac{N}{N-1}\theta(1-\theta)\binom{N-1}{\nu-1} \times$$
$$\left[\Delta^{\nu-1} \left\{\frac{1}{x} + \frac{N(x-3)}{x^2} log(\frac{N}{N-x}) + \frac{2}{x} \sum_{n=\nu}^{\infty} \frac{1}{n^2} \left(\frac{x}{N}\right)^{n-1}\right\} |_{x=0}\right] + \Phi_W E_n(\frac{1}{n}),$$

$$V_2(e_W) = \frac{\theta(1-\theta)}{N-1} \left[N E_n\left(\frac{1}{n}\right) - E_n\left(\frac{3n+1}{(n+1)^2}\right)\right] + \Phi_W E_n(\frac{1}{n}). \qquad (6.2.9)$$

In the case of direct response surveys with this inverse sampling scheme, Basu (1958) and Raj and Khamis (1958) compared the performances of the mean $\bar{y}(n)$ based on the sample s of all the n draws and the mean $\bar{y}(\nu)$ based on only ν distinct units in the sample, denoted by the set u, say. In RR surveys we study a similar problem. We consider some alternative estimators based on units in u as described below and compare them with e_W.

(ii) Let

$$e_{W1} = \frac{1}{\nu} \sum_{i \in u} r_i, \qquad (6.2.10)$$

writing r_i as used in (1.2.19), with only a single RR for every person sampled. Thus, e_{W1} is analogous to the estimator $\hat{\theta}_{W1}$ for fixed sample size, considered in (5.2.1).

(iii) Again as in Chapter 5, we now consider estimators based on the average of f_{is} RR's obtained from every distinct person i in sample s. With g_i as in (5.3.3), we consider the estimator

$$e_{W2} = \frac{1}{\nu} \sum_{i \in u} g_i. \tag{6.2.11}$$

This estimator is analogous to $\hat{\theta}_{W2}$ in (5.3.5). Then, it is clear that

$$E(e_{W1}) = \theta, \quad E(e_{W2}) = \theta$$

and so, e_{W1}, e_{W2} are unbiased for θ. Again, as ν is fixed here, one can show that

$$V(e_{W1}) = V_P \left[\frac{1}{\nu} \sum_{i \in u} y_i \right] + E_P \left[\frac{\Phi_W}{\nu} \right] = \left(\frac{1}{\nu} - \frac{1}{N} \right) \frac{N}{N-1} \theta(1-\theta) + \frac{\Phi_W}{\nu}, \tag{6.2.12}$$

$$V(e_{W2}) = V_P \left[\frac{1}{\nu} \sum_{i \in u} y_i \right] + E_P \left[\frac{1}{\nu^2} \sum_{i \in u} \frac{\Phi_W}{f_{is}} \right]$$

$$= \left(\frac{1}{\nu} - \frac{1}{N} \right) \frac{N}{N-1} \theta(1-\theta) + \frac{\Phi_W}{\nu^2} E_n E_{P|n} \left(\sum_{i \in u} \frac{1}{f_{is}} \right) \tag{6.2.13}$$

We may also consider the Horvitz and Thompson (1952) estimators, say, e_{WHT} and e_{W3} based on inverse sampling. These are as follows:

$$e_{WHT} = \frac{1}{N} \sum_{i \in u} \frac{r_i}{\pi_i} = \frac{\nu}{N \pi_i} e_{W1} \quad \text{and} \quad e_{W3} = \frac{1}{N} \sum_{i \in u} \frac{g_i}{\pi_i} = \frac{\nu}{N \pi_i} e_{W2} \tag{6.2.14}$$

where π_i denotes the inclusion probability of unit i in simple inverse sampling scheme; r_i and g_i are as in (1.2.19) and (5.3.3). Here, e_{W3} corresponds to $\hat{\theta}_{W3}$ in (5.3.9) for sampling with fixed sample size. It can be shown that

$$\pi_i = 1 - \sum_{n=\nu}^{\infty} P_i(n), \quad P_i(n) = \frac{(N-1)\binom{N-2}{\nu-1}}{N^n} \left[\Delta^{\nu-1} x^{n-1} |_{x=0} \right]$$

where $P_i(n)$ is the probability that n draws will be required to get ν distinct units except unit i obtained following the arguments as in (6.2.2).

Now,

$$\sum_{n=\nu}^{\infty} P_i(n) =$$

$$\frac{(N-1)\binom{N-2}{\nu-1}}{N\binom{N-1}{\nu-1}}\sum_{n=\nu}^{\infty}\frac{\binom{N-1}{\nu-1}}{N^{n-1}}\left[\Delta^{\nu-1}x^{n-1}|_{x=0}\right]=\sum_{n=\nu}^{\infty}P(n)\frac{(N-1)\binom{N-2}{\nu-1}}{N\binom{N-1}{\nu-1}}=\frac{N-\nu}{N},$$

since $\sum_{n=\nu}^{\infty}P(n)=1$. So,

$$\pi_i=\frac{\nu}{N}\quad\Rightarrow\quad e_{WHT}=e_{W1}\quad\text{and}\quad e_{W3}=e_{W2}.$$

Thus the Horvitz & Thompson (1952) versions do not lead to new estimators in this case.

6.2.2 Efficiency comparisons among e_W, e_{W1} and e_{W2}

From the observations at the end of the previous section, it is clear that we need to compare e_W, e_{W1} and e_{W2}.

(i) Comparing e_{W1} with e_{W2}:

From (6.2.12) and (6.2.13), on applying (5.3.13), we see that $V(e_{W2})\leq V(e_{W1})$. Thus

$$e_{W2}\succ e_{W1}\text{ uniformly.}$$

(ii) Comparing e_W with e_{W2}: From (6.2.9) and (6.2.13) we have

$$\begin{aligned}V(e_{W2})-V_2(e_W)&\\ &=\left(\frac{1}{\nu}-\frac{1}{N}\right)\frac{N\theta(1-\theta)}{N-1}+\frac{\Phi_W}{\nu^2}E_nE_{P|n}\left(\sum_{i\in u}\frac{1}{f_{is}}\right)\\ &\quad-\left[\frac{\theta(1-\theta)}{N-1}\left\{NE_n\left(\frac{1}{n}\right)-E_n\left(\frac{3n+1}{(n+1)^2}\right)\right\}+\Phi_WE_n(\frac{1}{n})\right]\\ &<\left(\frac{1}{\nu}-\frac{1}{N}\right)\frac{N}{N-1}\theta(1-\theta)+\frac{\Phi_W}{\nu}-\frac{\theta(1-\theta)}{N-1}NE_n\left(\frac{1}{n}\right)\\ &\quad+\frac{\theta(1-\theta)}{N-1}\frac{3}{\nu}-\Phi_WE_n(\frac{1}{n}),\end{aligned}$$

by (5.3.13) and since $E_n\left(\frac{3n+1}{(n+1)^2}\right)<\frac{3}{\nu}$.

Lanke (1975) showed that

$$\text{if }\nu\to\infty,\ N\to\infty,\frac{\nu}{N}\to f_0(0<f_0<1),\text{ then }NE_n\left(\frac{1}{n}\right)\to\frac{1}{log\left(\frac{1}{1-f_0}\right)}.\tag{6.2.15}$$

Applying this result, we may claim that, for large N

$$V(e_{W2}) - V_2(e_W) \;<\; \left(\frac{1}{\nu} - \frac{1}{N}\right)\frac{N}{N-1}\theta(1-\theta) + \frac{\Phi_W}{\nu} + \frac{\theta(1-\theta)}{N-1}\frac{3}{\nu} - \left[\frac{\theta(1-\theta)}{N-1} + \frac{\Phi_W}{N}\right]\frac{1}{log\left(\frac{1}{1-f_0}\right)}.$$

After some algebra it follows that $V(e_{W2}) - V_2(e_W) < 0$ if

$$\theta(1-\theta) \geq \frac{N-1}{N}\Phi_W\left[\frac{Nlog\left(\frac{1}{1-f_0}\right) - \nu}{\nu - (N+3-\nu)log\left(\frac{1}{1-f_0}\right)}\right].$$

We give some illustrations in Table 1 below.

Table 1: Some illustrations showing efficiency comparisons of the alternative estimators under inverse sampling with Warner's device.

Comparison	N	ν	f_0	p ($\neq 0.5$)	θ
$e_{W2} \succ e_{W1}$	any	any	—	any	any
$e_{W2} \succ e_W$	100	30	0.3	0.90, 0.10	$0.274 \leq \theta \leq 0.726$
	250	50	0.2	0.90, 0.10	$0.221 \leq \theta \leq 0.779$

6.2.3 Unbiased variance estimators

1. For e_W as in (6.2.3), we may provide unbiased estimators of $V(e_W)$ as follows: From (6.2.4) and (6.2.6), we note that

$$V(e_W) = E_n\left[\left\{\left(\frac{N-n}{Nn}\right) + \frac{(n-1)(n-2)}{n^2}\frac{[\Delta^{\nu-1}x^{n-2}|_{x=0}]}{[\Delta^{\nu-1}x^{n-1}|_{x=0}]}\right\}S^2 + \frac{\Phi_W}{n}\right]. \tag{6.2.16}$$

With r_k as in (1.2.19), let

$$\bar{r}(n) = \frac{1}{n}\sum_{k=1}^{n} r_k = e_W, \quad s_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y}(n))^2.$$

Then

$$\begin{aligned} E_R\left[\tfrac{1}{n(n-1)}\textstyle\sum_{k=1}^{n}(r_k - \bar{r}(n))^2\right] &= \tfrac{1}{n(n-1)}\left[\textstyle\sum_{k=1}^{n}\{V_R(r_k) + E_R^2(r_k)\}\right. \\ &\quad \left. - n\{V_R(\bar{r}(n)) + E_R^2(\bar{r}(n))\}\right] \\ &= \tfrac{1}{n(n-1)}\left[n\Phi_W + \textstyle\sum_{k=1}^{n} y_k^2 - n\frac{\Phi_W}{n} - n(\bar{y}(n))^2\right] \\ &= \tfrac{\Phi_W}{n} + \tfrac{s_n^2}{n} \end{aligned}$$

Hence

$$E_{P|n}E_R\left[\frac{1}{n(n-1)}\sum_{k=1}^{n}(r_k-\bar{r}(n))^2\right] = \frac{\Phi_W}{n}+\frac{1}{n}E_{P|n}(s_n^2)$$
$$= \frac{\Phi_W}{n}+\frac{1}{n}\left[1-\frac{n-2}{n}\frac{\left[\Delta^{\nu-1}x^{n-2}|_{x=0}\right]}{\left[\Delta^{\nu-1}x^{n-1}|_{x=0}\right]}\right]S^2,$$

since from Chikkagoudar (1966), it is known that

$$E_{P|n}(s_n^2)=\left[1-\frac{n-2}{n}\frac{[\Delta^{\nu-1}x^{n-2}|_{x=0}]}{[\Delta^{\nu-1}x^{n-1}|_{x=0}]}\right]S^2.$$

So, for a fixed n, S^2 is unbiasedly estimable by

$$\left[1-\frac{n-2}{n}\frac{[\Delta^{\nu-1}x^{n-2}|_{x=0}]}{[\Delta^{\nu-1}x^{n-1}|_{x=0}]}\right]^{-1}\left[\frac{1}{n-1}\sum_{k=1}^{n}(r_k-\bar{r}(n))^2-\Phi_W\right].$$

Hence from (6.2.16), an unbiased estimator for $V(e_W)$ is given by

$$\hat{V}(e_W)=\frac{\Phi_W}{n}+\left\{\left(\frac{N-n}{Nn}\right)+\frac{(n-1)(n-2)}{n^2}\frac{[\Delta^{\nu-1}x^{n-2}|_{x=0}]}{[\Delta^{\nu-1}x^{n-1}|_{x=0}]}\right\}\times$$

$$\left[1-\frac{n-2}{n}\frac{[\Delta^{\nu-1}x^{n-2}|_{x=0}]}{[\Delta^{\nu-1}x^{n-1}|_{x=0}]}\right]^{-1}\left[\frac{1}{n-1}\sum_{k=1}^{n}(r_k-\bar{r}(n))^2-\Phi_W\right].$$

2. For e_{W1} as in (6.2.10), an unbiased estimator of $V(e_{W1})$ as given in (6.2.12) may be obtained as follows. Since

$$E_R\left[\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\tfrac{1}{\nu-1}\textstyle\sum_{i\in u}(r_i-e_{W1})^2\right]=\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\tfrac{1}{\nu-1}\left[\textstyle\sum_{i\in u}E_R(r_i^2)-\nu E_R(e_{W1}^2)\right]$$
$$=\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\tfrac{1}{\nu-1}\left[\textstyle\sum_{i\in u}\{V_R(r_i)+(E_R(r_i))^2\}-\nu\{V_R(e_{W1})+(E_R(e_{W1}))^2\}\right]$$
$$=\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\tfrac{1}{\nu-1}\left[\textstyle\sum_{i\in u}\{\Phi_W+y_i^2\}-\nu\left\{\tfrac{\Phi_W}{\nu}-(\tfrac{\sum_{i\in u}y_i}{\nu})^2\right\}\right]$$
$$=\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\Phi_W+\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\tfrac{1}{\nu-1}\textstyle\sum_{i\in u}(y_i-\frac{\sum_{i\in u}y_i}{\nu})^2,$$

it follows that

$$E_PE_R\left[\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\tfrac{1}{\nu-1}\textstyle\sum_{i\in u}(r_i-e_{W1})^2|n\right]$$
$$=\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\Phi_W+E_P\left[\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\tfrac{1}{\nu-1}\textstyle\sum_{i\in u}(y_i-\frac{\sum_{i\in u}y_i}{\nu})^2|n\right]$$
$$=\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\Phi_W+\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\tfrac{1}{N-1}\textstyle\sum_{i=1}^{N}(y_i-\frac{\sum_{i=1}^{N}y_i}{N})^2$$
$$=\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\Phi_W+\left(\tfrac{1}{\nu}-\tfrac{1}{N}\right)\tfrac{N\theta(1-\theta)}{N-1}=V(e_{W1})-\tfrac{\Phi_W}{N}.$$

Hence,

$$\hat{V}(e_{W1})=\left(\frac{1}{\nu}-\frac{1}{N}\right)\frac{1}{\nu-1}\sum_{i\in u}(r_i-e_{W1})^2+\frac{\Phi_W}{N}.$$

3. For e_{W2} as in (6.2.11), an unbiased estimator of $V(e_{W2})$ as given in (6.2.13) can be obtained along lines similar to those used for e_{W1}. This will be given by

$$\hat{V}(e_{W2}) = \left(\frac{1}{\nu} - \frac{1}{N}\right) \frac{1}{\nu - 1} \sum_{i \in u} (g_i - e_{W2})^2 + \frac{\Phi_W}{N\nu} \left(\sum_{i \in u} \frac{1}{f_{is}} \right).$$

6.3 Estimation using Kuk's (1990) RR device in Simple Inverse Sampling with replacement

We now continue this study with Kuk's (1990) RR device applied to the persons selected via the inverse sample. Recalling this scheme from Section 5.4 and the notation used therein, we consider the following estimators corresponding to those studied with Warner's RRD in Section 6.2.

6.3.1 Some alternative estimators

Three estimators for θ based on RR's generated by this device may be denoted by e_K, e_{K1} and e_{K2}, and these are as follows. The notation used are as in (5.4.1) and (5.4.7).

$$e_K = \frac{1}{n} \sum_{k=1}^{n} \rho_k, \quad e_{K1} = \frac{1}{\nu} \sum_{i \in u} \rho_i \quad \text{and} \quad e_{K2} = \frac{1}{\nu} \sum_{i \in u} \gamma_i.$$

We can check on using (5.4.1) and (5.4.8) that

$$E(e_K) = E(e_K) = E(e_{K2}) = \theta,$$

and so these estimators are unbiased. Two expressions for variance of e_K and one each for the other two estimators may be obtained as

$$V_1(e_K) = \frac{N}{N-1} \theta(1-\theta) \binom{N-1}{\nu-1} \times$$

$$\left[\Delta^{\nu-1} \left\{ \frac{1}{x} + \frac{N(x-3)}{x^2} log(\frac{N}{N-x}) + \frac{2}{x} \sum_{n=\nu}^{\infty} \frac{1}{n^2} \left(\frac{x}{N}\right)^{n-1} \right\} |_{x=0} \right]$$

$$+(a\theta + b) E_n(\frac{1}{n}),$$

$$V_2(e_K) = \frac{\theta(1-\theta)}{N-1}\left[NE_n\left(\frac{1}{n}\right) - E_n\left(\frac{3n+1}{(n+1)^2}\right)\right] + (a\theta+b)E_n(\frac{1}{n}),$$

$$V(e_{K1}) = \left(\frac{1}{\nu} - \frac{1}{N}\right)\frac{N}{N-1}\theta(1-\theta) + \frac{a\theta+b}{\nu} \quad \text{and}$$

$$V(e_{K2}) = \left(\frac{1}{\nu} - \frac{1}{N}\right)\frac{N}{N-1}\theta(1-\theta) + \frac{1}{\nu^2}E_nE_{P|n}\left(\sum_{i\in u}\frac{ay_i+b}{f_{is}}\right).$$

6.3.2 Efficiency comparisons among e_K, e_{K1} and e_{K2}

(i) Comparing e_{K1} with e_{K2}: Following the arguments used in (5.3.1) we note that

$$\frac{1}{\nu^2}E_nE_{P|n}\left(\sum_{i\in u}\frac{ay_i+b}{f_{is}}\right) \leq \frac{a\theta+b}{\nu}. \tag{6.3.1}$$

So $V(e_{K2}) \leq V(e_{K1})$ and hence

$$e_{K2} \succ e_{K1} \text{ uniformly.}$$

(ii) Comparing e_K with e_{K2}: On applying (6.3.1) and since $E_n\left(\frac{3n+1}{(n+1)^2}\right) < \frac{3}{\nu}$, we note that

$$\begin{aligned} V(e_{K2}) - V_2(e_K) \;<\; & \left(\frac{1}{\nu} - \frac{1}{N}\right)\frac{N}{N-1}\theta(1-\theta) + \frac{a\theta+b}{\nu} \\ & -\frac{\theta(1-\theta)}{N-1}NE_n\left(\frac{1}{n}\right) + \frac{\theta(1-\theta)}{N-1}\frac{3}{\nu} - (a\theta+b)E_n(\frac{1}{n}). \end{aligned}$$

Further, using (6.2.15), we may claim that for large N

$$\begin{aligned} V(e_{K2}) - V_2(e_K) \;<\; & \left(\frac{1}{\nu} - \frac{1}{N}\right)\frac{N}{N-1}\theta(1-\theta) + \frac{a\theta+b}{\nu} \\ & +\frac{\theta(1-\theta)}{N-1}\frac{3}{\nu} - \left[\frac{\theta(1-\theta)}{N-1} + \frac{a\theta+b}{N}\right]\frac{1}{log\left(\frac{1}{1-f_0}\right)}. \end{aligned}$$

After some algebra it follows that $V(e_{K2}) - V_2(e_K) < 0$ if

$$\frac{\theta(1-\theta)}{N-1}\left[N\left(\frac{1}{\nu} - \frac{1}{N}\right) + \frac{3}{\nu} - \frac{1}{log\left(\frac{1}{1-f_0}\right)}\right] + a\theta\left(\frac{1}{\nu} - \frac{1}{Nlog\left(\frac{1}{1-f_0}\right)}\right)$$

$$\leq b\left(\frac{1}{Nlog\left(\frac{1}{1-f_0}\right)} - \frac{1}{\nu}\right).$$

We give some illustrations in Table 2.

Table 2: Some illustrations showing efficiency comparisons of the alternative estimators.

Comparison	N	ν	f_0	p_1	p_2	L	θ
$e_{K2} \succ e_{K1}$	any	any	—	any	any	any	any
$e_{K2} \succ e_K$	100	30	0.3	0.6	0.3	20	$0.218 \leq \theta \leq 0.759$
	250	50	0.2	0.6	0.3	20	$0.178 \leq \theta \leq 0.801$

6.3.3 Unbiased variance estimators

Proceeding as in Section 6.2, we obtain the following unbiased estimators for variances of e_K, e_{K1} and e_{K2}.

$$\hat{V}(e_K) = \frac{ae_K + b}{n} + \left\{ \left(\frac{N-n}{Nn} \right) + \frac{(n-1)(n-2)}{n^2} \frac{[\Delta^{\nu-1} x^{n-2}|_{x=0}]}{[\Delta^{\nu-1} x^{n-1}|_{x=0}]} \right\} \times$$

$$\left[1 - \frac{n-2}{n} \frac{[\Delta^{\nu-1} x^{n-2}|_{x=0}]}{[\Delta^{\nu-1} x^{n-1}|_{x=0}]} \right]^{-1} \left[\frac{1}{n-1} \sum_{k=1}^{n} (\rho_k - e_K)^2 - (ae_K + b) \right].$$

$$\hat{V}(e_{K1}) = \left(\frac{1}{\nu} - \frac{1}{N} \right) \frac{1}{\nu - 1} \sum_{i \in u} (\rho_i - e_{K1})^2 + \frac{ae_{K1} + b}{N}.$$

$$\hat{V}(e_{K2}) = \left(\frac{1}{\nu} - \frac{1}{N} \right) \frac{1}{\nu - 1} \sum_{i \in u} (\gamma_i - e_{K2})^2 + \frac{1}{N\nu} \left(\sum_{i \in u} \frac{a\gamma_i + b}{f_{is}} \right).$$

6.4 Estimation using Christofides's (2003) RR device in Simple Inverse Sampling with replacement

In this section, we use the same notation of transformed randomized responses obtained as $c_k, k = 1, \ldots, n$, $d_i, i \in u$ and $V_R(c_k) = \Phi_C$ defined in Section 5.5 of Chapter 5, and have observed that in Christofides's (2003) device the $V_R(c_k)$ turns out to be a constant involving only the parameters of the device, just as for Warner's (1965) device. So, the new estimators and the variance expressions for them are obtained straightforward following the same procedure as in Warner's device. These are shown below.

6.4.1 Some alternative estimators

(i) Let us consider the classical estimator analogous to the Warner's estimator for inverse sampling case

$$e_C = \frac{1}{n}\sum_{k=1}^{n} c_k = \bar{c}(n).$$

The two variance expression formulae $V_1(e_C)$ and $V_2(e_C)$ are as follows.

$$\begin{aligned} V_1(e_C) &= \frac{N\theta(1-\theta)}{N-1}\binom{N-1}{\nu-1} \times \\ &\quad \left[\Delta^{\nu-1}\left\{\frac{1}{x} + \frac{N(x-3)}{x^2} log\left(\frac{N}{N-x}\right) + \frac{2}{x}\sum_{n=\nu}^{\infty}\frac{1}{n^2}\left(\frac{x}{N}\right)^{n-1}\right\}|_{x=0}\right] \\ &\quad + \Phi_C E_n\left(\frac{1}{n}\right), \end{aligned}$$

$$V_2(e_C) = \frac{\theta(1-\theta)}{N-1}\left[N E_n\left(\frac{1}{n}\right) - E_n\left(\frac{3n+1}{(n+1)^2}\right)\right] + \Phi_C E_n\left(\frac{1}{n}\right).$$

(ii) We may consider the unbiased estimator based on only one response from the distinct respondents obtained in a simple inverse sample with replacement as

$$e_{C1} = \frac{1}{\nu}\sum_{i\in u} c_i.$$

The variance of this estimator $V(e_{C1})$ can be obtained as

$$V(e_{C1}) = V_P\left[\frac{1}{\nu}\sum_{i\in u} y_i\right] + E_P\left[\frac{\Phi_C}{\nu}\right] = \left(\frac{1}{\nu} - \frac{1}{N}\right)\frac{N}{N-1}\theta(1-\theta) + \frac{\Phi_C}{\nu}.$$

(iii) We now consider the unbiased estimator for θ based on multiple observations obtained from distinct respondents got in a simple inverse sample with replacement, namely e_{C2} and variance of it as follows.

$$e_{C2} = \frac{1}{\nu}\sum_{i\in u} d_i.$$

$$\begin{aligned} V(e_{C2}) &= V_P\left[\frac{1}{\nu}\sum_{i\in u} y_i\right] + E_P\left[\frac{1}{\nu^2}\sum_{i\in u}\frac{\Phi_C}{f_{is}}\right] \\ &= \left(\frac{1}{\nu} - \frac{1}{N}\right)\frac{N}{N-1}\theta(1-\theta) + \frac{\Phi_C}{\nu^2} E_n E_{P|n}\left(\sum_{i\in u}\frac{1}{f_{is}}\right). \end{aligned}$$

6.4.2 Efficiency comparisons among e_C, e_{C1} and e_{C2}

(i) Comparing e_{C1} with e_{C2}:

On applying (5.3.13) on the difference in between $V(e_{C2})$ and $V(e_{C1})$ we note that $V(e_{C2}) \leq V(e_{C1})$. Thus

$$e_{C2} \succ e_{C1} \text{ uniformly.}$$

(ii) Comparing e_C with e_{C2}: Proceeding as for Warner's device in Section 6.2 and using (6.2.15), we may conclude that if $\nu \to \infty$, $N \to \infty$, $\frac{\nu}{N} \to f_0 (0 < f_0 < 1)$, then for large N,

$$V(e_{C2}) - V_2(e_C) < 0 \text{ if}$$

$$\theta(1-\theta) \geq \frac{N-1}{N} \Phi_C \left[\frac{N log\left(\frac{1}{1-f_0}\right) - \nu}{\nu - (N+3-\nu) log\left(\frac{1}{1-f_0}\right)} \right].$$

6.4.3 Unbiased variance estimators

1. For e_C proceeding as in Section 6.2, we may provide unbiased estimator of $V(e_C)$ as follows.

$$\hat{V}(e_C) = \frac{\Phi_C}{n} + \left\{ \left(\frac{N-n}{Nn} \right) + \frac{(n-1)(n-2)}{n^2} \frac{[\Delta^{\nu-1} x^{n-2}|_{x=0}]}{[\Delta^{\nu-1} x^{n-1}|_{x=0}]} \right\} \times$$

$$\left[1 - \frac{n-2}{n} \frac{[\Delta^{\nu-1} x^{n-2}|_{x=0}]}{[\Delta^{\nu-1} x^{n-1}|_{x=0}]} \right]^{-1} \left[\frac{1}{n-1} \sum_{k=1}^{n} (c_k - \bar{c}(n))^2 - \Phi_C \right].$$

2. For e_{C1} , an unbiased estimator of $V(e_{C1})$ will be obtained as:

$$\hat{V}(e_{C1}) = \left(\frac{1}{\nu} - \frac{1}{N} \right) \frac{1}{\nu - 1} \sum_{i \in u} (c_i - e_{C1})^2 + \frac{\Phi_C}{N}.$$

3. For e_{C2}, an unbiased estimator of $V(e_{C2})$ can be obtained as:

$$\hat{V}(e_{C2}) = \left(\frac{1}{\nu} - \frac{1}{N} \right) \frac{1}{\nu - 1} \sum_{i \in u} (d_i - e_{C2})^2 + \frac{\Phi_C}{N\nu} \left(\sum_{i \in u} \frac{1}{f_{is}} \right).$$

Chapter 7

Modifying classical randomized response techniques with provision for true response

Abstract

In this chapter we consider some well-known RR models other than those considered in the previous two chapters, for instance, the unrelated question model, the forced response model, a model for quantitative responses, etc. Traditionally these models are applied to respondent chosen by SRSWR. We show that a modification may be applied to them and we examine how, irrespective of how a sample is drawn, we may gain in efficiency in estimation. We illustrate our findings through a numerical exercise.

7.1 Introduction

In Chapters 5 and 6, we focussed on the randomized response models due to Warner (1965), Kuk (1990) and Christofides (2003), and based on the randomized responses generated by these devices, we proposed some new estimators for estimating the proportion of people bearing a stigmatizing attribute in a population. However, in the RR literature, several other models have been proposed and studied. In this chapter, we focus on some of these models, namely, the *unrelated question model* of Horvitz et al. (1967), Greenberg et al. (1969), the *unknown repeated trial model* of Singh and Joarder's (1997) model, and the *forced response model* due to Chaudhuri

and Mukerjee (1988). All the above mentioned models were developed with an SRSWR sample of respondents.

While studying the estimators in Chapters 5 and 6, we considered that respondents are chosen by SRSWR; in keeping with the usual assumption in the literature on randomized responses where most of the results are developed employing only SRSWR for selecting the respondents. In this chapter, we relax this restriction of SRSWR sampling and allow respondents to be chosen by any varying probability sampling scheme; which, in particular, will also include SRSWR.

Mangat and Singh (1990) proposed a modification to Warner's (1965) method and studied conditions under which this modification led to improved efficiency of estimation. In this chapter we show that *irrespective of how a sample is drawn*, Mangat and Singh's (1990) modification may also be profitably applied to the models considered in this chapter.

Again, in Chapters 5 and 6 we only considered the case where the stigmatizing variable is a qualitative one. However, as described in Section 1.2, there are situations where the stigmatizing variable y is a quantitative one and one would like to estimate the population mean $\bar{Y}$. RR techniques are available for this problem, for instance, the one developed by Eichhorn and Hayre (1983) and many others; these being mainly based on SRSWR samples. We conclude this chapter by applying Mangat and Singh's (1990) modification to such a model and do not restrict to only SRSWR samples.

7.2 Mangat and Singh's modification to Warner's model

We will generically write r_i to denote the transformed RR under any of the models we will consider in this chapter. After applying Mangat and Singh's (1990) modification to a model, the corresponding transformed RR will again, be generically denoted by r_i'. We will show in Section 7.3.1 that for all of the models considered by us, when the sample s of respondents is chosen with probabilities $p(s)$, p being any arbitrary sampling design, the estimator is a function of r_i for the original model and the same function of r_i' under its modified version. More importantly, the variance of each of these estimators will be a sum of two terms, the first term being a constant

depending on the design p and the second part a function of the variances of either r_i or r_i', as the case may be. So, in order to compare the variances of the estimators under the original model and its modified version, it is enough to compare the variances of r_i and r_i'. We now proceed to carry out this comparison.

We recall Warner's RR model from section 1.2 and the quantities y_i, r_i and I_i as defined therein. Suppose we label the box of cards used in Warner's RR device as Box 1. In Mangat and Singh's (1990) modification, as described in Section 1.2, a second box is constituted with two types of cards as follows: a proportion $T(0 < T < 1)$ of cards being marked 'True' and the remaining ones marked 'RR'. We label this box as Box 2. In Mangat and Singh's (1990) modification to Warner's model, as detailed in Section 1.2, a sampled person i is asked to first draw a card out of Box 2. He/she is requested to report the true value y_i if a 'True' marked card is drawn, and otherwise to use Box 1 to produce a response I_i as in Warner's technique, I_i being defined in (1.2.17). Thus, the response from the i^{th} person sampled is

$$\begin{aligned} z_i &= y_i \text{ if a 'True' card is drawn from Box 2,} \\ &= I_i \text{ if an 'RR' card is drawn instead.} \end{aligned}$$

Since the interviewer does not witness the steps of the trial carried out by the respondent to arrive at his/her response, the respondent's privacy is protected. Clearly,

$$E_R(z_i) = Ty_i + (1-T)\left[py_i + (1-p)(1-y_i)\right]. \tag{7.2.1}$$

Then, forming the transformed response as

$$r_i' = \frac{z_i - (1-T)(1-p)}{T + (1-T)(2p-1)}$$

where T and p are such that $T + (1-T)(2p-1) \neq 0$, we can show that

$$E_R(r_i') = y_i \quad \text{and} \quad V_R(r_i') = \frac{(1-T)(1-p)(T+p-pT)}{\left[T + (1-T)(2p-1)\right]^2} \tag{7.2.2}$$

Now, on comparing the variances of the transformed responses r_i and r_i' under Warner's model and its modification, as given in (1.2.19) and (7.2.2), respectively, we have

$$V_R(r_i') \leq V_R(r_i) \quad \forall\, i \in U \text{ if one chooses } T \geq \frac{1-2p}{1-p} \tag{7.2.3}$$

7.2.1 Modification applied to the unrelated question model

As described in Section 1.2, the unrelated question model studied by Horvitz et al. (1967), Greenberg et al. (1969) use another innocuous human attribute B not correlated with the stigmatizing attribute A under study, and uses two boxes as the RR devices. Let these two boxes be labeled Box 3 and Box 4; Box 3 consisting of cards marked either A or B in proportions $p_1 : (1 - p_1);\ (0 < p_1 < 1)$ and Box 4 containing similarly marked cards in proportions $p_2 : (1 - p_2);\ (0 < p_2 < 1)$, with $p_1 \neq p_2$.

Analogous to the variable y defined in Section 1.2 in connection with the stigmatizing attribute A, let x be an indicator variable such that x_i is 1 or 0 corresponding to whether unit i bears the innocuous attribute B or not. A sampled person is asked to draw a card independently from each of the Boxes 3 and 4 and then give a truthful response y_i or x_i, according as whether the card type drawn is A or B. Similar to I_i in (1.2.17), let these two RR's be denoted by I_i for Box 3 and J_i for Box 4. Then, clearly,

$$E_R(I_i) = p_1 y_i + (1 - p_1)x_i, \qquad E_R(J_i) = p_2 y_i + (1 - p_2)x_i$$

and we may form the transformed RR's as

$$r_i = \frac{(1 - p_2)I_i - (1 - p_1)J_i}{p_1 - p_2}, \text{ where } p_1 \neq p_2.$$

On simplification it follows that

$$E_R(r_i) = y_i \text{ and } V_R(r_i) = \frac{(1-p_1)(1-p_2)\{p_1(1-p_2) + p_2(1-p_1)\}}{(p_1-p_2)^2}(y_i-x_i)^2 \tag{7.2.4}$$

To modify this procedure along the line of Mangat and Singh (1990), our suggestion is that a sampled person i be requested to first draw one card from Box 2. If an 'RR'-marked card appears, the respondent is to produce two independent responses I_i and J_i using Box 3 and 4, respectively. Otherwise, i.e., if the card drawn from Box 2 is marked 'True', then the true response y_i is to be given. We write the RR in this case as

$$\begin{aligned} z_i &= y_i \text{ if a 'True' marked card is drawn from Box 2,} \\ &= I_i \text{ if an 'RR' marked card is drawn from Box 2 and} \\ &\quad \text{then Box 3 is used} \\ z_i' &= y_i \text{ if a 'True' marked card is drawn from Box 2} \\ &= J_i \text{ if an 'RR' marked card is drawn from Box 2 and} \\ &\quad \text{then Box 4 is used.} \end{aligned}$$

Then,

$$\begin{aligned} E_R(z_i) &= Ty_i + (1-T)\left[p_1 y_i + (1-p_1)x_i\right] \\ E_R(z_i') &= Ty_i + (1-T)\left[p_2 y_i + (1-p_2)x_i\right]. \end{aligned}$$

This leads to the transformed RR

$$r_i' = \frac{(1-p_2)z_i - (1-p_1)z_i'}{(p_1-p_2)} \text{ with } p_1 \neq p_2,$$

for which

$$E_R(r_i') = y_i \quad \text{and}$$

$$V_R(r_i') = \frac{(1-T)(1-p_1)(1-p_2)\left[(1-p_2)(T+p_1-p_1T)+(1-p_1)(T+p_2-p_2T)\right]}{(p_1-p_2)^2}(y_i-x_i)^2. \tag{7.2.5}$$

On comparing (7.2.4) and (7.2.5), it follows that

$$V_R(r_i') \leq V_R(r_i)\ \forall\ i \in U \text{ if one chooses } T \geq 1 - \frac{p_1(1-p_2)+p_2(1-p_1)}{2(1-p_1)(1-p_2)}. \tag{7.2.6}$$

7.2.2 Modification applied to the unknown repeated trial model

We may recall from Section 1.2, Singh and Joarder's (1997) unknown repeated trial model where they recommend that a sampled person is asked to generate an RR using Box 1 mentioned in Section 7.2. If he bears A^c, he is to report this RR. However, if he bears A and his response is 'yes', implying 'matching' of the true trait and the card type drawn, then he/she is to report 'yes'. Otherwise, he/she should generate another RR using Box 1 and then report this RR.

Writing z_i as the final response by this RR technique, we have

$$\begin{aligned} E_R(z_i) &= P(z_i = 1) = (1-y_i)(1-p) + y_i\left[p + (1-p)p\right] \\ &= E_R\left[I_i\right] + p(1-p)y_i \\ &= \left[(2p-1)y_i + (1-p)\right] + p(1-p)y_i \\ &= \left[(2p-1) + p(1-p)\right]y_i + (1-p), \end{aligned}$$

where I_i is as in (1.2.17). This leads to the transformed RR

$$r_i = \frac{z_i - (1-p)}{\alpha} \text{ with } E_R(r_i) = y_i \text{ and } V_R(r_i) = \frac{p(1-p)}{\alpha^2}\left[1 - \alpha y_i\right]. \tag{7.2.7}$$

on writing $\alpha = (2p-1) + p(1-p)$.

Applying Mangat and Singh's (1990) modification on this RR technique by using as before, Box 2 described in Section 7.2, to permit a 'True' answer as well, the RR may be written as

$$\begin{aligned} z_i' &= y_i \text{ with probability } T, \\ &= z_i \text{ as above, with probability } (1-T). \end{aligned}$$

Thus,

$$\begin{aligned} E_R(z_i') &= P(z_i' = 1) = Ty_i + (1-T)\left[\alpha y_i + (1-p)\right] \\ &= y_i\alpha' + (1-T)(1-p), \end{aligned}$$

writing $\alpha' = T + (1-T)\alpha$. This gives us the transformed RR for the modified model as

$$r_i' = \frac{z_i' - (1-T)(1-p)}{\alpha'},$$

with

$$E_R(r_i') = y_i,\ V_R(r_i') = \frac{(1-T)(1-p)}{(\alpha')^2}\left[(T + p - pT) - y_i p\alpha'\right]. \qquad (7.2.8)$$

Comparing (7.2.7) and (7.2.8), we can see that Singh and Joarder's (1997) RR device can be improved upon by Mangat and Singh's (1990) modification to ensure

$$V_R(r_i') \leq V_R(r_i) \quad \forall i \in U$$

if T is so chosen that

$$T \geq max\left[\frac{(1-3p+p^2)(1-p+p^2)}{(1-p)[1-2p+3p^2-p^3]},\quad \frac{(1-3p+p^2)(1+p-p^2)}{(1-p)^2[1+2p-p^2]}\right] \qquad (7.2.9)$$

7.2.3 Modification applied to forced response model

As described in Section 1.2, in Chaudhuri & Mukerjee's (1988) 'Forced Response Model', a sampled person i is given a box with three kinds of cards marked respectively 'yes', 'no' and 'A' in proportions $p_1 : p_2 : (1-p_1-p_2)$; with $p_1 + p_2 < 1$. He is requested to answer I_i as 1 if a card bearing 'yes' is drawn; I_i as 0 for a 'no' card; otherwise to truthfully report $I_i = 1$ or 0 according as whether he bears A or not. Then we get,

$$E_R(I_i) = Prob(I_i = 1) = p_1 + (1-p_1-p_2)y_i$$

leading to

$$\begin{aligned} r_i &= \tfrac{I_i - p_1}{1-p_1-p_2} \\ E_R(r_i) &= y_i \\ V_R(r_i) &= \tfrac{(p_2-p_1)(1-p_1-p_2)y_i + p_1(1-p_1)}{(1-p_1-p_2)^2}. \end{aligned} \qquad (7.2.10)$$

Using Box 2 to apply Mangat and Singh's (1990) modification on this model, we get the RR

$$\begin{aligned} z_i &= y_i \text{ with probability } T, \\ &= I_i \text{ as above with probability } (1-T). \end{aligned}$$

Then,

$$E_R(z_i) = Ty_i + (1-T)\left[(1-p_1-p_2)y_i + p_1\right]$$

leading to the transformed RR

$$r_i' = \frac{z_i - p_1(1-T)}{T + (1-T)(1-p_1-p_2)}$$

with

$$\begin{aligned} E_R(r_i') &= y_i \\ V_R(r_i') &= \frac{\left[(1-p_1-p_2)+T(p_1+p_2)\right](1-T)(p_2-p_1)y_i + p_1(1-T)(1-p_1+Tp_1)}{(1-p_1-p_2+Tp_1+Tp_2)^2}. \end{aligned} \tag{7.2.11}$$

To examine when the Forced Response model with Mangat and Singh's (1990) modification is superior to the classical one, we study when $V_R(r_i') < V_R(r_i)$. For this, we first note that $(1-p_1-p_2+p_1T+p_2T)^2 > (1-p_1-p_2)^2$ yielding

$$\frac{1}{(1-p_1-p_2+p_1T+p_2T)^2} < \frac{1}{(1-p_1-p_2)^2}.$$

So for $V_R(r_i') \leq V_R(r_i)$, we need to compare only the coefficients of y_i's and the constant terms in the numerators separately. Now, after some algebra we see that

$$V_R(r_i') < V_R(r_i) \quad \forall i \in U$$

if we take $p_2 > p_1$ and T satisfies

$$T \geq max\left[\frac{2(p_1+p_2)-1}{(p_1+p_2)}, \quad \frac{2p_1-1}{p_1}\right] \tag{7.2.12}$$

Clearly, in (7.2.12), $T \geq \{2(p_1+p_2)-1\}/(p_1+p_2)$ for $p_1+p_2 \leq 0.5$ while $T \geq 2(p_1-1\}/p_1$ for $p_1 \leq 0.5$.

7.2.4 Modification applied to model for quantitative stigmatizing variable

Let y_i be the value of the quantitative stigmatizing variable y, for a sampled person i. In this RR model, person i is given two boxes, one containing cards

marked $a_1, \ldots, a_j, \ldots, a_M$ and another with cards marked $b_1, \ldots, b_k, \ldots, b_L$. The person is requested to draw one card from each of these two boxes independently, and to respond $z_i = a_j y_i + b_k$ if he/she happens to choose a_j and b_k marked cards. Let

$$\mu_A=\frac{1}{M}\sum_{j=1}^{M} a_j,\ \mu_B=\frac{1}{L}\sum_{k=1}^{L} b_k,\ \sigma_A^2=\frac{1}{M}\sum_{1}^{M}(a_j-\mu_A)^2,\ \text{and}\ \sigma_B^2=\frac{1}{L}\sum_{1}^{L}(b_k-\mu_B)^2.$$

Then, we may obtain a transformed RR as

$$r_i = \frac{z_i - \mu_B}{\mu_A},$$

for which

$$E_R(r_i) = y_i \quad \text{and} \quad V_R(r_i) = \frac{1}{\mu_A^2}\left(y_i^2\sigma_A^2 + \sigma_B^2\right) \tag{7.2.13}$$

Applying Mangat and Singh's (1990) modification to this model, the corresponding RR from the i^{th} person is

$$\begin{aligned} z_i' &= y_i \text{ with probability } T, \\ &= z_i \text{ as above with probability } (1-T), \end{aligned}$$

with $E_R(z_i') = [T + (1-T)\mu_A]y_i + (1-T)\mu_B$. Now, we may form the transformed RR as

$$r_i' = \frac{z_i' - (1-T)\mu_B}{T + (1-T)\mu_A},\ \text{with}\ E_R(r_i') = y_i \text{ and}$$

$$V_R(r_i') = \frac{(1-T)\left[y_i^2\left\{T+\sigma_A^2+T\mu_A^2-2T\mu_A\right\} + y_i\left\{2T\mu_B(\mu_A-1)\right\} + (\sigma_B^2+T\mu_B^2)\right]}{\left[T+(1-T)\mu_A\right]^2} \tag{7.2.14}$$

Without loss of generality, we may take $\mu_A = 1$ and $\mu_B = 0$. Then, from (7.2.13) and (7.2.14) we get that

$$V(r_i) = y_i^2\sigma_A^2 + \sigma_B^2 \text{ and } V_R(r_i') = (1-T)\left[y_i^2\sigma_A^2 + \sigma_B^2\right] \tag{7.2.15}$$

Now, since $0 < T < 1$, (7.2.15) implies that

$$V_R(r_i') < V_R(r_i) \ \ \forall\ i \in U.$$

Hence, with the choices $\mu_A = 1$ and $\mu_B = 0$, the modified version of the above RR technique for quantitative stigmatizing variable is always superior to the original RR technique.

7.3 Unbiased estimators and variance estimators based on RR's obtained from samples chosen by varying probabilities

With y_i as defined in the preceding sections, when we are studying a qualitative characteristic A, then the population mean $Y = \sum_{i=1}^{N} y_i/N$ is the proportion of persons bearing A, and for a quantitative stigmatizing variable, Y is the population total of interest. So, in either case, we consider the problem of estimating Y based on a sample s chosen from U with probabilities $p(s)$ by any suitable sampling design p.

If direct responses y_i's are available on the respondents in s, one may use a homogeneous linear unbiased estimator of the form

$$\hat{Y}_b = \sum_{i \in s} y_i b_{si}, \tag{7.3.1}$$

where b_{si}'s are free of y_i's and satisfy $\sum_{s \ni i} p(s) b_{si} = 1, \forall\ i \in U$. Then it is well known that

$$E_P(\hat{Y}_b) = Y \text{ and } V_P(\hat{Y}_b) = \sum_{i=1}^{N} y_i^2 c_i + \sum \sum_{i,j=1, i \neq j}^{N} y_i y_j c_{ij} \tag{7.3.2}$$

where $c_i = E_P(b_{si}^2 I_{si}) - 1$ and $c_{ij} = E_P(b_{si} b_{sj} I_{sij}) - 1$ where I_{si} and I_{sij} are as in (1.2.3). Let c_{si} and c_{sij} be such that $E_P(c_{si} I_{si}) = c_i$ and $E_P(c_{sij} I_{sij}) = c_{ij}$. Then it follows that

$$E_P \left[\sum_{i \in s} y_i^2 c_{si} + \sum \sum_{i \neq j, i, j \in s} y_i y_j c_{sij} \right] = V_P(\hat{Y}_b) = E_P[\hat{V}_P(\hat{Y}_b)], \tag{7.3.3}$$

say. Thus $\hat{V}_P(\hat{Y}_b) = \sum_{i \in s} y_i^2 c_{si} + \sum \sum_{i \neq j, i, j \in s} y_i y_j c_{sij}$ is an unbiased estimator for $V_P(\hat{Y}_b)$.

7.3.1 Unbiased estimators based on RR's and their variances

When instead of y_i's, only the RR's under any of the models discussed in Section 7.2 are available, we use the corresponding transformed RR's to form the following estimator for Y

$$e_b = \sum_{i \in s} r_i b_{si}. \tag{7.3.4}$$

Since $E_R(r_i) = y_i$ for each model, from (7.3.1) and (7.3.2) it follows that $E_P E_R(e_b) = Y$. Thus e_b is unbiased for Y.

The estimator based on the RR's obtained from the modified version of the corresponding model will be

$$e_b' = \sum_{i \in s} r_i' b_{si} \tag{7.3.5}$$

and again, as $E_R(r_i') = y_i$ for each model, e_b' is unbiased for Y.

To obtain the variance of e_b we note that

$$V(e_b) = E[e_b - Y]^2 = E_P E_R[e_b - Y]^2 = V_P(\hat{Y}_b) + E_P(\sum_{i \in s} b_{si}^2 V_R(r_i)), \tag{7.3.6}$$

where the first term is as in (7.3.2) and depends on the design p, while the $V_R(r_i)$ in the second term is specific to the model used and is as given in (1.2.19), (7.2.4), (7.2.7), (7.2.10) or (7.2.13). Similarly, for the modified version of the model,

$$V(e_b') = V_P(\hat{Y}_b) + E_P(\sum_{i \in s} b_{si}^2 V_R(r_i')), \tag{7.3.7}$$

where the $V_R(r_i')$ for the modified versions of the different models are as given in (7.2.2), (7.2.5), (7.2.8), (7.2.11) or (7.2.14). From (7.3.6) and (7.3.7) it is now evident that in order to study the relative performances of e_b and e_b' by comparing their variances, it is enough to compare $V_R(r_i)$ with $V_R(r_i')$, as was done in Section 7.2.

7.3.2 Unbiased variance estimators

For Warner's model, it is clear from (1.2.19) and (7.2.2) that both $V_R(r_i)$ and $V_R(r_i')$ are constants depending on the parameters of the devices, namely p and T. So for Warner's model, both the variances $V_R(r_i)$ and $V_R(r_i')$ are known. Hence, from (7.3.6), to estimate $V(e_b)$ and $V(e_b')$ we now first need to estimate $V_R(r_i)$ and $V_R(r_i')$ for the other models considered so far in this chapter. For this, we will again use the notation $\hat{V}_R(r_i)$ and $\hat{V}_R(r_i')$ to generically denote the unbiased variance estimators under the different models.

Towards this, we observe from section 7.2.1–7.2.3 that for all these models, $E_R(r_i) = y_i = E_R(r_i')$, $y_i^2 = y_i$, and so, $V_R(r_i) = E_R[r_i(r_i - 1)], V_R(r_i') = E_R[r_i'(r_i' - 1)]$ So, for each of these three models,

$$\hat{V}_R(r_i) = r_i(r_i - 1) \text{ and } \hat{V}_R(r_i') = r_i'(r_i' - 1). \tag{7.3.8}$$

For the model for quantitative variable, we can see from (7.2.13) and (7.2.15) that both $V_R(r_i)$ and $V_R(r_i')$ may be written in the form

$$V_R(r_i) = ay_i^2 + b \text{ and } V_R(r_i') = a'y_i^2 + b',$$

where a, a', b, b' are known quantities. Clearly,

$$E_R\left[\frac{ar_i^2 + b}{(1+a)}\right] = V_R(r_i), \quad E_R\left[\frac{a'r_i'^2 + b'}{(1+a')}\right] = V_R(r_i'),$$

and so, under this model,

$$\hat{V}_R(r_i) = \frac{ar_i^2 + b}{(1+a)} \text{ and } \hat{V}_R(r_i') = \frac{a'r_i'^2 + b'}{(1+a')}. \tag{7.3.9}$$

Using the form of $\hat{V}_P(\hat{Y}_b)$ in (7.3.3), let

$$v_1 = \sum_{i\in s} r_i^2 c_{si} + \sum \sum_{i\neq j, i,j\in s} r_i r_j c_{sij} + \sum_{i\in s} \hat{V}_R(r_i)(b_{si}^2 - c_{si}),$$

and

$$v_2 = \sum_{i\in s} r_i^2 c_{si} + \sum \sum_{i\neq j, i,j\in s} r_i r_j c_{sij} + \sum_{i\in s} \hat{V}_R(r_i) b_{si}. \tag{7.3.10}$$

Then, $E_P E_R(v_1) = V(e_b) = E_P E_R(v_2)$. So both v_1 and v_2 are unbiased variance estimators for e_b, where the $\hat{V}_R(r_i)$ correspond to the particular model being referred to.

Similarly, replacing r_i by r_i' and $\hat{V}_R(r_i)$ by $\hat{V}_R(r_i')$ in (7.3.10), unbiased variance estimators for e_b' can be obtained.

7.4 Numerical illustrations showing the gains in efficiencies

In order to demonstrate how our method can be effectively applied in a practical survey situation, we present a numerical study. For this, using (7.2.3), (7.2.6), (7.2.9) and (7.2.12), we first derive the admissible ranges of T in Mangat and Singh's (1990) device corresponding to some values of the classical device parameters p or p_1 and p_2 for each of the four models considered in Sections 7.2, 7.2.1, 7.2.2 and 7.2.3 , to ensure gain in efficiency in modifying by their technique. For the quantitative variable model, we recall

from Section 7.2.4, that considering $\mu_A = 1$ and $\mu_B = 0$, the modification is always superior to the classical one for all T.

For our numerical study, we consider the data on household expenditure and household size for the state of West Bengal, India. These data are obtained from a sample survey conducted by National Sample Survey Organization, India in its 61st round. We take a random sample of 417 households from this data and assume this to be the population of interest. This means that our population comprises $N = 417$ households and for each of these, the household size, together with its monthly expenditures (in Indian Rupees) on food and on non-food items are available. We may suppose that household members usually will hesitate to reveal their actual monthly expenditure on non-food items because this amount includes their expenditure on sensitive variables like alcoholic drinks, excessive luxurious entertainment, etc. We consider the households spending more than Rs. 1500 as bearing the sensitive characteristic. To illustrate our work, we estimate the proportion of households, say, θ, bearing this sensitive character in the said community, based on a sample drawn by an unequal probability without replacement sampling scheme from that population. We also estimate the actual mean expenditure $\bar{Y}$ on non-food items per household. Note that, as remarked before, with our definition of y_i used in this chapter, both problems reduce to that of estimating Y.

We also consider the data x on some innocuous variable unrelated to sensitive qualitative variable describing $x_i = 1$ if the head of the household is a business-man, else $x_i = 0$.

We draw a sample of size $n = 105$ from the above population of size $N = 417$, employing Rao, Hartley and Cochran's sampling scheme (RHC, 1962) with household size as the size measure. The details of this scheme is given in Section 1.2, with the estimator and variance estimator as shown in (1.2.12) and (1.2.14).

However, the y_i's are not directly ascertainable in randomized response surveys. So, under this RHC sampling scheme, unbiased estimators for Y under the classical models and also their modified versions are obtained from (7.3.4) and (7.3.5) as

$$e = \sum_n \frac{Q_i}{p_i} r_i, \text{ and } e' = \sum_n \frac{Q_i}{p_i} r_i',$$

respectively, with $b_{si} = Q_i/p_i$, and r_i and r_i' corresponding to the model

being considered.

An unbiased variance estimator for e is then as obtained from (7.3.10). Using the form v_2 in (7.3.10), (1.2.14) and a result in Chaudhuri, Adhikary and Dihidar (2000), we get this as

$$v(e) = \frac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2} \Sigma_n \Sigma_n Q_i Q_j \left(\frac{r_i}{p_i} - \frac{r_j}{p_j} \right)^2 + \sum_n \hat{V}_R(r_i) \frac{Q_i}{p_i},$$

where again, the r_i and $\hat{V}_R(r_i)$ correspond to the model being considered. By replacing r_i by r_i' and $\hat{V}_R(r_i)$ by $\hat{V}_R(r_i')$ in $v(e)$ above, we obtain an unbiased variance estimator for e'.

We use the standard error (se) and the coefficient of variation (cv) as the criteria for our efficiency comparisons. The cv is defined in (1.1.2) and it is well known that less the cv, the more efficient is the corresponding estimator. In the following tables, for some chosen values of the device parameters, we present the model-wise performance of the estimator e under the classical model and also the estimator e' under the corresponding modified version. For the qualitative models considered in this chapter, we have chosen the value of T as belonging to the admissible range as shown in Table 1.

Qualitative models

Table 1. Model-wise performances for qualitative data by the classical method and the modified method

Warner's model

Device p	Admissible Range of T	T used	e	e'	se of e	se of e'	cv(%) of e	cv(%) of e'
0.20	≥ 0.750	0.80	0.292	0.279	0.082	0.069	28.21	24.87
0.25	≥ 0.667	0.70	0.322	0.375	0.095	0.089	29.59	23.67
0.30	≥ 0.571	0.60	0.450	0.442	0.120	0.112	26.75	25.32

Unrelated question model

Device p_1	Device p_2	Admissible Range of T	T used	e	e'	se of e	se of e'	cv(%) of e	cv(%) of e'
0.30	0.45	≥ 0.377	0.40	0.551	0.530	0.238	0.205	43.11	38.65
0.50	0.30	≥ 0.286	0.30	0.363	0.473	0.158	0.157	43.61	33.23
0.50	0.60	Any T	0.20	0.502	0.488	0.212	0.164	42.18	33.59

Unknown repeated trials model

Device p	Admissible Range of T	T used	e	e'	se of e	se of e'	cv(%) of e	cv(%) of e'
0.15	≥ 0.769	0.80	0.263	0.342	0.084	0.081	31.75	23.73
0.20	≥ 0.649	0.65	0.284	0.364	0.105	0.094	37.03	25.91
0.25	≥ 0.504	0.55	0.420	0.426	0.143	0.110	34.07	25.74

Forced response model

Device p_1	Device p_2	Admissible Range of T	T used	e	e'	se of e	se of e'	cv(%) of e	cv(%) of e'
0.20	0.25	Any T	0.35	0.330	0.271	0.090	0.066	27.40	24.24
0.30	0.40	≥ 0.571	0.60	0.575	0.396	0.165	0.064	28.64	16.03
0.30	0.35	≥ 0.462	0.50	0.480	0.326	0.143	0.066	29.75	20.40

Quantitative model

For this model, y is the amount in Indian rupees spent for non-food items by each household. We may want to estimate $\bar{Y}$. We construct the RR devices with $M = 10$ and $L = 12$ and

$(a_1, \ldots, a_{10})$ = (0.935, 0.759, 0.764, 1.124, 1.172, 1.048, 0.817, 1.196, 1.223, 0.923) and $(b_1, \ldots, b_{12})$ = (-35, -50, 31, -17, -29, -15, 43, 10, 23, 30, 18, -9). Thus, $\mu_A = 0.9961 \approx 1$ and $\mu_B = 0$, as we had assumed in Section 7.2.4. For these μ_A and μ_B values we have seen in our theoretical derivation in Section 7.4 that our modified estimator performs better than the classical one, for all T, $(0 < T < 1)$. In Table 2, we choose some specific values of T to compare the performances of estimators e and e'.

We may remark here that for our population we see from the data that the range of non-food expenditure is (Rs.90.00, Rs.17887.33), whereas for above choice of the values of a_j and b_k, the range of $a_j y_i + b_k$ happen to be (Rs.18.31, Rs.21919.20). So, by these choices, the device covers the range of the y_i's. In practical survey situations, the range of the y_i's may not be known in advance, and so, the devices need to be made based on some prior guess.

Table 2. Performance for quantitative model by the classical method and by the modified method

Admissible Range of T	T used	e	e'	se of e	se of e'	cv(%) of e	cv(%) of e'
Any T	0.20	1546.23	1511.42	184.873	161.807	11.96	10.71
Any T	0.25	1559.93	1536.23	172.450	152.619	11.06	9.94
Any T	0.40	1615.54	1589.04	186.525	167.090	11.55	10.52

7.5 Concluding remarks

We observe from Tables 1 and 2 that Mangat and Singh's (1990) technique can be profitably applied in all the above mentioned models. Some cases of course show large values of estimated coefficient of variation, but they are still lower than the estimated coefficient of variations for the available classical method. So, one should not hesitate to give a chance to the respondent to reveal truthfully their true value by lottery method unnoticed by the interviewer. The device for that lottery method, i.e. Box 2, should be prepared by following the admissible ranges of T values, in order to achieve better results.

References

Abul-Ela, Abdel-Latif, A., Greenberg, B.G. and Horvitz, D.G.(1967). A multiproportions randomized response models. *Journal of the American Statistical Association.* **62**, 990-1008.

Arnab, R. (1999). On use of distinct respondents in RR surveys. *Biometrical Journal.* **41(4)**, 507-513.

Asok, C. and Sukhatme, B.V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association.* **71**, 912-918.

Basu, D. (1958). On sampling with and without replacement. *Sankhya.* **20**, 287-294.

Berg, S. (1974). A note on the RHC method. *Scandinavian Journal of Statistics.* **57**, 108-114.

Biyani, S.H. (1980). On inadmissibility of the Yates-Grundy variance estimator in unequal probability sampling. *Journal of the American Statistical Association.* **75**, 709-712.

Bose, M., Chaudhuri, A., Dihidar, K. and Das, S. (2009). Model-cum-design based estimation of the prevalence rate of a disease in a locality using spatial smoothing. Accepted for publication in *Statistics.*

Bourke, P.D. (1978). Randomized response designs with symmetric response for multiproportions situations. *Statisics Tidskrift.* **16**, 197-207.

Brewer, K.R.W. (1990). Review of 'Unified theory and strategies of survey sampling' by Chaudhuri, A. and Vos, J.W.E. *Journal of Official Statistics.* **6**, 101-104.

Brewer, K.R.W. (2001). Deriving and estimating an approximate variance for the Horvitz-Thompson estimator using only first order inclusion probabilities. *In ICES - II, Proceedings of the Second International Conference on Establishment Surveys.*

Brewer, K.R.W. and Hanif, M. (1983). Sampling with unequal probabilities. Springer-Verlag. New York.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika.* **63**, 615-620.

Chambers, R., Chambers, R.L. and Skinner, C.J. (2003). Analysis of survey data. John Wiley and Sons, Ltd.

Chaudhuri, A. (2000). Network and adaptive sampling with unequal probabilities. *Calcutta Statistical Association Bulletin.* **50**, 237-253.

Chaudhuri, A. (2001). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *Journal of Statistical Planning and Inference.* **94**, 37-42.

Chaudhuri, A., Adhikary, A.K. and Dihidar, S. (2000). Mean square error estimation in multi-stage sampling. *Metrika.* **52**, 115-131.

Chaudhuri, A., Bose, M. and Dihidar, K. (2005). Sample-size-restrictive adaptive sampling: an application in estimating localized elements. *Journal of Statistical Planning and Inference.* **134**, 254-267.

Chaudhuri, A., Bose, M. and Dihidar, K. (2009a). Estimating sensitive proportions by Warner's randomized response technique using multiple randomized responses from distinct persons sampled. Accepted for publication in *Statistical Papers.* Now available online through journal's website: DOI: 10.1007/s00362-009-0210-3.

Chaudhuri, A., Bose, M. and Dihidar, K. (2009b). Estimation of a sensitive proportion by Warner's randomized response data through inverse sampling. Accepted for publication in *Statistical Papers.* Now available online through journal's website: DOI: 10.1007/s00362-009-0234-8.

Chaudhuri, A., Bose, M. and Dihidar, K. (2009c). Rao-Hartley-Cochran sampling with competitive estimators. *6th Triennial Symposium Proceedings Volume, Calcutta Statistical Association Bulletin.* **61**, 227-242.

Chaudhuri, A. Bose, M. and Ghosh, J.K. (2004). An application of adaptive sampling to estimate highly localized population segments. *Journal of Statistical Planning and Inference.* **121**, 175-189.

Chaudhuri, A., Dihidar, K. and Bose, M. (2006). On the feasibility of basing Horvitz & Thompson's estimator on a sample by Rao, Hartley & Cochran's scheme. *Communications in Statistics. Theory and Methods.* **35**, 2039-2044.

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

Chaudhuri, A. and Pal, S. (2008). Estimating sensitive proportions from Warner's randomized responses in alternative ways restricting to only distinct units sampled. *Metrika.* **68**, 147-156.

Chaudhuri, A. and Stenger, H.(2005) Survey Sampling: Theory and Methods (2nd edition). Chapman and Hall, New York.

Chikkagoudar, M.S. (1966). A note on inverse sampling with equal probabilities. *Sankhya, Series A.* **28**, 93-96.

Chikkagoudar, M.S. (1967). Sampling with preliminary random stratification. *Australinal Journal of Statistics.* **9**, 57-60.

Chow, L.P. and Liu, P.T. (1973) A new randomized response technique: the multiple answer model. Department of Population Dynamics, John Hopkins University, Baltimore, Md.

Christofides, T. C. (2003). A generalized randomized response technique. *Metrika.* **57**, 195-200.

Christofides, T. C. (2005). Randomized response in stratified sampling. *Journal of Statistcal Planning and Inference.* **128**, 303-310.

Cliff, A.D. and Ord, J.K. (1973). Spatial autocorrelation. Pion, London.

Cliff, A.D. and Ord, J.K. (1981). Spatial Processes : Models and applications. Pion, London.

Cochran, W.G. (1963, 1977). Sampling Techniques, 2nd and 3rd Edition. John Wiley & Sons, New York.

Cressie, N.A.C. (1989). Geostatistics. *American Statistician.* **43**, 197-202.

Cressie, N.A.C. (1993). Statistics for spatial data. Wiley, New York.

Cressie, N.A.C. and Chan, N. H. (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association.* **84**, 393-401.

Cressie, N.A.C. and Read, T.R.C. (1989) Spatial data analysis of regional counts. *Biometrical Journal.* **31**, 699-719.

Deshpande, M.N. (1984). A note on Rao, Hartley and Cochran's method. *Journal of the Indian Society of Agricultural Statistics.* **36(3)**, 114-116.

Dihidar, K. (2009). Modifying classical randomized response techniques with provision for true response. *Indian Statistical Institute Technical Report.* No. ISI/ASD/2008/8.

Dowling, T.A. and Shachtman, R. (1975). On the relative efficiency of randomized response models. *Journal of the American Statistical Association.* **70**, 84-87.

Duffy, J.C. and Waterton, J.J. (1984). Randomized response models for estimating the distribution function of a quantitative character. *International Statistical Review.* **52**, 165-171.

Eichhorn, B.H. and Hayre, L.S.(1983). Scrambled randomized response method for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference.* **7**, 307-316.

Eriksson, S.A.(1973). A new model for randomized response. *International Statistical Review.* **41**, 101-113.

Folsom, R.E., Greenberg, B.G., Horvitz, D.G. and Abernathy, J.R. (1973). The two alternate questions randomized response model for human surveys. *Journal of American Statistical Association.* **68**, 525-530.

Franklin, L.A. (1989). A comparison of estimators for randomized response sampling with continuous distributions from a dichotomous population. *Communications in Statistics - Theory and Methods.* **18(2)**, 489-505.

Geary, R.C. (1954). The contiguity ratio and statistical mapping. *Incorporated Statistician*, **5**, 115-145.

Ghosh, M. (1987). On admissibility and uniform admissibility in finite population sampling. In *Applied Probability, Stochastic Processes and Sampling Theory.* (MacNeil, I.B. and Umphrey, G.J. eds.) 197-213.

Godambe, V.P. (1960). An admissible estimate for any sampling design. *Sankhya.* **22**, 285-288.

Godambe, V.P. and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations I. *The Annals of Mathematical Statistics.* **36**, 1707-1722.

Godambe, V.P. (1980). Estimation in randomized response trials. *International Statistical Review.* **48**, 29-32.

Gould, A.L., Shah, B.U. and Abernathy, J.R. (1969). Unrelated question randomized response techniques with two trials per respondent. *Proceedings of the Social Statistics Section, American Statistical Association.*

Greenberg, B.G., Abul-Ela, Abdel-Latif, A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question RR model: theoretical framework. *Journal of the American Statistical Association.* **64**, 520-539.

Greenberg, B.G., Kubler, R.R., Abernathy, J.R. and Horvitz, D.G. (1971). Applications of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association.* **66**, 243-250.

Hartley, H.O. and Rao, J.N.K.(1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics.* **33**, 350-374.

Hartley, H.O. and Ross, A.(1954). Unbiased ratio estimators. *Nature.* **174**, 270-271.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association.* **47**, 663-685.

Horvitz, D.G., Shah, B.V. and Simmons, W.R. (1967). The unrelated question RR model. *Proceedings of the Social Statistics Section, American Statistical Association.*, 65-72.

Horvitz, D.G., Shah, B.V. and Abernathy, J.R. (1976). Randomized response: a data gathering device for sensitive questions. *International Statistical Review.* **44**, 181-196.

Jessen, R.J. (1969). Some methods of probability non-replacement sampling. *Journal of the American Statistical Association.* **64**, 175-193.

Korwar, R.M. and Serfling, R.J. (1970). On averaging over distinct units in sampling with replacement. *The Annals of Mathematical Statistics.* **41(6)**, 2132-2134.

Kuk, Anthony Y.C. (1990). Asking sensitive questions indirectly. *Biometrika.* **77(2)**, 436-438.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimators. *Bulletin of the International Statistical Institute.* **33(2)**, 133-140.

Lanke, J. (1975). Some contributions to the theory of survey sampling. *Ph.D. thesis, University of Lund.*

Liu, P.T. and Chow, L.P. (1976). The efficiency of the multiple trial randomized response technique. *Biometrics.* **32**, 607-618.

Mangat, N.S. and Singh, R. (1990). An alternative randomized response procedure. *Biometrika.* **77(2)**, 439-442.

Mangat, N.S., Singh, R., Singh, S., Bellhouse, D.R. and Kashani, H.B. (1995). On efficiency of estimator using distinct respondents in randomized response survey. *Survey Methodology.* **21**, 21-23.

Moors, J.J.A. (1971). Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association.* **66**, 627-629.

Moran, P.A.P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika.* **37**, 17-33.

Mukerjee, R. (1981). Inference on confidential characters from survey data. *Calcutta Statistical Association Bulletin.* **30**, 77-88.

Mukhopadhyay, P. (1996). Inferential problems in survey sampling. New Age International Publishers, Calcutta.

Murthy, M.N. (1967). Sampling theory and methods. Statistical Publishing Society, Calcutta.

Ohlsson, E. (1989). Variance estimation in the Rao-Hartley-Cochran procedure. *Sankhya, Series B.* **51**, 348-361.

Pathak, P.K. (1962). On simple random sampling with replacement. *Sankhya, Series A.* **24**, 287-302.

Raj, Des (1956). Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association.* **51**, 269-284.

Raj, Des (1954). Ratio estimation in sampling with equal and unequal probabilities. *Journal of the Indian Society of Agricultural Statistics.* **6(2)**, 127-138.

Raj, Des (1968). Sampling Theory. Mc-graw Hill, N.Y. Inc.

Raj, Des. and Khamis, Salem H. (1958). Some remarks on sampling with replacement. *Annals of Mathematical Statistics.* **39**, 550-557.

Rao, J.N.K.(1979). On deriving mean square errors and other non-negative unbiased estimators in finite population sampling. *Journal of the Indian Statistical Association.* **17**, 125-136.

Rao, J.N.K., Hartley, H.O., and Cochran, W.G. (1962).On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society.* **24**, 482-491.

Rao, J.N.K. and Vijayan, K. (1977). On estimating the variance in sampling with probability proportional to aggregate size. *Journal of the American Statistical Association.* **72**, 579-584.

Rao, T.J., Sinha, B.K. and Srivenkataramana, T. (2003). On order relations between selection and inclusion probabilities in RHC sampling scheme. *Journal of Applied Statistical Science.* **12(1)**, 67-73.

Salehi, M. M. and Seber, G.A.F. (1997). Adaptive cluster sampling with networks selected without replacement. *Biometrika.* **84(1)**, 209-219.

Salehi, M. M. and Seber, G.A.F. (2002) Unbiased estimators for restricted adaptive cluster sampling. *Australian & New Zealand Journal of Statistics.* **44(1)**, 63-74.

Samiuddin, M. and Asad, H. (1981) A simple procedure of unequal probability sampling. *Biometrika.* **68(3)**, 728-731.

Särndal, C.E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association.* **91 (435)**, 1289-1300.

Särndal, C.E. (1980). A two-way classification of regression estimation strategies in probability sampling. *Canadian Journal of Statistics.* **8 (2)**, 165-177 (1981).

Särndal, C.E., Swenson, B. and Wretman, J.H. (1992). Model assisted survey sampling. Springer-Verlag, New York.

Singh, S. and Joarder, A.H. (1997). Unknown repeated trials in randomized response sampling. *Journal of the Indian Society of Agricultural Statistics.* **50(1)**, 103-105.

Singh, S., Mahmud, M. and Tracy, D.S. (2001). Estimation of mean and variance of stigmatized quantitative variable using distinct units in randomized response sampling. *Statistical Papers.* **42**, 403-411.

Skinner, C.J., Holt, D. and Smith, T.M.F. (eds) (1989). Analysis of complex surveys. Chichester, Wiley.

Tam, S.M. (1988). Asymptotically design-unbiased predictors in survey sampling. *Biometrika.* **75 (1)**, 175-177.

Tamhane, A.C. (1981). Randomized response techniques for multiple sensitive attributes. *Journal of the American Statistical Association.* **76**, 916-923.

Thompson, M.E. (1997). Theory of sample surveys. Chapman & Hall. London.

Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association.* **85**, 1050-1059.

Thompson, S.K. (1992). Sampling. Wiley & Sons. New York.

Thompson, S.K. and Seber, G.A.F. (1996). Adaptive Sampling Wiley & Sons. New York.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association.* **60**, 63-69.

Yates, F. & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the American Statistical Association.* **75,** 206-211.