Revised Version

CONTRIBUTIONS TO EMERGING TECHNIQUES IN SURVEY SAMPLING

Sanghamitra Pal

A Thesis submitted to the Indian Statistical Institute, Kolkata in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Statistics

2003

River Research Institute
Government of West Bengal
Jalasampad Bhavan, Salt Lake
Kolkata - 700 091, India

Acknowledgement and Declaration

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

No part of this thesis was submitted to any other Institute for any degree, diploma, certificate etc. However, this thesis includes materials which in some form constitute contents of the papers jointly with Arijit Chaudhuri entitled

(i) "On Certain alternative mean square error estimators in Complex survey sampling" *Jour. Stat. Plan and Inf.*, 2002, **104**(2), 363-375.

(ii) "On a version of cluster sampling and its practical use", *Jour. Stat. Plan. Inf.*, 2003, **113**(1), 25-34.

(iii) "Estimating proportions from unequal probability samples using randomized responses by Warner's and other devices", *Jour. Ind. Soc. Agri. Stat.*, 2002, **55**(2), 174-183.

(iv) "Systematic sampling: 'Fixed' versus 'Random' sampling interval", *Pak. Jour. Stat.*, 2003, **19**(2), 259-271.

In preparing this Ph.D. thesis I am deeply grateful to my supervisor Professor Arijit Chaudhuri, who taught me survey sampling and helped me in every stage of preparing this.

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

I am also thankful to two anonymous examiners for their helpful comments in preparing this revised write-up.

Sanghamitra Pal

Date : 27/6/03
June, 2003

Sanghamitra Pal
River Research Institute
Government of West Bengal
Jalasampad Bhavan
9th Floor, Western Block
Salt Lake, Kolkata - 91
India

Contributions to Emerging Techniques in Survey Sampling

Sanghamitra Pal

Preface

This dissertation contains seven Chapters. The contents in the respective Chapters may be briefly recounted as follows.

A topic of classical interest in survey sampling is how to ensure the existence of a uniformly non-negative (UNN) unbiased estimator for the mean square error (MSE) of a homogeneous linear estimator (HLE) for a finite survey population total. Hájek (1958), Vijayan (1975) , Rao and Vijayan (1977) and Rao (1979) developed a number of results which boil down to the following as narrated in the monograph by Chaudhuri and Stenger (1992).

If there exist non-zero constants w_i and the unknown values y_i of the variable of interest y be such that for a given sampling design the MSE of an HLE for the finite population total Y, the sum of y_i's over all the N population units, takes the value zero if $\frac{y_i}{w_i}$ is a constant for every $i = 1, \cdots, N$, then the MSE for arbitrary values of y_i's can be expressed in a specific form. When the above-mentioned 'constraint' holds, then this specific MSE-form leads to a specific form of a homogeneous quadratic unbiased estimator (HQUE) for the MSE if it is to have the above noted 'UNN' - property. In the particular case of the Horvitz and Thompson's (1952) estimator the above 'constraint' induces the requirement that every sample with a positive probability of selection should have a constant number of distinct units in it. When this condition is satisfied, Yates and Grundy's (1953) variance estimator is available with simple conditions for its 'UNN-property', which has been examined in the literature to be satisfied for several celebrated schemes of sampling. Two other classical variance estimators for the Horvitz and Thompson's (1952) estimator are given by Horvitz and Thompson (1952) and Ajgaonkar (1967). But no simple results are available to ensure their 'UNN' property. Chaudhuri (2000a) has added a correction term to Yates and Grundy's (1953) variance estimator in case the 'constraint' is relaxed giving simple conditions for the 'UNN'-property of his resulting variances estimator. In the first Chapter of this thesis (i) sampling schemes have been identified satisfying Chaudhuri's (2000a) conditions. Certain theorems have been established to cover general HLE's with relaxed 'constraints'. Further extensions have been implemented to cover multi-stage sampling and randomized response (RR) surveys relevant to this context. The details are presented in the Chapter 1 and have appeared in Chaudhuri and Pal (2002a).

The Chapter 2 deals with a specific version of cluster sampling found appropriate in certain kinds of surveys in practice. In this case the samples vary in size and hence the variance estimators of the kind presented in Chapter 1 turn out relevant. We encounter a practical survey situation when there are two kinds of sampling units -- some are 'small' units like 'Primary health centres' (PHC) in Indian villages and some are "bigger" ones -- the so called BPHC's. Around each BPHC there are a few PHC's that are geographically contiguous and each serves exclusive groups of villagers. Each such set composed of a BPHC along with the allied PHC's in its close proximity may be called a 'cluster'. In a survey implemented by Indian Statistical Institute (ISI) and sponsored by UNICEF in 1998 on infant, child and maternal mortality it was found useful to have representations in samples of both PHC's and BPHC's. In order to have adequate geographical coverage it was found convenient to adopt, within each stratum of villages in a given district of interest, a version of cluster sampling where (a) initially a simple random sample (SRS) of PHC's is to be selected without replacement (WOR), adding (b) to the sample a BPHC from a cluster whenever a PHC from the latter happened to be selected. In spite of this equal probability sampling of the

PHC's the over-all sampling of the clusters involves 'unequal probability selection' with varying sample-sizes. Consequently, a wide variety of choices of estimators is possible leading to the need for new theoretical developments. The details are presented in the Chapter 2. The contents have appeared in Chaudhuri and Pal (2003a).

A simple, yet popular, scheme of sampling is the well-known circular systematic sampling (CSS) with probabilities proportional to sizes (PPS) with a random start but a 'pre-assigned sampling interval', as determined by the consideration of the sample-size required and the known normed size-measures. This scheme has the following well-known disadvantages:
(1) the realized sample-size, in certain situations, falls short of the 'intended one', (2) the inclusion-probabilities of certain units fail to be proportional to the sizes and their ascertainments sometimes become too hard, especially if the size-measures be very high in magnitude as is often the case, (3) inclusion-probabilities of many pairs of units become zero leading to the non-availability of an 'unbiased estimator' for the variance of a linear estimator for a finite population total. Following the earlier approaches by Das (1982) and Ray and Das (1997) we develop encouraging results for a 'modified' CSSPPS scheme that admits a 'random' choice of a sampling interval as a random number between 1 and $(X - 1)$, where X is the aggregated value of all the population size-measures, which are adjusted to be positive integers. In particular, this scheme admits positive inclusion-probabilities for all the pairs of population units. In terms of efficiency also this scheme competes well against an analogous scheme of drawing two 'independent' CSSPPS samples with a number of draws fixed in a comparable way. The details are narrated in Chapter 3. The contents are to appear in Chaudhuri and Pal (2003b).

Many followers of Warner (1965) restrict to sample selection by 'simple random sampling with replacement' (SRSWR) in developing estimators for proportions of people bearing socially stigmatizing characteristics by generating 'randomized responses' (RR) from the people sampled on adopting ingeneous devices. We illustrate how the theory may be extended to cover 'general sampling schemes' by referring to particular RR procedures given by Singh and Joarder (1997), Franklin (1989a, 1989b) and Singh and Singh (1992, 1993). Our developments as follow-ups of Singh and Joarder's (1997) works have appeared in Chaudhuri and Pal (2002b). Such extensions are necessary because it is hard to find sponsors for large-scale social surveys exclusively to tackle 'sensitive issues' and especially by the very restrictive SRSWR selection method. In practice the sampling schemes are more complex, allowing stratification, clustering, selection in two or more stages, with varying probabilities and selection 'without replacement'. Also, in such surveys numerous items of enquiry are covered and only a few of them may relate to sensitive features. So, methods of estimation are needed based on 'direct responses' (DR) relating to 'innocuous' and 'randomized responses' (RR) relating to 'stigmatizing' issues from the same persons selected by a common complex sampling procedure. The subject is dealt with in our Chapter 4.

In survey sampling one procedure used in common to construct a confidence interval (CI) for a parameter like a finite population total is to use a pivot, which is 'an estimator minus the parameter' divided by an estimated standard error of the estimator, treat the pivot as a standard normal deviate - an assumption that may be valid for large samples - and thereby consulting the normal table work out the CI with a pre-assigned confidence coefficient. In terms of the 'generalized regression' (greg) estimator such a procedure is easy to employ. But if one intends to avoid this assumption of normality alternative ways to construct CI's are to employ the bootstrap technique in diverse ways. Treating the greg estimator, motivated by a 'regression modelling through the origin with a single regressor', as a non-linear function of four different Horvitz and Thompson's (HT, 1952) estimators of the population totals of four different variables it is easy to apply Rao and Wu's (1988) bootstrap technique to construct CI's for population totals employing non-linear functions of statistics. But this technique applies only under the two conditions that (1) every sample with a positive selection – probability has a common number of distinct units in it and (2) the product of inclusion probabilities of any two distinct units must not be less than the joint inclusion-probability of these two units. We present modifications needed in methods when either of (1) and (2) is violated at a time.We carry out certain simulation-based numerical calculations for comparative studies in our Chapter 5.

In Chapter 6 we describe a situation where an initial sample may not yield enough observations on a variable of interest to ensure adequate level of efficacy for generalized regression estimators and for empirical Bayes estimators developed therefrom for several non-overlapping domain totals. In such a case we examine how adaptive sampling technique may be suitably applied to effectively enhance the relevant information content by dint of gathering supplementary sample observations by appropriate formation of networks so that the above two types of estimators may acquire improved accuracy levels. We utilize Chaudhuri's (2000a) results concerning 'Adaptive' sampling which we adopt in this context.

Särndal (1996) introduced the striking idea of bypassing the preponderance of the terms involving the 'inclusion-probabilities' of paired units in estimates of variances of Horvitz and Thompson's (1952) estimator and of the generalized regression estimator derived therefrom for a population total because (i) they are hard to calculate and (ii) they often destabilize the variance estimators. Deville (1999), Brewer (1999,2000) and Brewer and Gregoire (2000) have interesting contributions as follow-ups of this approach.

The concluding Chapter 7 pursues with this topic adding a few results and presents a few numerical evaluations for competing estimation procedures.

Contents

0 A Critical Review of the Literature 1

1 Alternative Mean Square Error Estimators in Complex Survey Sampling 12

1.1 Introduction 13

1.2 Alternative MSE estimators 14

1.3 An Illustrative sampling scheme for which the 'Constraint C' does not hold but alternatives to HT, YG Estimators have the 'UNN' property 17

1.4 MSE estimation for greg predictor 20

1.5 Multi-stage sampling and randomized response surveys 24

1.6 A numerical exercise on efficacy in estimation 27

1.6.1 Comments on numerical findings and recommendations 32

2 On a version of cluster sampling and its practical use 34

2.1 Introduction 35

2.2 Estimators and Variance Estimators 38

2.3 Numerical Evaluation of Relative Efficacies of Various Procedures by Simulation 45

2.4 Concluding Remarks and Recommendations 49

3 Systematic sampling: 'Fixed' versus 'Random' Sampling Interval 52

3.1 Introduction 53
3.2 Alternative Circular Systematic Sampling Schemes and Respective Estimators 54
3.2.1 CSSPPS Scheme of Sample Selection 54
3.2.2 CSSPPS with 'random sampling interval' 58
3.3 Conclusion and Recommendation. 68

4 Estimating numbers/proportions of people with stigmatizing features from randomized responses by specific devices through complex survey sampling 69
4.1 Introduction 70
4.2 Unbiased Estimators and Variance Estimators 72
4.3 A Comparative Study with Numerical Illustrations 77
4.4 Repeated randomized response techniques of Franklin (1989a,b) and of Singh and Singh (1992, 1993): 81
4.4.1 Estimation using SRSWR-based repeated RR's 82
4.4.2 Estimation using Repeated RR's in Complex Surveys 87
4.5 Conclusion and recommendation 100

5 Bootstrap Procedures for Generalized Regression Estimators 101
5.1 Introduction 102
5.2 Bootstrap Procedures 103
5.3 Simulation for a numerical study of efficacies 107
5.4 A Discussion on the Simulated Results and Concluding Remarks 110

6 Estimating Domain-wise Distribution of Scarce Objects by Adaptive Sampling and Model-based Borrowing of Strength 112
6.1 Introduction 114
6.2 Sampling and Estimation Methods 115
6.3 Simulation - based Numerical Evaluation of Relative Efficacies 120

6.4 Concluding Remarks and Recommendations 122

7 Simplified variance and mean square error estimation avoiding inclusion-probabilities of paired units 124

7.1 Introduction 125

7.2 Numerical comparison of efficacies of a few alternative estimators of a population total 131

Bibliography 140

Chapter 0

A Critical Review of the Literature

In the traditional design-based approach to the problem of inference in sampling from a finite survey population of a given number of labelled units we are aware of a few well-known 'negative' or 'non-existence' results which briefly are as marked I and II below.

I. In the class of design-unbiased estimators (UE) of a finite population total or mean there does not exist one with the 'uniformly smallest variance' so long as a non-census design is employed. Basu (1971) reinforces this theorem of his by citing his famous 'Circus' or 'elephant' example to point out how the celebrated estimator given by Horvitz and Thompson (HT, 1952) as also, earlier by Narain (1951), ceases to be effective, in spite of its theoretical properties of Admissibility (vide Godambe (1960), Godambe and Joshi (1965)), hyper-admissibility (vide Hanurav (1968)), Necessary bestness (vide Ajgaonkar (1967)) etc. unless it is based on an appropriate sampling design.

II. Godambe (1955) earlier showed the non-existence of a uniformly minimum variance (UMV) estimator of a finite population total in the class of homogeneous linear unbiased estimators (HLUE) for a general class of sampling designs. That this does not apply to what are characterized as 'uni-cluster designs' (UCD) has been pointed out by Hege (1965) and Hanurav (1966).

Stenger (1977) further showed that an exception is also provided by 'informative' and especially 'sequential' sampling designs for which UMV estimator in the HLUE class is available for a population total or mean.

III. That the HT estimator is the unique and hence the UMV estimator in the HLUE class of estimators for a population total when one employs a UCD is demonstrated by Hanurav (1966) by appealing to the concepts of "sufficiency" and 'minimal sufficiency' in survey sampling. Basu and Ghosh (1967) and Basu (1958, 1969) discuss how the combination of the set of distinct units in a sample ignoring their order of appearance along with the unit-wise variate - values constitutes the 'minimal sufficient statistic' based on 'elementary survey data'. Murthy (1957) earlier laid the foundation for the construction of a 'complete class' of statistics by showing that given an unbiased estimator of a finite population parameter that is not a function of this 'minimal sufficient statistic' there is available one with a less variance which is a function of this 'minimal sufficient statistic'. Unfortunately this 'complete class' is not adequately narrow to yield serviceable estimators, many of which remain competitive among themselves. Basu (1971) again pointed out that when the parametric space in the context of inference in finite populations is allowed to be 'wide enough' with each value of a variable of interest permitted to be any real number, thus $\underset{\sim}{0} = (0, 0, \ldots, 0)$, for example, being a possible realization for the vector $\underset{\sim}{Y} = (y_1, \ldots, y_N)$ of values for the N units of a finite population for a variable of interest y, then of course not only the HT estimator $t_H = \sum_{i \in s} \frac{y_i}{\pi_i}$, with $\pi_i(> 0)$ as the inclusion probability of a unit i in a sample s but also the 'infinitely many' others like $t_A = \sum_{i \in s} \frac{y_i - a_i}{\pi_i} + \sum_1^N a_i$, with $\underset{\sim}{A} = (a_1, \ldots, a_N)$ as any element in the parametric space of $\underset{\sim}{Y}$ with real co-ordinates are admissible for $Y = \sum_1^N y_i$. But if the parametric space is narrowed down to a 'close neighbourhood' of $\underset{\sim}{A}$, then t_H is inferior to t_A if $\underset{\sim}{A}$ excludes $\underset{\sim}{0}$ and hence is 'inadmissible'. Yet the HT estimator still occupies a central position in the 'survey sampling' literature though many of its competitors enjoy attention and are also emerging anew supported by modern 'model assistance' and 'calibration' approaches, which we shall briefly narrate in what follows.

About estimators for population totals a major concern is to derive their suitable variance or mean square error (MSE) formulae by way of getting estimators for them as estimators of measures of errors in estimation. More importantly, one needs to get estimators of standard errors (SE) of the estimators of the totals so as to construct plausible confidence intervals (CI) with high enough Confidence Coefficients (CC). In order to achieve this one needs to ensure the MSE - estimators to be non-negative. The problem of getting uniformly non-negative (UNN) unbiased estimators for the MSE of a homogeneous linear estimator (HLE) for a finite population total has been addressed by many survey sampling researchers like Hájek (1958), Sen (1953), Raj (1954), Yates and Grundy (1953), Vijayan (1975), Rao and Vijayan (1977), Rao (1979) among many others. In case one employs a sampling design admitting samples only with a common number of distinct units each, Sen, Yates and Grundy's (SYG) unbiased estimator for the variance of the HT estimator is UNN for many sampling designs for each of which the condition "$\pi_i \pi_j \geq \pi_{ij}$ for every $i, j (i \neq j)$" is satisfied, writing π_{ij} as the inclusion-probability of the pair (i, j) of units in a sample. But no plausible result covering the "unequal sample - size designs" exists in the literature. So, we first address this problem in our Chapter 1 with certain ramifications because a plenty of sampling schemes admitting variable sample sizes often are realistic in practice as we shall presently illustrate in this thesis. In the context of survey sampling we need to briefly describe

(1) Super-population modelling approach (cf. Cochran, 1939, 1946),

(2) Prediction approach (cf. Brewer (1963), Royall (1970)) and

(3) Model assisted approach (cf. Särndal, Swensson and Wretman (1992)).

Since the unified survey sampling theory that treats $\underset{\sim}{Y} = (y_1, \ldots, y_i, \ldots, y_N)$ as a vector of fixed constants is not adequately selective of right strategies for sample-selection and estimation in a suitably optimal way, one way to sort this out is by supposing $\underset{\sim}{Y}$ to be a random vector allowing wide possibilities for the nature of its probability distribution suitably modelled in terms of certain unknown low order moments and possibly with independence or zero correlations.

A population generated by such a probability distribution is called a super-population. An advantage for this is to apply a new criterion of controlling the 'model-expected design-variance' often called the 'Anticipated Variance' (AV) (cf. Isaki and Fuller (1982), Fuller and Isaki (1981)) of a design - unbiased estimator and derive optimal strategies.

A prediction approach starts with $\underset{\sim}{Y}$ as a random vector and consequently Y as a random variable and hence instead of estimating Y attacks the problem of predicting Y by adding $\sum\limits_{i \in s} y_i$ to a predictor of $\sum\limits_{i \notin s} y_i$ based on a suitable modelling of $\underset{\sim}{Y}$. This approach does not need probability sampling. Rather, purposive sampling may yield the most desirable predictor so long as a postulated model is simple as well as correct. Here 'robustness' is an important requirement because one needs a strategy to work in a desirable way not only when a postulated model is tenable but also when it is 'not'. Brewer (1979) and Särndal (1980, 1981, 1982) have recommended going for a predictor which is model unbiased or more importantly is asymptotically design unbiased (ADU) and asymptotically design consistent (ADC) so that it may perform well at least when the sample-size is large irrespective of the correctness or otherwise of the model. This leads to the 'model assisted approach' of Särndal, Swensson and Wretman (SSW, 1992) under which a postulated model "suggests a particular form of an estimator" but its efficiency and accuracy are gauged in terms of its design-based properties of being ADU and/or ADC.

A measure of its error is also design-based and a design-based performance of an estimator of this measure is in question in evaluating its accuracy. A model-based study of a measure of error and its estimation is also of course recommended and pursued with in the literature by many researchers.

A central position is occupied by Cassel, Särndal and Wretman's (CSW, 1976) generalized regression (greg, in brief) estimator, rather predictor for a finite population total in the literature on 'Model assisted survey sampling' approach, a principal reference for which is Särndal, Swensson and Wretman (SSW, 1992). Our work relating to HT estimator in Chapter 1 is also relevant to the Greg estimator which is a generalization upon it. Hence, the Chapter 1 pays attention also to the Greg predictor in a way similar to that paid to the HT and other related homogeneous linear estimators.

Stratified, multi-stage and cluster sampling and related methods of estimation of finite population totals and means are the immediately next steps in the development of the classical theory of design-based estimation covering strategies directly relating to a finite unstratified population of its units. The Chapter 2 deals with a novel method of cluster sampling and related estimation methods necessitated by a specific practical survey sampling situation deemed crucially worthy of attention.

Our proposed cluster sampling permits sample sizes to vary and hence we had a scope to try our methods developed in Chapter 1 to take care of the situation.

Systematic sampling is another classical form of cluster sampling and is traditionally used in Indian 'National Sample Surveys' (NSS) with probabilities of selection of the first stage units (fsu) proportional to their sizes. Here also a pre-assigned sample - size may not be realized, vindicating the role of some of the results in our Chapter 1 in case one employs the HT method of estimation.

Since π_{ij} may be zero for certain pairs (i, j), $i \neq j$ for the traditional PPS (probability proportional to size) systematic sampling with a 'fixed interval', taking a cue from Das (1982) and Ray and Das (1997) we eliminate this shortcoming on allowing 'varying intervals' as is needed in variance estimation Here also sample size need not be a constant calling for our results in Chapter 1. The issues are covered in our Chapter 3, the content of which is now accepted for publication in Pak. Jour. of Stat., Vol. 19(2), 2003, as a joint paper by Chaudhuri, A. and Pal, S.

In multistage sampling the units i in $U = (1, \ldots, i, \ldots, N)$ are treated as the first stage units (fsu) with y_i-values supposed to be unascertainable and is itself to be composed of a certain number M_i of second stage units (ssu) with y_{ij} $(j = 1, \ldots, M_i)$ as the value of the jth second stage unit (ssu) in the ith fsu namely ij with y_i as the sum of all these y_{ij}'s, the ijth ssu in its turn being composed of a certain number of third stage units (tsu) and so on. In order to estimate $Y = \Sigma y_i$ one starts on taking a sample s of fsu's, following up by sampling ssu's independently from the respective fsu's sampled, ascertaining the sampled y_{ij} values and repeating the same procedure in the subsequent stages like-wise if y_{ij}'s themselves are not ascertainable as well.

Thus one may visualize that from such a multistage sample r_i's are available as independent quantities satisfying

(i) $E_L(r_i) = y_i$, (ii) $V_L(r_i) = V_i$ or V_{si} and numbers v_i or v_{si}'s are available such that either (iii) $E_L(v_i) = V_i$ or (iv) $E_L(v_{si}) = V_{si}$, writing E_L, V_L as operators for expectation, variance with respect to sampling at stages 'later' than the first. Then an estimator $e = t(s, r_i | i \in s)$ corresponding to $t = t(s, y_i | i \in s)$ for Y is available along with estimators of variance or MSE of e in terms of $(s, r_i, v_i (i \in s))$ or $(s, r_i, v_{si} (i \in s))$. Raj (1968), Rao (1975) and Chaudhuri, Adhikary and Dihidar (2000) provide the research materials relating to these.

These ideas extend themselves straight forwardly to help one in addressing the problem of handling randomized response (RR) usable in estimating Y when y_i's relate to sensitive issues like earnings by gambling, amounts of taxes evaded, expenses on alcoholism, numbers of days of drunken driving etc. or in estimating proportions of people addicted to stigmatizing habits and practices. Warner (1965) introduced the RR technique to estimate the unknown proportion of people bearing a sensitive characteristic on choosing a simple random sample with replacement (SRSWR) and gathering data from selected persons about their bearing a stigmatizing characteristic A or its complement A^c not as a direct response (DR) but by dint of a randomization device in order to protect a respondent's privacy. A spurt of research ensued as documented by numbers of published papers and a monograph by Chaudhuri and Mukerjee (1988). Estimation of the proportion parameter noted above is mostly based on SRSWR's. But estimation of totals of quantitative variables is based on general sampling schemes as well. Chaudhuri (1987) showed that the above formulation for multistage sampling is available to take care of RR's, replacing E_L, V_L by E_R, V_R respectively namely the expectation and variance operators for 'randomized' response gathering instead of direct questioning.

He could visualize expressing V_i in the form

$$V_i = \alpha_i y_i^2 + \beta_i y_i + \theta_i$$

with $\alpha_i, \beta_i, \theta_i$ as known constants determined by specific RR devices leading to

$$v_i = \frac{1}{1+\alpha_i}(\alpha_i r_i^2 + \beta_i r_i + \theta_i), \text{ provided } 1 + \alpha_i \neq 0$$

as satisfying $E_R(v_i) = V_i$. This yields easy solutions for estimation including variance estimation.

In our Chapter 4 we present a few results concerning RR's applicable under unequal probability sampling, a portion of which is published as a paper by Chaudhuri, A. and Pal, S. in *Jour. Ind. Soc. Agri. Stat.* (2002b).

In constructing a confidence interval (CI) for Y a traditional procedure is to start with a point estimator e for Y along with an estimator for the variance or the MSE of e as say v, and straightaway regard the pivotal quantity

$$\frac{e-Y}{\sqrt{v}}$$

as a standardized normal deviate and hence take $(e-1.96\sqrt{v}, e+1.96\sqrt{v})$ as a 95% CI for Y. Godambe (1998), Rao and Wu (1987) and Woodruff (1952) presented alternatives to this approach including the application of the theory of estimating function.

Another approach is to avoid normality assumption altogether and use bootstrap samples to construct CI's either by (1) Percentile method or (2) the Double boostrap method or other alternatives combining bootstrap statistics with jackknife statistics as recommended by Rao and Wu (1988). Starting with the greg predictor motivated by a linear regression model through the origin for a single regressor as e and v as its estimated MSE as provided by Särndal (1982) for a fixed sample-size design and treating it as a non-linear function of HT estimators for four separate finite population totals one may easily apply the traditional and Rao and Wu's (1988) bootstrap sampling procedures to construct 95% CI's for Y respectively with and without normality assumptions and compare the two procedures on examining the lengths of the respective CI's.

In our Chapter 5 we cover the case of 'non-fixed sample size' designs utilizing developments in our Chapter 1 relevant to this case and modifying Rao and Wu's (1988) bootstrap sampling technique which applies only with fixed sample-size designs in the present context. Our comparison of the CI's by the two alternative approaches is however simulation-based.

In the broad area of survey sampling research one emerging sub-area of promise is concerned with a situation when for many population units y_i's are each zero but there are many other units with y_i's not only positive but also with some of them quite large. Consequently Y is large too but unless the sample captures enough of the latter units its information content may be low and an accurate estimation of Y may be a tough problem.

Thompson (1990, 1992), Thompson and Seber (1996) did a substantial pioneering research in tackling this, introducing their Adaptive sampling technique which is one of capturing further units with positive y_i's by additional sampling starting with an initial one. Chaudhuri (2000) showed how an estimator based on an initial sample with any design may be revised to accomodate additional data gathered from an adaptive sample realized through the initial one. Also Chaudhuri, Bose and Ghosh (2002) demonstrated its efficacy in capturing rare units bearing the features of interest and in deriving efficient estimators.

We already mentioned the role of model-based prediction approach, where no specific form of the distribution of $\underset{\sim}{Y}$ is postulated. If one is prepared as well to postulate a specific parametric model for the random vector $\underset{\sim}{Y}$, for example normality, independence in suitable ways, a simple Bayes estimator for Y is also available starting with an appropriate initial estimator. As this Bayes estimator generally involves unknown parameters one may instead employ empirical Bayes (EB) estimator (EBE) replacing the model parameters by their suitable estimators say, by the method of moments. Fay and Herriot (1979) and Prasad and Rao (1990) provide basic tools for using the EBE's along with estimates of their measures of error.

These methods are particularly appropriate when one intends to effectively estimate not just the population total itself but also the totals of its various non-overlapping segments, called domains on drawing a sample designed without taking account of the domains when in particular there is low representation of the sample observations among some of these domains, called small domains. In order to correct for low efficacy levels for the estimators for these small domains one tries for improvements on borrowing strength from outside the domain - specific parts of the sample but from other parts of the sample on proper postulation of models to reflect affinity among the domains. Thus, instead of using the ordinary greg estimators for the domain totals one may employ their 'synthetic' versions with borrowed data from the entire sample in estimating the postulated regression slope parameters. One may further improve thereupon by the EB technique with further model- specifications. In the Chapter 6 of our thesis we illustrate how adaptive sampling may fare in the context of such domain estimation by synthetic greg and EBE's derived therefrom. Besides the 'model assisted' or 'model-motivated' justification for the greg predictor a 'model free' property is also provided for it in the literature.

To rid the HTE of some of its shortcomings one may replace the weights $\frac{1}{\pi_i}$ of t_H by some other weights $\frac{g_{si}}{\pi_i}$, say, close to $\frac{1}{\pi_i}$ with a well-defined concept of a 'distance' to quantify this closeness. Using some auxiliary observations x_i's well related to y_i's one may impose a condition like

$$\sum \frac{x_i}{\pi_i} g_{si} = X,$$

which is called a 'calibration equation', which is somewhat a 'side condition on the weights' to be chosen.

One possible choice is

$$g_{si} = 1 + \left(X - \sum_{i \in s} \frac{x_i}{\pi_i} \right) \frac{\sum_{i \in s} y_i x_i Q_i}{\sum_{i \in s} x_i^2 Q_i}$$

for many suitable choices of $Q_i (> 0)$.

But $\sum_{i \in s} \frac{y_i}{\pi_i} g_{si}$ is a 'greg' estimator and this is a 'calibration' estimator with no 'backing up' by any super-population modelling.

This is an additional qualification for the greg estimator for it to remain in a central position. Because the greg estimator is a calibration estimator and hence with an interpretation as an exclusively design-based estimator with no appeal to any model we may be satisfied with the design-based estimators for its mean square error in constructing a CI for Y based on it as a point estimator.

Särndal (1996), Deville (1999) and others propagated the idea that a usual estimator of the variance of the HT estimator may be unstable as it involves too many cross-product terms with coefficients which are difficult to calculate as they involve π_{ij}'s which for many sampling schemes are hard to compute. So, Hájek's (1981) Poisson sampling scheme provides a relief as it is free of cross product terms because $\pi_i \pi_j - \pi_{ij} = 0$ for this scheme. But as the sample size for this scheme may vary widely over 0 to N, instead of the HTE, some other estimators need to be used to ensure stability in estimation.

One such proposed estimator turns out to be a particular form of a greg estimator. Brewer (1999, 2000) and Brewer and Gregoire (2000) introduced certain approximations connecting π_i and π_{ij}'s. Hartley and Rao's (1962) scheme involving PPS Circular Systematic Sampling with a prior random permutation of the population units also provides some approximate formulae for π_{ij}'s which all turn out positive.

In the Chapter 7 of our thesis we examine numerically the performance of several of these alternatives including a few suggested by ourselves.

Chapter 1

Alternative Mean Square Error Estimators in Complex Survey Sampling

Abstract

JNK Rao (1979) gave a 'necessary form' for an unbiased mean square error (MSE) estimator to be 'uniformly non-negative' (UNN). The MSE is of a homogeneous linear estimator (HLE) 'subject to a specified constraint', for a survey population total of a real variable of interest. A corresponding theorem is presented when the 'constraint' is relaxed. Certain results are added presenting formulae for estimators of MSE's when the variate-values for the sampled individuals are not ascertainable. Though not ascertainable, they are supposed to be suitably estimated either by (1) randomized response techniques covering sensitive issues or by (2) further sampling in subsequent stages in specific ways when the initial sampling units are composed of a number of sub-units, rather 'subsequent stage units'.

1.1 Introduction

Let us consider a survey population $U = (1, \cdots, i, \cdots, N)$ of a known number N of identifiable units labeled $i = 1, \cdots, N$. On it is defined a real variable of interest y with values y_i with a population total $Y = \Sigma y_i$, writing Σ to denote sum over i in U. We suppose that a sample s is drawn from U with a probability $p(s)$ and the values y_i are ascertained for the units i in s. A homogeneous linear estimator (HLE) for Y is to be employed. For such an estimator written as

$$t_b = \Sigma y_i b_{si} I_{si} \tag{1.1}$$

with b_{si}'s as constants free of $\underline{Y} = (y_1, \cdots, y_i, \cdots, y_N)$ and $I_{si} = 1$ if $i \epsilon s; 0$ if $i \not\epsilon s$; then the MSE is

$$M(t_b) = E_1(t_b - Y)^2 = \Sigma\Sigma d_{ij} y_i y_j. \tag{1.2}$$

Here, E_1 denotes expectation with respect to the design p corresponding to $p(s)$ above; $\Sigma\Sigma$ denotes sum over i, j in U with no restrictions :

$$d_{ij} = E_1(b_{si} I_{si} - 1)(b_{sj} I_{sj} - 1).$$

Rao (1979) considered a sub-class of estimators t_b in (1.1) for which the following

"Condition, say C" holds:

'If there exist $w_i (\neq 0)$ as constants free of $\underline{Y}$, then

$$M(t_b) = 0, \text{ if } z_i = \frac{y_i}{w_i} \text{ for every } i \text{ in } U \text{ is a constant'}. \tag{1.3}$$

Under "C" it follows that $M(t_b)$ may be written as

$$M(t_b) = -\sum_{i<j}\sum d_{ij} w_i w_j \left(\frac{y_i}{w_i} - \frac{y_j}{w_j}\right)^2; \tag{1.4}$$

here $\underset{i<j}{\Sigma\Sigma}$ denotes sum over $i, j (i < j)$ in U. This was enunciated by Rao (1979) following the approach of Hájek (1958). Rao (1979) then deduced that in the class of homogeneous quadratic unbiased estimators (HQUE) of $M(t_b)$ one that may be uniformly non-negative (UNN) is "necessarily of the form"

$$m(t_b) = -\underset{i<j}{\Sigma\Sigma} d_{sij} I_{sij} w_i w_j \left(\frac{y_i}{w_i} - \frac{y_j}{w_j} \right)^2; \tag{1.5}$$

here d_{sij}'s are constants free of $\underline{Y}$ and $I_{sij} = I_{si} I_{sj}$

$$\text{subject to } E_1(d_{sij} I_{sij}) = d_{ij} \text{ for every } i, j \text{ in } U. \tag{1.6}$$

This is Rao's (1979) theorem as a further development following Vijayan's (1975) and Rao and Vijayan's (1977) earlier works. In the next section we present certain results when the 'Constraint C' is relaxed.

1.2 Alternative MSE estimators

Rao (1979) illustrated several classical sampling strategies for which the above theory applies. For example, by denoting $\pi_i = \underset{s}{\Sigma} p(s) I_{si}$,

writing $\underset{s}{\Sigma}$ as the sum over all samples s with $p(s) > 0$, as the inclusion-probability of i, in a sample, the Horvitz and Thompson's (HT, 1952) estimator (HTE) given by $t_H = \Sigma y_i \frac{I_{si}}{\pi_i}$, assuming "$\pi_i > 0$ for every i in U - a necessary condition for the existence of a design-unbiased estimator for Y", satisfies 'C' if '$\nu(s) = \Sigma I_{si}$, the number of distinct units in s, is a constant for every s with $p(s) > 0$'.

Since t_H is unbiased for Y, its MSE equals its variance which is, writing $\pi_{ij} = \Sigma p(s) I_{sij}$, as given by HT (1952), on taking $I_{sij} = I_{si} I_{sj}$,

$V_1(t_H) = \Sigma y_i^2 \frac{1-\pi_i}{\pi_i} + 2 \underset{i<j}{\Sigma\Sigma} y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$, in line with (1.2). If $p(s) > 0 \Rightarrow \nu(s)$ is a constant, then 'C' holds and in accordance with (1.4), $V_1(t_H)$ equals

$$V_2(t_H) = \underset{i<j}{\Sigma\Sigma} (\pi_i \pi_j - \pi_{ij}) (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2$$

a form of the variance of t_H as given by Yates and Grundy (1953). Accordingly, if 'C' holds, t_H has its unbiased variance estimator given by Yates and Grundy (YG, 1953), as

$$v_{YG} = \sum_{i<j}\sum(\pi_i\pi_j - \pi_{ij})(\tfrac{I_{sij}}{\pi_{ij}})(\tfrac{y_i}{\pi_i} - \tfrac{y_j}{\pi_j})^2, \text{ assuming } \pi_{ij} > 0 \quad \forall\ i, j.$$

This is 'UNN' if $\pi_i\pi_j \geq \pi_{ij} \quad \forall i, j$.

If 'C' does not hold, then $V_1(t_H)$ may be unbiasedly estimated by

$$v_{HT} = \Sigma y_i^2(\frac{1-\pi_i}{\pi_i})\frac{I_{si}}{\pi_i} + 2\sum_{i<j}\sum y_i y_j(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j})\frac{I_{sij}}{\pi_{ij}}, \qquad (2.1)$$

as was proposed by HT (1952).

Unfortunately, it is difficult to work out conditions for 'UNN' property of v_{HT}, except noting that π_i, π_{ij}'s should be such that v_{HT} must be a "non-negative definite" as a quadratic form in terms of y_i's for i in s for every s with $p(s) > 0$. This is too hard a condition to check for most sampling designs.

We intend to provide another alternative unbiased variance estimator for t_H when 'C' fails and yet it is easy to present conditions for its 'UNN' property. Before that we present a theorem covering the general estimator t_b for which 'C' fails.

Theorem 1. Let there exist non-zero constants w_i free of $\underline{Y}$, for $i \epsilon U$. Then, writing $z_i = \frac{y_i}{w_i}$, it follows that

$$M(t_b) = -\sum_{i<j}\sum d_{ij}w_iw_j(z_i - z_j)^2 + \Sigma\frac{y_i^2}{w_i}\alpha_i, \text{ where } \alpha_i = \sum_{j=1}^{N} d_{ij}w_j. \qquad (2.2)$$

Proof: $-\sum_{i<j}\sum d_{ij}w_iw_j(z_i - z_j)^2 = -\frac{1}{2}\sum_{i\neq j}\sum d_{ij}w_iw_j(\frac{y_i^2}{w_i^2} + \frac{y_j^2}{w_j^2} - \frac{2y_iy_j}{w_iw_j})$

$$= \sum_{i\neq j}\sum d_{ij}y_iy_j - \sum_i \frac{y_i^2}{w_i}(\sum_{j=1}^{N} d_{ij}w_j - d_{ii}w_i)$$

$$= \Sigma\Sigma d_{ij}y_iy_j - \sum_i \frac{y_i^2}{w_i}\alpha_i.$$

This completes the proof with one obvious step.

Corollary 1. Two unbiased estimators of $M(t_b)$ immediately suggested by Theorem 1 are:

$$m_1(t_b) = -\sum\sum_{i<j} d_{sij} I_{sij} w_i w_j (\frac{y_i}{w_i} - \frac{y_j}{w_j})^2 + \Sigma \frac{y_i^2}{w_i} \alpha_i \frac{I_{si}}{\pi_i}$$

with d_{sij} subject to (1.6) and

$$m_2(t_b) = -\frac{1}{p(s)} [\sum\sum_{i<j} c_{sij} w_i w_j (\frac{y_i}{w_i} - \frac{y_j}{w_i})^2 - \Sigma \frac{y_i^2}{w_i} \alpha_i c_{si}]$$

writing $c_{sij} = \frac{I_{sij} d_{ij}}{\sum\limits_s I_{sij}}, c_{si} = \frac{I_{si}}{\sum\limits_s I_{si}}$.

This $m_2(t_b)$ is motivated somewhat by Ajgaonkar's (1967) estimator of $V(t_H)$, which is $v_A(t_H) = \frac{1}{p(s)}[\frac{1}{\binom{N-1}{n-1}} \Sigma y_i^2 I_{si} + \frac{1}{\binom{N-2}{n-2}} \sum\sum_{i<j} y_i y_j I_{sij}]$ when $\nu(s) = n \forall s$ with $p(s) > 0$.

Remark I. Conditions for the 'UNN' properties of $m_1(t_b)$ and $m_2(t_b)$ are obviously

(i) $w_i w_j d_{sij} I_{sij} \leq 0, w_i \alpha_i I_{si} \geq 0$ for
the former and

(ii) $w_i w_j c_{sij} \leq 0, w_i \alpha_i I_{si} \geq 0$ for the latter.

Remark II. $\pi_i > 0 \quad \forall\, i \Rightarrow \sum\limits_s I_{si} > 0 \;\forall\; i$

and $\pi_{ij} > 0 \quad \forall i, j \Rightarrow \sum\limits_s I_{sij} > 0 \;\forall\; i, j$.

The easy proofs are omitted. The Corollary 2 given below was given by Chaudhuri (2000a).

Corollary 2. If 'C' does not hold, writing $\nu = \Sigma \nu(s) p(s)$, the variance of t_H equals, on allowing $\nu(s)$ to vary with s, $p(s) > 0$,

$$V_3(t_H) = \sum\sum_{i<j} (\pi_i \pi_j - \pi_{ij})(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 + \Sigma \frac{y_i^2}{\pi_i} \beta_i \tag{2.3}$$

writing

$$\beta_i = (1 + \frac{1}{\pi_i} \sum_{j \neq i} \pi_{ij} - \nu), i \epsilon U. \tag{2.4}$$

Proof: Easy, on recalling $\nu = \Sigma\pi_i$.

Corollary 3. Two unbiased estimators of $V_3(t_H)$ are

$$v_1(t_H) = \underset{i<j}{\Sigma\Sigma}\frac{I_{sij}}{\pi_{ij}}(\pi_i\pi_j - \pi_{ij})(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 + \Sigma\frac{I_{si}}{\pi_i}\frac{y_i^2}{\pi_i}\beta_i \tag{2.5}$$

and

$$v_2(t_H) = \frac{1}{p(s)}[\underset{i<j}{\Sigma\Sigma}I_{sij}(\frac{\pi_i\pi_j - \pi_{ij}}{\sum_s I_{sij}})(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 + \Sigma c_{si}\frac{y_i^2}{\pi_i}\beta_i] \tag{2.6}$$

Remark III. Conditions for the 'UNN' properties of both of $v_1(t_H)$ and $v_2(t_H)$ are:

"$\pi_i\pi_j \geq \pi_{ij} \quad \forall\ i, j(i \neq j), \beta_i \geq 0 \quad \forall i$".

In section 1.3 we illustrate situations when (A) $\nu(s)$ varies with s when $p(s) > 0$, but the (B) conditions for 'UNN' properties of $v_1(t_H)$ and $v_2(t_H)$ hold.

1.3 An Illustrative sampling scheme for which the 'Constraint C' does not hold but alternatives to HT, YG Estimators have the 'UNN' property

Brewer (1963) gave a sampling scheme when normed size-measures $p_i(0 < p_i < 1 \quad \forall i, \Sigma p_i = 1)$are available for $i\epsilon U$. Here, on the first draw, the unit i is chosen with a probability proportional to $q_i = \frac{p_i(1-p_i)}{(1-2p_i)}$ and leaving aside the unit i so chosen, a second unit $j(\neq i)$ is chosen in the second draw with a probability $\frac{p_j}{1-p_i}$. Writing

$D = \Sigma(\frac{p_i}{1-2p_i})$ he showed that for this scheme the 'inclusion-probability' $\pi_i(2)$ for i equals $2p_i$ and 'that' for (i, j), denoted by $\pi_{ij}(2)$ equals $\frac{2p_ip_j}{(1+D)}(\frac{1}{1-2p_i} +$

$\frac{1}{1-2P_j}$). It is further known that

$$\Delta_{ij}(2) = \pi_i(2)\pi_j(2) - \pi_{ij}(2) \geq 0 \quad \forall i, j(i \neq j) \text{ in } U. \tag{3.1}$$

We use '2' within parentheses to emphasize that this scheme uses 2 draws. Let the sample chosen as above be augmented by adding to the 2 distinct units so drawn as above, $(r-2)$ further distinct units from the remaining $(N-2)$ units of U by simple random sampling (SRS) without replacement (WOR). For such a scheme introduced by Seth (1966) admitting r distinct units in each sample, the inclusion-probabilities $\pi_i(r)$ for i and $\pi_{ij}(r)$ for $(i,j)(i \neq j)$, involving $r(>2)$ draws, are respectively

$$\begin{aligned}
\pi_i(r) &= \frac{1}{N-2}[(r-2) + (N-r)\pi_i(2)], \\
\pi_{ij}(r) &= \pi_{ij}(2) + (\frac{r-2}{N-2})(\pi_i(2) + \pi_j(2) - 2\pi_{ij}(2)) \\
&\quad +(\frac{r-2}{N-2})(\frac{r-3}{N-3})(1 - \pi_i(2) - \pi_j(2) + \pi_{ij}(2)).
\end{aligned}$$

Let us slightly modify this sampling scheme of Seth (1966) by allowing $(r-2)$ to be (1) a number $(n-2)$ to be chosen with a pre-assigned probability $w(0 < w < 1)$ and (2) a number $(n-1)$ to be chosen with the complementary probability $(1-w)$. Then a sample s so drawn will have a size 'n with probability w' and '$(n+1)$ with probability $(1-w)$'. Thus $\nu(s)$ is either n or $(n+1)$. Putting $n, (n+1)$ by turn in $\pi_i(r), \pi_{ij}(r)$ above we get the inclusion probabilities π_i^*, say, for i and π_{ij}^* for (i,j) for this 'modified Seth's sampling scheme', as

$$\begin{aligned}
\pi_i^* &= w\pi_i(n) + (1-w)\pi_i(n+1) \\
\text{and } \pi_{ij}^* &= w\pi_{ij}(n) + (1-w)\pi_{ij}(n+1).
\end{aligned}$$

Then, we have

Theorem 2. $\pi_i^*\pi_j^* \geq \pi_{ij}^* \quad \forall i, j \ (i \neq j)$ in U.

Proof: On simplifications we get, using the results of this section,

$$\begin{aligned}\pi_i^*\pi_j^* &= \pi_i(2)\pi_j(2)+\frac{(n-1-w)}{(N-2)}(\pi_i(2)+\pi_j(2)-2\pi_i(2)\pi_j(2)) \\ &\quad +[\frac{(n-1-w)}{(N-2)}]^2(1-\pi_i(2)-\pi_j(2)+\pi_i(2)\pi_j(2))\end{aligned} \tag{3.2}$$

and

$$\begin{aligned}\pi_{ij}^* &= \pi_{ij}(2)+\tfrac{(n-1-w)}{(N-2)}(\pi_i(2)+\pi_j(2)-2\pi_{ij}(2)) \\ &\quad +\frac{(n-2)(n-1-2w)}{(N-2)(N-3)}(1-\pi_i(2)-\pi_j(2)+\pi_{ij}(2)).\end{aligned} \tag{3.3}$$

Subtracting (3.3) from (3.2) and using (3.1), on further simplications the theorem is immediately proved.

Next, let us note the

Lemma. For any design p

$$\sum_{s\ni i}\nu(s)p(s)=\sum_{j\neq i}\pi_{ij}+\pi_i$$

Proof : $\pi_{ij}=\sum_s p(s)I_{sij}$. So, $\sum_{j\neq i}\pi_{ij}=\sum_s p(s)(\nu(s)-1)I_{si}=\sum_{s\ni i}p(s)\nu(s)-\pi_i$. This gives the result.

Incidentally, using this Lemma, β_i in (2.4) may be alternatively written also as

$$\beta_i=\frac{1}{\pi_i}\sum_{s\ni i}\nu(s)p(s)-\nu,\ i\in U. \tag{3.4}$$

For the scheme of sampling given by Seth (1966) as modified above, the formula for β_i turns out to be

$\beta_i^*=1+\frac{1}{\pi_i^*}\sum_{j\neq i}\pi_{ij}^*-\Sigma\pi_i^*.$

The above Lemma yields

Theorem 3. $\beta_i^*>0\ \forall\ i\epsilon U.$

Proof: $\sum\limits_{s\ni i} p(s)\nu(s) = wn\pi_i(n) + (1-w)(n+1)\pi_i(n+1)$. So,

$$\begin{aligned}\beta_i^* &= \frac{1}{\pi_i^*}[wn\pi_i(n) + (1-w)(n+1)\pi_i(n+1)] - [wn + (1-w)(n+1)] \\ &= \frac{1}{\pi_i^*}[w(1-w)(\pi_i(n+1) - \pi_i(n)] > 0\end{aligned}$$

because $\pi_i(n+1) - \pi_i(n) = \frac{(1-\pi_i(2))}{(N-2)} > 0$.

Since for this scheme $\nu(s)$ is either n with a probability w or $(n+1)$ with a probability $(1-w)$, Rao's (1979) 'Constraint C' here is violated. Yet, for t_H based on this scheme, our proposed new estimators $v_1(t_H), v_2(t_H)$ for $V_3(t_H)$ are 'uniformly non-negative'.

It seems to us that the celebrated generalized regression (greg) estimator or predictor given by Cassel, Särndal and Wretman (CSW, 1976) for Y needs a discussion in the present context. Let us take it up below.

1.4 MSE estimation for greg predictor

Let $\underline{X} = (x_1, \cdots, x_i, \cdots, x_N)$ with x_i as the positive value of an auxiliary variable x for i in U with a known total $X = \Sigma x_i$. Then, choosing numbers R_i suitably for example as $\frac{1}{x_i}$ or $\frac{1}{x_i^2}$ or $\frac{1}{\pi_i x_i}$ or $\frac{1-\pi_i}{\pi_i x_i}$ or $\frac{1}{x_i^g}$, (with suitably specified g in (0,2)) etc., the greg predictor for Y given by CSW (1976) is

$$t_g = \Sigma \frac{y_i}{\pi_i} I_{si} g_{si} = \Sigma \frac{y_i}{\pi_i} I_{si} + b_R(X - \Sigma \frac{x_i}{\pi_i} I_{si}),$$

writing $g_{si} = 1 + (X - \Sigma \frac{x_i}{\pi_i} I_{si}) \frac{x_i \pi_i R_i}{\Sigma x_i^2 R_i I_{si}}$,

$$b_R = \frac{\Sigma y_i x_i R_i I_{si}}{\Sigma x_i^2 R_i I_{si}}.$$

Let

$$B_R = \frac{\Sigma y_i x_i R_i \pi_i}{\Sigma x_i^2 R_i \pi_i}, e_i = y_i - b_R x_i, E_i = y_i - B_R x_i.$$

Then, t_g is a (design-) biased predictor for Y but its bias may be neglected for large

samples. Assuming large samples and applying Taylor series neglecting terms involving second and higher order derivatives, the following formulae for MSE of t_g and estimators of MSE are well-known, especially from Särndal, Swensson and Wretman (SSW,1992).

$$M_1 = M(t_g) = \Sigma E_i^2 \tfrac{1-\pi_i}{\pi_i} + 2\underset{i<j}{\Sigma\Sigma} E_i E_j \tfrac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}$$

$=$ variance of $\Sigma \frac{E_i I_{si}}{\pi_i}$, vide HT (1952),

$M_2 = M(t_g) = \underset{i<j}{\Sigma\Sigma}(\pi_i\pi_j - \pi_{ij})(\frac{E_i}{\pi_i} - \frac{E_j}{\pi_j})^2$, following YG (1953), applicable

when '$\nu(s)$ is a constant for every s with $p(s) > 0$' – an example where Rao's (1979) 'constraint C' is imposed taking

(i) $E_i = a\pi_i \forall i$ with 'a' as a constant and

(ii) $\nu(s) = \nu$ for every s with $p(s) > 0$.

Estimators of M_1 are well-known (vide SSW (1992)) to be

$$m_{kg} = \Sigma(a_{ki}e_i)^2(\frac{1-\pi_i}{\pi_i})\frac{I_{si}}{\pi_i} + 2\underset{i<j}{\Sigma\Sigma}(a_{ki}e_i)(a_{kj}e_j)(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j})\frac{I_{sij}}{\pi_{ij}}, \tag{4.1}$$

$k = 1, 2;\ a_{1i} = 1, a_{2i} = g_{si}.$

Estimators of M_2 are well-known to be

$$r_{kg} = \underset{i<j}{\Sigma\Sigma}(\pi_i\pi_j - \pi_{ij})((\frac{a_{ki}e_i}{\pi_i} - \frac{a_{kj}e_j}{\pi_j})^2)\frac{I_{sij}}{\pi_{ij}}, k = 1, 2. \tag{4.2}$$

For m_{kg}, conditions for 'uniform non-negativity' are difficult to check, but they are usable even if $\nu(s)$ varies with s. On the contrary, r_{kg} is 'UNN' if $\pi_i\pi_j \geq \pi_{ij}$ for $i \neq j$, but its use is recommended if '$\nu(s)$ is a constant for every s with $p(s) > 0$'. If $\nu(s)$ varies with s we have the following alternative approximate formula for $M(t_g)$ based on Taylor series expansion:

$$M_3 = M(t_g) = \sum_{i<j}\sum(\pi_i\pi_j - \pi_{ij})((\frac{E_i}{\pi_i} - \frac{E_j}{\pi_j})^2) + \Sigma\frac{E_i^2}{\pi_i}\beta_i \tag{4.3}$$

To see this let us write

$t_1 = \Sigma\frac{y_i}{\pi_i}I_{si}, t_2 = \Sigma\frac{x_i}{\pi_i}I_{si}, t_3 = \Sigma y_i x_i R_i I_{si}, t_4 = \Sigma x_i^2 R_i I_{si}, \underline{t} = (t_1, t_2, t_3, t_4)$,
$\theta_1 = \Sigma y_i = Y, \theta_2 = \Sigma x_i = X, \theta_3 = \Sigma y_i x_i R_i \pi_i, \theta_4 = \Sigma x_i^2 R_i \pi_i, \underline{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$.

Then, $E_1(t_j) = \theta_j,\ j = 1, \cdots, 4$,

Also, $t_g = t_1 + (X - t_2)\frac{t_3}{t_4} = f(\underline{t})$, say;

$f(\underline{\theta}) = Y$.

$\lambda_1 = \frac{\delta}{\delta t_1}f(\underline{t})|_{\underline{t}=\underline{\theta}} = 1, \lambda_2 = \frac{\delta}{\delta t_2}f(\underline{t})|_{\underline{t}=\underline{\theta}} = -B_R, \lambda_3 = \frac{\delta}{\delta t_3}f(\underline{t})|_{\underline{t}=\underline{\theta}} = 0$,

$\lambda_4 = \frac{\delta}{\delta t_4}f(\underline{t})|_{\underline{t}=\underline{\theta}} = 0$.

Then, $t_g = f(\underline{t})$ may be approximated by

$$\begin{aligned} f(\underline{\theta}) + \lambda_1(t_1 - \theta_1) + \lambda_2(t_2 - \theta_2) &= Y + (t_1 - Y) - B_R(t_2 - \theta_2) \\ &= Y + (\Sigma E_i\frac{I_{si}}{\pi_i} - \Sigma E_i) \end{aligned}$$

Then, $M(t_g)$ is approximately equal to

$M_3(t_g) = E_1[\Sigma E_i\frac{I_{si}}{\pi_i} - \Sigma E_i]^2$

which is the variance of $\Sigma E_i\frac{I_{si}}{\pi_i}$.

Now applying Corollary 2, we get analogously to (2.2),

$$M_3(t_g) = \sum_{i<j}\sum(\pi_i\pi_j - \pi_{ij})((\frac{E_i}{\pi_i} - \frac{E_j}{\pi_j})^2) + \Sigma\frac{E_i^2}{\pi_i}\beta_i$$

So, our proposed estimators for $M_3(t_g)$ are

$$\begin{aligned} v_k(t_g) = & \sum\sum_{i<j}(\pi_i\pi_j - \pi_{ij})((\frac{a_{ki}e_i}{\pi_i} - \frac{a_{kj}e_j}{\pi_j})^2)\frac{I_{sij}}{\pi_{ij}} \\ & + \Sigma(a_{ki}e_i)^2\beta_i\tfrac{I_{si}}{\pi_i^2}, \\ k = & \; 1, 2. \end{aligned} \tag{4.4}$$

By way of justification of $v_k(t_g)$ we may observe that

$g_{si}e_\imath = \{1 + (X - \Sigma x_i \frac{I_{si}}{\pi_i})(\frac{x_i\pi_i R_i}{\Sigma x_i^2 R_i I_{si}})\}(y_i - \frac{\Sigma y_i x_i R_i I_{si}}{\Sigma x_i^2 R_i I_{si}} x_i)$

may be approximated, through Taylor Series expansion, by $E_i, i\epsilon U$ on approximating t_j by $\theta_j, j = 2, 3, 4$. The rest follows as in corollary 2 in Section 1.2.

Next we consider application of the above results when y_i is not directly ascertainable. This happens, for example, (1) when y relates to a sensitive characteristic like number of days of drunken driving, amount of tax evaded etc so that it is embarrassing to ask for and get direct responses (DR) concerning such questions and instead 'randomized responses' (RR) by dint of ingeneous devices may be generated to yield 'estimates' for y_i, i in s and (2) when i itself contains a large number of further sub-units or second-stage units of which only a sample may be observed or more generally multi-stage sampling may seem feasible in a given context. We present relevant results in the next section. In these 2 cases (1) and (2), appropriate devices and designs are employed by an investigator to derive suitable estimatiors for $y_i, i \in s$ to be subsequently used in estimating $Y = \Sigma y_i$. A third possibility of a super-population model-based approach of dealing with the situation when y_i is subject to measurement and observational errors as treated by Fuller (1987) and Bolfarine and Zacks (1992) among others is not considered here.

1.5 Multi-stage sampling and randomized response surveys

Let E_2 denote the operator for taking expectation and V_2 that for variance with respect either to (a) randomized response (RR) gathering from a person in the population or (b) sampling at later stages of sampling. Let 'independent' observations r_i be available in either case along with sample-based observations v_i such that

$$(i) \quad E_2(r_i) = y_i, (ii) \quad V_2(r_i) = V_i \text{ and } (iii) \quad E_2(v_i) = V_i, i \epsilon U. \qquad (5.1)$$

We may further assume that

$$E_1 E_2 = E_2 E_1. \qquad (5.2)$$

Raj (1968), Chaudhuri (1987) considered this set-up in the contexts respectively, of multi-stage and RR given in (5.1). Chaudhuri, Adhikary and Dihidar (2000) use (5.2). Let E, V denote the over-all expectation, variance operators. Then, $E = E_1 E_2 = E_2 E_1$ and

$V = E_1 V_2 + V_1 E_2 = E_2 V_1 + V_2 E_1$ by (5.2).

Writing any of the above-noted estimators or predictors for Y based on $y_i, i \epsilon s$ as

$$t = t(s, \underline{Y}),$$

we may write

$$e = e(s, \underline{r}),$$

where $\underline{r} = (r_1, \cdots, r_i, \cdots, r_N)$, to denote the function

$t(.,.)$ in which y_i is replaced by r_i for $i \epsilon s$. We shall next write $R = \Sigma r_i$ and $\underline{V} = (v_1, \cdots, v_i, \cdots, v_N)$. Then, the MSE of e about Y will be taken as

$$M_1^*(e) = E(e - Y)^2 = E_1 E_2[(e - t) + (t - Y)]^2 = E_1 E_2 (e - t)^2 + M(t) \quad (5.3)$$

where $M(t) = E_1(t - Y)^2$,

and

$$\begin{aligned} M_2^*(e) = E(e - Y)^2 &= E_2E_1[(e - R) + (R - Y)]^2 \\ &= E_2M(e) + \Sigma V_i \end{aligned} \tag{5.4}$$

Remark IV. $M(e) = E_1(e - R)^2 = E_1(t - Y)^2|_{\underline{Y}=\underline{R}} = M(t)|_{\underline{Y}=\underline{R}}$,
$e_b = t_b|_{\underline{Y}=\underline{R}}, M(e_b) = M(t_b)|_{\underline{Y}=\underline{R}} = -\underset{i<j}{\Sigma\Sigma} d_{ij}w_iw_j(\frac{r_i}{w_i} - \frac{r_j}{w_j})^2 + \Sigma\frac{r_i^2}{w_i}\alpha_i$, from (2.2),
$e_H = t_H|_{\underline{Y}=\underline{R}}$,
$V_3(e_H) = V_3(t_H)|_{\underline{Y}=\underline{R}} = \underset{i<j}{\Sigma\Sigma}(\pi_i\pi_j - \pi_{ij})(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j})^2 + \Sigma\frac{r_i^2}{\pi_i}\beta_i$ from (2.3),
$e_g = t_g|_{\underline{Y}=\underline{R}}$,
$M_3(e_g) = M_3(t_g)|_{\underline{Y}=\underline{R}} = \underset{i<j}{\Sigma\Sigma}(\pi_i\pi_j - \pi_{ij})((\frac{E_i(r)}{\pi_i} - \frac{E_j(r)}{\pi_j})^2) + \Sigma\frac{E_i^2(r)}{\pi_i}\beta_i$, writing $E_i(r) = E_i|_{\underline{Y}=\underline{R}}$, from (4.3).

Next, let us write

$$e_i(r) = e_i|_{\underline{Y}=\underline{R}} = r_i - b_R(r)x_i, \tag{5.5}$$

writing

$$b_R(r) = \frac{\Sigma r_ix_iR_iI_{si}}{\Sigma x_i^2R_iI_{si}} = b_R|_{\underline{Y}=\underline{R}}.$$

Our proposed estimators of MSE's are the following:

For $M_1^*(e_b)$ the proposed unbiased estimators are:

$$\begin{aligned} m_1^*(e_b) &= m_1(e_b) + \underset{i<j}{\Sigma\Sigma} d_{sij}I_{sij}w_iw_j(\frac{v_i}{w_i^2} + \frac{v_j}{w_i^2}) \\ &\quad -\Sigma\frac{v_i}{w_i}\alpha_i\frac{I_{si}}{\pi_i} + \Sigma v_ib_{si}^2I_{si} \end{aligned} \tag{5.6}$$

$$\begin{aligned} m_2^*(e_b) &= m_2(e_b) + \frac{1}{p(s)}[\underset{i<j}{\Sigma\Sigma} c_{sij}w_iw_j(\frac{v_i}{w_i^2} + \frac{v_j}{w_j^2}) \\ &\quad -\Sigma\frac{v_i}{w_i}\alpha_ic_{si}] + \Sigma v_ib_{si}^2I_{si}, \end{aligned} \tag{5.7}$$

writing $m_k(e_b) = m_k(t_b)|_{\underline{Y}=\underline{R}}, k = 1, 2$ given in Corollary 1.

It is easy to see that

$$E(m_k^*(e_b)) = M_1^*(e_b) = E_1E_2(e_b - t_b)^2 + M(t_b), k = 1, 2.$$

For $M_2^*(e_b)$ the proposed unbiased estimators are

$$\hat{m}_k(e_b) = m_k(e_b) + \Sigma v_i b_{si} I_{si}, k = 1, 2. \tag{5.8}$$

It is easy to check that

$$E\hat{m}_k(e_b) = M_2^*(e_b), k = 1, 2.$$

For $V_1^*(e_H) = E_1E_2[(e_H - t_H)^2] + V(t_H) = E_1[\Sigma V_i \frac{I_{si}}{\pi_i^2}] + V(t_H)$

our proposed unbiased estimators are:

$$v_1^*(e_H) = v_1(e_H) - \mathop{\Sigma\Sigma}_{i<j} \frac{I_{sij}}{\pi_{ij}}(\pi_i\pi_j - \pi_{ij})(\frac{v_i}{\pi_i^2} + \frac{v_j}{\pi_j^2}) + \Sigma v_i(1 - \beta_i)\frac{I_{si}}{\pi_i^2} \tag{5.9}$$

and

$$v_2^*(e_H) = v_2(e_H) - \frac{1}{p(s)}[\mathop{\Sigma\Sigma}_{i<j} c_{sij}(\pi_i\pi_j - \pi_{ij})(\frac{v_i}{\pi_i^2} + \frac{v_j}{\pi_j^2}) + \Sigma c_{si}\frac{v_i}{\pi_i}\beta_i] + \Sigma v_i \frac{I_{si}}{\pi_i^2} \tag{5.10}$$

writing $v_k(e_H) = v_k(t_H)|_{\underline{Y}=\underline{R}}, k = 1, 2$, given in Corollary 3.

For $V_2^*(e_H) = E_2E_1[(e_H - R) + (R - Y)]^2 = E_2V_1(e_H) + \Sigma V_i$

where $V_1(e_H) = V_1(t_H)|_{\underline{Y}=\underline{R}} = V(t_H)|_{\underline{Y}=\underline{R}}$,

our proposed unbiased estimators are

$$\hat{v}_k(e_H) = v_k(e_H) + \Sigma\frac{v_i}{\pi_i}I_{si}, k = 1, 2. \tag{5.11}$$

Next our proposed estimators for

$$M_2^*(e_g) = E_2M_3(e_g) + \Sigma V_i$$

are

$$m_k^*(e_g) = v_k(g)|_{\underline{Y}=\underline{R}} + \Sigma v_i\frac{I_{si}}{\pi_i}, k = 1, 2, . \tag{5.12}$$

Here $v_k(e_g)|_{\underline{Y}=\underline{R}}$ equals $v_k(t_g)|_{e_i=e_i(r)}$ as given in (4.4), $k = 1, 2$.

Remark V. Rao (1975), in the context of multi-stage sampling, illustrated a situation where (ii) should be replaced by $(ii)'V_2(r_i) = V_{si}, v_i$ by v_{si} and (iii) by $(iii)'E_2(v_{si}) = V_{si}$ when $i \epsilon s$, corresponding to (5.1) given earlier. Chaudhuri et al (2000) noted that in this situation (5.2) is not applicable.

When $(ii)', (iii)'$ are assumed and (5.2) is ruled out our proposals are the following:

(I) Replace v_i by v_{si} in the formulae (5.6), (5.7), (5.9), (5.10);

(II) Rule out the uses of (5.8), (5.11), (5.12).

Remark VI. For $M_1^*(e_g) = E_1E_2(e_g - t_g)^2 + M(t_g)$ we do not propose any estimator because no elegant estimator seems to be available.

Remark VII. To prove uniform non-negativity of an unbiased estimator of the mean square error (MSE) of a linear estimator of a finite population total when the size of a sample is allowed to vary across the samples is not easy.

We are able to prove this so far only for one situation with Seth's (1966) scheme with our modifications on it through our Theorems 2 and 3 above.

This illustration of course is too artificial. Unfortunately we are yet to hit upon a more natural one. Also, we do not come across any better one in the literature to date.

1.6 A numerical exercise on efficacy in estimation

Applying the modified sampling scheme of Seth (1966) we illustrate how the estimators given by (2.1), (2.4), (4.1) and (4.4) fare in yielding estimated coefficients of variation (CV) of estimates of totals. From SSW (1992, Appendix C, pp. 660-661) we take the first $N = 29$ clusters of municipalities as our illustrative population for which 'size' - values are taken as size-measures to apply the modified Seth (1966) scheme with $w = 0.4$ and the values of the total population in 1985 and 1975 are taken as respectively the values of y, the variable of interest and of x, the auxiliary correlated variable. We take $n = 9$. Table 1 presents, for 10 replicates of samples drawn as above from this population, the values of

$$a_1 = 100\frac{\sqrt{v_{HT}}}{t_H}, a_2 = 100\frac{\sqrt{v_1(t_H)}}{t_H}, b_k = \frac{100\sqrt{m_{kg}}}{t_g}, k = 1, 2, c_k = 100\frac{\sqrt{v_{kg}}}{t_g}, k = 1, 2.$$

Table 1

Performance of $v_1(t_H)$ VS v_{HT} and v_{kg} VS m_{kg} in terms of the criteria $a_1, a_2, b_k, c_k, (k = 1, 2)$ based on modified Seth (1966) Scheme using SSW (1992) data. We take $w = .4$.

Sample	Sample	a_1	a_2	b_1	c_1	b_2	c_2
number	size						
	realized						
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	9		29.95		1.34		1.20
2	9		14.84		1.31		1.43
3	9		18.17	0.74	2.38	0.87	2.79
4	10		27.32		0.33		0.28
5	10		11.41		0.80		0.70
6	10		11.77	0.28	0.38	0.40	0.53
7	9		15.11		0.27		0.27
8	10		18.34		0.36		0.43
9	9		13.63		1.08		1.24
10	9		16.13	0.59	1.15	0.54	1.07

Table 2

Performance of $v_1(t_H) VS\ v_{HT}$ and $v_{kg} VS\ m_{kg}$ in terms of the criteria $a_1, a_2, b_k, c_k (k = 1, 2)$ based on 'modified Seth (1966) scheme' using Indian census 1991 data.

Sample number	Sample size realized	a_1	a_2	b_1	c_1	b_2	c_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	6		20.36		7.86		7.77
2	7		17.42	0.00	8.50	0.00	7.88
3	7		12.37		12.15		10.15
4	7		18.34	6.73	6.52	9.01	8.70
5	7		9.56	3.08	3.47	4.99	5.62
6	7		18.23	6.33	10.31	5.32	8.52
7	7		27.95	1.93	10.84	2.09	11.74
8	7		12.73	10.53	10.05	11.16	10.65

We apply the same method taking $n = 6$ and $w = 0.4$ as before to draw samples from 23 villages in a particular district for which the household size is taken as the size measure, y as the area in hectare and x as the total population size, the source for each being the Indian population census, 1991.

For 8 replicates of samples the values of a_1, a_2, b_k, c_k are presented in Table 2.

Note: Absence of an entry in Tables 1,2 and 3 signifies 'negative' values of v_{HT} or m_{kg}. Everywhere R_i is taken as $(1-\pi_i)/\pi_i x_i,\ i \epsilon U$.

Table 3

Performance of $v_1(t_H) VS\ v_{HT}$ and $v_{kg} VS\ m_{kg}$ via the criteria a_1, a_2, b_k, c_k, $(k = 1, 2)$ based on PPS circular systematic samples repeated twice using data from SSW (1992).

Sample number	realized sample size	a_1	a_2	b_1	c_1	b_2	c_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	8	41.61	164.24		4.68		10.38
2	9	31.04	107.24		2.85		5.12
3	10	22.19	84.80		1.54		3.19
4	10	29.09	101.60	0.26	1.73	0.57	3.82
5	10	27.77	94.01		2.39		4.42
6	7	37.07	124.40		1.43		3.73
7	10	30.82	106.58	0.09	1.49	0.16	2.76
8	10	32.56	149.02		3.11		3.36

Finally we illustrate instead of the rather artificial sampling scheme above a realistic one which in fact is usually applied in Indian annual national sample surveys covering many socio-ecomomic issues. This is the circular systematic sampling (CSS) with probability proportional to size (PPS) in a pre-assigned number of draws but the entire draw is 'independently' repeated twice. The draw is repeated because for many pairs (i, j) the inclusion-probabilities π_{ij}'s turn out to be zero in case of a 'single' draw of the sample. For the CSSPPS sampling repeated twice, each sample being of size n, the realized number of distinct units $\nu(s)$ varies between n and $2n$, the inclusion-probabilities, ψ_i, say, of i are determined in terms of the normed size measures p_i's and the inclusion-probabilities ψ_{ij}'s of (i, j)'s, say, turn out positive. So, for this sampling scheme v_{HT} and $v_1(t_H)$ are competitors and so are m_{kg} vis-a-vis $v_{kg}, (k = 1, 2)$. Using the same 29 values of size-measures, y and x as in Table 1 and taking $n = 5$ for each CSSPPS repeated twice the similar exercise as in Tables 1 and 2 above is presented for 8 replicated samples in Table 3.

1.6.1 Comments on numerical findings and recommendations

For the 'modified Seth (1966) scheme' v_{HT} is throughout negative justifying thoroughly the introduction of $v_1(t_H)$. For CSSPPS repeated twice however this is not the case and $v_1(t_H)$ leads to loss in efficiency.

For both the schemes, both $m_{kg}, (k = 1, 2)$ turn out negative, justifying the proposal for $v_{kg}(k = 1, 2)$ as their competitors. However when they turn out positive, they often yield higher efficiencies compared to $v_{kg}(k = 1, 2)$.

Incidentally, for CSSPPS repeated twice, (1) $(\psi_i\psi_j - \psi_{ij})$ have variable signs but (2) $\beta_i > 0$ for every i, while (3) our proposed MSE estimators turn out positive for each sample.

Our recommendation is that when employing t_g and t_H based on 'varying sample size sampling schemes' one should employ, respectively, v_{kg} as possible competitors against $m_{kg}(k = 1,2)$ and $v_1(t_H)$ against v_{HT} irrespective of whether theorems like Theorems 2 and 3 apply for the sampling scheme employed or not.

Chapter 2

On a version of cluster sampling and its practical use

Abstract

We consider a practical sample survey problem in which we require to reach households in villages which in batches are served by certain primary health centres (PHC) or by somewhat 'bigger' primary health centres (BPHC) located in various territorial cross-sections in a district near Calcutta. We present here a comparative study of alternative schemes of sampling the PHC's and the BPHC's and estimating the totals of variate values related to all the BPHC's and PHC's in the district. The situation demands adequate representation of the BPHC's in the sample and geographically, each BPHC separately has an exclusive group of PHC's contiguous and hence attached to it. So, a BPHC together with the neighbouring PHC's may be supposed to constitute a cluster, each cluster disjoint

with every other and values within a cluster are supposed to be well and positively correlated. Some of the possible alternative ways of sampling are: (1) Sampling from the pool of all the BPHC's and PHC's taken together, (2) Sampling independently from the two separate strata of BPHC's and PHC's and (3) Two-stage sampling with the clusters mentioned above as the first stage units and choosing from each selected cluster the BPHC in it and a sample from the PHC's contiguous to this BPHC. But we recommend a fourth alternative that selects first a sample of PHC's from all the PHC's in a cross-section and attaches to each selected PHC the BPHC in the cluster to which it belongs. This yields several alternative estimators of interest. By a simulation exercise we present numerical findings of relative performances of these various procedures. Though it is difficult to identify a best procedure theoretically, empirically our proposed procedure appears to be quite encouraging. It ensures a desirably wide territorial coverage of PHC's in the sample and reducing a stage of sampling promises to yield efficient estimators.

2.1 Introduction

Recently, Indian Statistical Institute (ISI), Calcutta collaborated with UNICEF, Calcutta to examine the extent of Infant and Maternal mortality experiences in a district near Calcutta. It was found convenient to identify several cross-sections of the geographical coverage as various strata classified according to varying levels of accessibility to them from the adjoining cities. Within each stratum households were to be sampled within selected villages. The villages themselves were observed to be served in exclusive batches either by a separate primary health centre (PHC) or a 'bigger' PHC, say, a BPHC. To each BPHC were contiguous and hence attached a number of PHC's. So, it was felt that each BPHC combined with its associated PHC's might be regarded as a 'Cluster' and values of variables of interest pertaining to each PHC and BPHC in such a cluster to be well correlated. Each such cluster is disjoint from every other. In this chapter we consider appropriate ways of sampling the BPHC's and the PHC's within each separate cross-section or stratum and in the entire district as well Of course, from each selected PHC and each BPHC we further take samples of villages separately served by them and select households from the chosen villages. In other words at the PHC and BPHC levels we handle only 'estimated totals' of values related to them. But in our presentation here we ignore this aspect and treat these 'estimated totals' as true PHC- or BPHC- specific values in developing our analytic results. This is because our intention here is only to discuss appropriate ways of sampling the PHC's and the BPHC's and using values related to them to produce serviceable estimates of totals of all the PHC- and BPHC-specific values in the respective strata and in the district.

It was recognized that medical facilities were available in greater abundance from the BPHC's than from the PHC's. So, adequate representation of BPHC's in the sample seemed to be an important requirement. So, the following two sampling schemes appeared to be worth trying:

I. A two-stage cluster sampling. We could treat each cluster as a 'first stage unit' (fsu) and take a sample of them. From each selected cluster one may then take the BPHC and a sample of PHC's from all the PHC's in it.

II. A single-stage cluster sampling. In order to achieve a wide geographical coverage a sample of PHC's may first be chosen from all the PHC's in a 'cross-sectional' stratum and for each chosen PHC the BPHC in the cluster to which it belongs may be added in the sample.

We actually adopted the scheme II in the ISI survey because we felt the inherent 'correlation' among the values 'within the cluster' could thereby be well exploited in estimation and the sample would achieve a wide territorial representation.

We find that the scheme II permits several alternative estimators of totals and variance estimators. For simplicity the PHC's were chosen by the method of simple random sampling (SRS) without replacement (WOR). Yet the BPHC's turned out to have varying probabilities of inclusion.

Two other possible ways of sampling might be:

III. Treating the BPHC's and the PHC's as two separate strata within each cross-sectional stratum a stratified SRSWOR method may be tried;

IV. An SRSWOR from the pool of all the BPHC's and the PHC's may be selected independently from each cross-section.

In section 2.2 we present some details of the features of the scheme II, several alternative estimators of totals and their variance estimators. In section 2.3 we present some numerical results based on simulations to indicate how the various alternative procedures may fare in respect of two well-known criteria for comparison on the basis of such empirical studies. The section 2.4 gives some concluding remarks and recommendations.

An alternative prediction method of two-stage sampling as discussed by Bolfarine and Zacks (1992), however is not considered as a possible competitor here.

2.2 Estimators and Variance Estimators

Our focus is on scheme II and so let us first present the material related to this. We need the following notations to start with. For a typical cross-sectional stratum let there be k BPHC's labeled i and N_i PHC's be attached to the ith BPHC labeled $i1, \cdots, ij, \cdots, iN_i$ with $i = 1, \cdots, k$. Let y_i and y_{ij} be the values of a variable y of interest for the ith BPHC and for the jth PHC attached to the ith BPHC respectively.

Let $Y_1 = \sum_{i=1}^{k} y_i, Y_2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i} y_{ij}$ and $Y = Y_1 + Y_2$. Our task is to estimate Y using y_i and y_{ij}-values for a sample of PHC's and BPHC's. For the scheme II let n_i denote the number of PHC's that happen to be selected from among the N_i PHC's that together with the ith BPHC constitute the ith cluster $(i = 1, \cdots, k)$ when an SRSWOR of n PHC's is chosen out of the total of $N = N_1 + \cdots + N_i + \cdots + N_k$ PHC's in the above stratum. Then, the value of y_i may be recorded n_i times. By $E(.), V(.), C(.,.)$ we shall denote the operators for expectation, variance and covariance with respect to an adopted scheme of sampling. Let s denote a sample of PHC's coupled with the chosen BPHC's for the scheme II.

$$\begin{aligned} \text{Let } I_s(\alpha) &= 1, \text{ if a PHC or BPHC labeled } \alpha \text{ is in } s, \\ &= 0, \text{ else; here } \alpha = i1, \cdots, ij, \cdots, iN_i; i = 1 \cdots, k; \end{aligned}$$

$\pi_\alpha = E(I_s(\alpha))$ = the inclusion probability of α in a sample for scheme II,

$\pi_{\alpha,\alpha'}$ = the inclusion-probability of BPHC's α, α' in a sample,

$\pi'_{\alpha,\alpha'}$ = the same for PHC's α, α',

$\pi''_{\alpha,\alpha'}$ = the same for BPHC α and PHC α',

$\pi^*_{\alpha,\alpha'}$ is a common notation reserved for $\pi_{\alpha,\alpha'}, \pi'_{\alpha,\alpha'}, \pi''_{\alpha,\alpha'}$ the one meant for to be clear from a given context;

$\nu(s)$= the number of distinct units i.e. PHC's, BPHC's in s; $\nu = E(\nu(s))$; $r_i = \frac{n_i}{N_i}$; d = the number of distinct BPHC's sampled, M_j's are the numbers N_i's when arranged in the increasing order; T_j's are N_i's arranged in the decreasing order; thus

$M_1 \leq \cdots \leq M_i \leq \cdots \leq M_k; T_1 \geq \cdots \geq T_j \geq \cdots \geq T_k; C_i = M_1 + \cdots + M_i;$ $D_i = T_1 + \cdots + T_i; i = 1, \cdots, k; P_i = \frac{N_i}{N}, Q_i = 1 - P_i; p_i = \frac{n_i}{n}; q_i = 1 - p_i;$ $p(s_d)$ = the probability of selecting a particular sample s_d of BPHC's with exactly d distinct BPHC's in it;

s_d^* = the collection of samples of BPHC's each having a common set of distinct units as in a given s_d;

$\sum'$= the sum over samples like s_d ignoring 'order' and/or 'multiplicity' of BPHC's in it,

$p(d)$= the probability that d is the number of distinct BPHC's that happen to be sampled;

g_i= the frequency of the ith BPHC in the set of all s_d^*'s with various values of d.

By $\hat{\theta}$ we shall denote a typical estimator for a parameter θ, by v an estimator for its variance, regard $\delta = \frac{\hat{\theta}-\theta}{\sqrt{v}}$ as a standard normal deviate and hence treat $(\hat{\theta}-1.96\sqrt{v}, \hat{\theta}+1.96\sqrt{v})$ or $(\hat{\theta}\pm 1.96\sqrt{v})$ in brief as a 95 per cent confidence interval (CI) for θ. We present below four alternative unbiased estimators for Y based on scheme II and their unbiased variance estimators.

Let $$e_1 = \sum_{i=1}^{k} y_i \frac{I_s(i)}{\pi_i}, \qquad e_2 = \sum_{i=1}^{k} y_i \frac{r_i}{E(r_i)} = \frac{N}{n}\sum_{1}^{k} y_i(\frac{n_i}{N_i}) = e_2(s_d),$$
$$e_3 = e_3(s_d) = \frac{1}{p(s_d)} \sum_{i\epsilon s_d} \frac{y_i}{g_i}, \quad e_4 = \frac{\sum' e_2(s_d)p(s_d)}{\sum' p(s_d)},$$
$$f_1 = \sum_{i=1}^{k}\sum_{j=1}^{N_i} y_{ij}\frac{I_s(ij)}{\pi_{ij}} = \frac{N}{n}\sum_{i=1}^{k}\sum_{j=1}^{N_i} y_{ij}I_s(ij).$$

Then, Horvitz and Thompson's (HT, 1952) unbiased estimator for Y for the scheme II is $t_1 = e_1 + f_1$.

Three other obviously unbiased estimators for Y are

$$t_2 = e_2 + f_1, t_3 = e_3 + f_1 \text{ and } t_4 = e_4 + f_1.$$

Remark I. The number of distinct units in s is a random variable. It is

$$\nu(s) = n + d.$$

The range of variation in d is given by the

Proposition 1. $m \leq d \leq M$.

Here $M = \min(n, k)$;

It may be checked that m is either (i) the minimum value of r for which $n - C_r \leq 0$ or (ii) the minimum value of r for which $n - D_r \leq 0$.

Because $\nu(s)$ is not a constant the Yates and Grundy's (YG,1953) form of variance estimator for the HT estimator t_1 is not available. We shall consider (1) the HT form of the variance estimator of t_1 which is

$$v_1 = v(t_1) = \sum_{\alpha} y_{\alpha}^2 \frac{(1-\pi_{\alpha})}{\pi_{\alpha}} \frac{I_s(\alpha)}{\pi_{\alpha}} + \sum_{\alpha \ddagger \alpha'}\sum y_{\alpha} y_{\alpha'} (\frac{\pi^*_{\alpha,\alpha'} - \pi_{\alpha}\pi_{\alpha'}}{\pi_{\alpha}\pi_{\alpha'}}) \frac{I_s(\alpha,\alpha')}{\pi^*_{\alpha,\alpha'}}$$

and (2) the alternative form of variance estimator of t_1 given by Chaudhuri(2000a) and as discussed in details in the Chapter 1 of this dissertation, namely

$$v_1' = v'(t_1) = \underset{\alpha<\alpha'}{\Sigma\Sigma} (\frac{\pi_{\alpha}\pi_{\alpha'} - \pi^*_{\alpha,\alpha'}}{\pi^*_{\alpha,\alpha'}})(\frac{y_{\alpha}}{\pi_{\alpha}} - \frac{y_{\alpha'}}{\pi_{\alpha'}})^2 I_s(\alpha,\alpha') + \sum_{\alpha} \frac{y_{\alpha}^2}{\pi_{\alpha}^2} \beta_{\alpha} I_s(\alpha)$$

where $\beta_{\alpha} = 1 + \frac{1}{\pi_{\alpha}} \sum_{\alpha'(\neq\alpha)}\sum \pi^*_{\alpha,\alpha'} - \sum_{\alpha} \pi_{\alpha}$.

Similarly, for e_1 as an unbiased estimator for Y_1 and f_1 as that for Y_2 the corresponding HT forms of variance estimators may be denoted by $v(e_1)$ and $v(f_1)$. For f_1 however $v(f_1)$ is also of the YG form.

We may observe the following formulae which are easy to check:

$\pi_{ij} = \frac{n}{N}$ for every $j = 1, \cdots, N_i$ for the respective $i = 1, \cdots, k$.

$\pi'_{ij,i'j'} = \frac{n(n-1)}{N(N-1)}$ for every $ij (= i1, \cdots, iN_i;\ i = 1, \cdots k)$ different from $i'j' (= i'1, \cdots, i'N_i';\ i' = 1, \cdots, k)$

$\pi_i = 1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}}$;

$\pi_{i,i'} = \pi_i + \pi_{i'} - 1 + \frac{1}{\binom{N}{n}} \binom{N-N_i-N_{i'}}{n}$.

$$\pi''_{i,\alpha} = \frac{n}{N} \text{ for } \alpha = i1, \cdots, iN_i \text{ and } i = 1, \cdots, k;$$

$$= \quad \pi_i + \frac{n}{N} - 1 + \frac{1}{\binom{N}{n}}\binom{N - N_i - 1}{n},$$
$$\text{for } \alpha = i'1, \cdots, i'N_{i'} \text{ but } i' \neq i; i = 1, \cdots, k.$$

In the Appendix we add a few numerical illustrations concerning the well-known consistency relations that the inclusion-probabilities should obey to show that they in fact do so for the scheme II. To work out the variance estimators for t_2, t_3, t_4 we note the following.

$V(t_2) = V(e_2) + V(f_1) + 2C(e_2, f_1)$.

$V(e_2) = (\frac{N}{n})^2[\sum\limits_1^k y_i^2 V(r_i) + \sum\limits_{i \neq i'}^{k}\sum^{k} y_i y_{i'} C(r_i, r_{i'})]$.

$V(r_i) = (\frac{n}{N_i})^2 V(p_i) = (\frac{n}{N_i})^2(\frac{N-n}{Nn})\frac{NP_iQ_i}{(N-1)}$,

$C(r_i, r_{i'}) = -\frac{n^2}{N_iN_{i'}}(\frac{N-n}{Nn})\frac{NP_iP_{i'}}{(N-1)}$,

$\bar{y} = \frac{1}{n}\sum\limits_1^k\sum\limits_1^{N_i} y_{ij} I_s(ij), s^2 = \dfrac{1}{(n-1)}\sum\limits_1^k\sum\limits_1^{N_i} I_s(ij)(y_{ij} - \bar{y})^2$.

Though both are known, $V(r_i)$ and $C(r_i, r_{i'})$ may also respectively be unbiasedly estimated by

$\hat{V}(r_i) = (\frac{n}{N_i})^2(\frac{N-n}{Nn})\frac{np_iq_i}{(n-1)}, \hat{C}(r_i, r_{i'}) = -\frac{n^2}{N_iN_{i'}}(\frac{N-n}{Nn}) \times \frac{np_ip_{i'}}{(n-1)}$

The product Y_1Y_2 may be estimated by

$$(\widehat{Y_1Y_2})_1 = \sum_\alpha\sum_{\alpha'}\frac{y_\alpha y_{\alpha'}}{\pi^*_{\alpha,\alpha'}} I_s(\alpha, \alpha')$$

and also by

$$(\widehat{Y_1Y_2})_{2j} = e_1 f_1 - \hat{C}_j(e_1, f_1), j = 1, 2,$$

writing

$$\hat{C}_1(e_1, f_1) = v(t_1) - v(e_1) - v(f_1)$$

and

$$\hat{C}_2(e_1, f_1) = v'(t_1) - v(e_1) - v(f_1)$$

on observing that

$$C(e_1, f_1) = E(e_1 f_1) - Y_1 Y_2.$$

Similarly, $C(e_2, f_1)$ may be unbiasedly estimated by

$$\hat{C}_1(e_2, f_1) = e_2 f_1 - (\widehat{Y_1 Y_2})_1$$

or by

$$\hat{C_{2j}}(e_2, f_1) = e_2 f_1 - (\widehat{Y_1 Y_2})_{2j}, j = 1, 2.$$

$V(e_2)$ may be unbiasedly estimated by

$$v_1(e_2) = (\frac{N}{n})^2 [\sum_1^k y_i^2 V(r_i) \frac{I_{si}}{\pi_i} + \sum_{i \neq i'}^{k} \sum^{k} y_i y_{i'} C(r_i, r_{i'}) \frac{I_s(i, i')}{\pi_{i,i'}}]$$

and by

$$v_2(e_2) = (\frac{N}{n})^2 [\sum_1^k y_i^2 \hat{V}(r_i) I_s(i) + \sum_{i \neq i'}^{k} \sum^{k} y_i y_{i'} \hat{C}(r_i, r_{i'}) I_s(i, i')].$$

An unbiased estimator for $V(f_1)$ is $v(f_1) = N^2 (\frac{N-n}{Nn}) s^2$.

So, four unbiased estimators for $V(t_2)$ are:

$$v_{\theta\lambda}(2) = v_\theta(e_2) + v(f_1) + 2\hat{C}_\lambda(e_2, f_1); \theta = 1, 2; \lambda = 1, 2.$$

Two more unbiased estimators for $V(e_2)$ and hence $v_{\xi\lambda}(2), \xi = 3, 4;\ \lambda = 1, 2,$ as four more unbiased estimators for $V(t_2)$ may be proposed as presented below following Rao (1979).

Writing $d_{ii} = V(r_i), d_{ii'} = C(r_i, r_{i'}), \hat{d}_{ii} = \hat{V}(r_i), \hat{d}_{ii'} = \hat{C}(r_i, r_{i'})$ we have

$$V(e_2) = (\frac{N}{n})^2 (\sum_1^k \sum_1^k y_i y_{i'} d_{ii'}).$$

We may observe that if

$y_i \infty N_i$, then e_2 equals Y_2 and hence $V(e_2)$ equals zero. Then, from Rao (1979) it follows that

$$V(e_2) = -(\frac{N}{n})^2 \sum_{i<i'}^{kk}\sum N_i N_{i'} (\frac{y_i}{N_i} - \frac{y_{i'}}{N_{i'}})^2 d_{ii'}.$$

This observation yields two unbiased estimators for $V(e_2)$ as

$$v_3(e_2) = -(\frac{N}{n})^2 \sum_{i<i'}\sum N_i N_{i'} (\frac{y_i}{N_i} - \frac{y_{i'}}{N_{i'}})^2 d_{ii'} \frac{I_s(i,i')}{\pi_{ii'}}$$

and

$$v_4(e_2) = -(\frac{N}{n})^2 \sum_{i<i'}\sum N_i N_{i'} (\frac{y_i}{N_i} - \frac{y_{i'}}{N_{i'}})^2 \hat{d}_{ii'} I_s(i,i').$$

So,

$$v_{\xi\lambda}(2) = v_\xi(2) + v(f_1) + 2\hat{C}_\lambda(e_2, f_1), \xi = 3, 4; \lambda = 1, 2j(j = 1, 2)$$

are several further unbiased estimators of $V(t_2)$.

In order to unbiasedly estimate

$$V(t_3) = V(e_3) + V(f_1) + 2C(e_3, f_1),$$

we first note that respective unbiased estimators for $Y_1^2, V(e_3)$ and $C(e_3, f_1)$ are

$\hat{Y}^2(1) = e_1^2 - v(e_1)$ and $\hat{Y}^2(2) = e_1^2 - v'(e_1)$

$v_j(e_3) = e_3^2 - \hat{Y}^2(j), j = 1, 2$; and

$$\hat{C}_1(e_3, f_1) = e_3 f_1 - (\widehat{Y_1 Y_2})_1$$

and also

$$\hat{C}_{2j}(e_3, f_1) = e_3 f_1 - (\widehat{Y_1 Y_2})_{2j}, j = 1, 2.$$

Consequently,

$$\hat{V}(t_3)_{\theta\xi} = v_\theta(e_3) + v(f_1) + 2\hat{C}_\xi(e_3, f_1),$$

say, $\theta = 1, 2; \xi = 1, 2j (j = 1, 2)$ emerge as several unbiased estimators of $V(t_3)$.

To unbiasedly estimate $V(t_4)$ we observe that e_4 is derived from e_2 by Rao-Blackwellization and note from Chaudhuri and Stenger (1992) that we have

$$V(e_2) - V(e_4) = E(e_2 - e_4)^2$$

and so using any one of the unbiased estimators $v_\xi(e_2)$ for $V(e_2)$ one may employ an unbiased estimator

$$v_\xi(e_4) = v_\xi(e_2) - (e_2 - e_4)^2 \text{ for } V(e_4), \xi = 1, 2, 3, 4.$$

Similarly we may unbiasedly estimate $C(e_4, f_1)$ by

$$\hat{C}_\lambda(e_4, f_1) = e_4 f_1 - (\widehat{Y_1 Y_2})_\lambda, \lambda = 1, 2j (j = 1, 2).$$

So, $V(t_4) = V(e_4) + V(f_1) + 2C(e_4, f_1)$ may be unbiasedly estimated by $v(t_4)$ which is, say,

$$v_{\xi\lambda}(4) = v_\xi(e_4) + v(f_1) + 2\hat{C}_\lambda(e_4, f_1), \xi = 1, 2, 3, 4; \lambda = 1, 2j (j = 1, 2).$$

2.3 Numerical Evaluation of Relative Efficacies of Various Procedures by Simulation

As mentioned at the introduction in Section 2.1 the problem treated here originated when an actual survey was undertaken. The scheme II was adopted from a pragmatic consideration. But the theory discussed above was developed even before the actual survey was executed. Since our plan here is to anticipate how this scheme may compete with others which cannot all be implemented in the same survey, instead of waiting for the actual survey results to be gathered we are curious to evaluate below its possible performance with fictitious data through a simulation study. Of course, the survey results are now at hand. But these are ignorable in arriving at the kind of conclusions we plan to reach concerning the efficacy of scheme II relative to its possible competitors. The criteria for comparison employed cannot be calculated from 'Sample Survey Data' alone.

So, for the sake of illustration of how the various procedures may compete we present in the Table 1 below some arbitrarily chosen values for $N, k, N_i (i = 1, \cdots, k), n, y_i, y_{ij}$, number of strata H and use the notation L_h for the hth stratum total Y's.

Table 1
Some fictitious data about strata-wise BPHC's, PHC's.

Stratum serial number h	BPHC values y_i	PHC values y_{ij}	L_h	k	N_i	n	N
1	103.00	(196.00, 151.00, -97.00, -39.00)			4		
	-100.00	(74.00, 0.00, 141.00, 131.00, 63.00, 22.00)			6		
	116.00	(5.00, 80.00, 179.00, 52.00, 10.00, 2.00)			6		
	-127.00	(38.00, -152.00, 29.00, 42.00, 99.00, 10.00, 34.00, -87.00)			8		
			975.00	4		5	24
2	72.00	(93.00, 126.00, -38.00)			3		
	33.00	(46.00, 133.00, -121.00), 107.00, 51.00, 84.00)			6		
	-75.00	(9.00, 71.00, 101.00, 7.00)			4		
	-84.00	(5.00, 28.00, 75.00, 51.00, -35.00, 22.00, -269.00)			7 7		
			495.00	4		6	20
3	62.00	(0.00, 30.00, -62.00)			3		
	65.00	(126.00, 36.00, -71.00), 110.00, 51.00, 106.00) 34.00, -77.00)			8		
	-39.00	(3.00, 0.00, 69.00, 12.00, 271.00, -29.00)			6		
	-37.00	(-25.00, 50.00, 57.00, 44.00, 19.00, -62.00 5.00, 32.00)			8		
	-54.00	(39.00, -84.00, 26.00, 52.00, 35.00, -15.00, 60.00)			7		
			839.00	5		7	32
4	98.00	(53.00, 0.00, 0.00) (1.00, 129.00, -127.00),			6		
	-85.00	(78.00, 0.00, 13.00, 846.00) -10.00, 83.00, 29.00, -2.00)			8		
	-88.00	(125.00, 27.00, 66.00, -31.00, 77.00, 67.00, -77.00)			7		
	-84.00	(1.00, 65.00, 60.00, 104.00, 6.00, -43.00)			6		
			1381.00	4		8	27
5	20.00	(67.00, 100.00, 6.00, -73.00)			4		
	22.00	(32.00, 141.00, -5.00), 24.00, -2.00)			5		
	-12.00	(40.00, 3.00, 0.00),			3		
	-16.00	(31.00, 4.00, 73.00, -4.00, 3.00, 41.00, 3.00, -12.00)			8		
	-28.00	(48.00, 41.00, 132.00, -26.00)			4		
			653.00	5		7	24

A Remark: We have included negative values to cover situations considering changes over time which may reflect growth as well as decay.

To examine the performance of any pair of an estimator e for a parameter θ admitting a positive variance estimator v we take $R = 1000$ independently replicated samples and calculate the 'Actual coverage percentage' i.e.

(I) ACP = the percentage of the replicates for which the interval (CI) $(e - 1.96\sqrt{v}, e + 1.96\sqrt{v})$ covers the parameter θ - the closer it is numerically to 95 the better;
and the Average coefficient of variation

(II) ACV = the average, over the R replicates, of the values of $100 \times \frac{\sqrt{v}}{|e|}$ – this reflects the length of CI – the smaller it is the better.

In this case $\theta = \sum_{h=1}^{5} L_h$ and the sample is chosen according to scheme II from each stratum and e stands for t_1, t_2, t_3, t_4 and v for their various alternative variance estimators discussed in section 2. To bring the Scheme I under comparison we repeat this exercise for it likewise. From each stratum a preassigned fraction r_1, say, of k BPHC's is chosen by SRSWOR method and from each cluster consisting of the selected BPHC and the associated PHC's a pre-assigned fraction, r_2, say, of the PHC's in the cluster is chosen by SRSWOR and to the selected PHC's the BPHC in the cluster is added to give the sample from the stratum. Then, the standard estimation of stratum total and variance estimation formulae, and therefrom the estimator for the district total and its variance estimator, say, e_I for θ and v_I for $v(e_I)$ are employed. We omit the explicit formulae to save space. We choose r_1 and r_2 judiciously to keep the over-all sample sizes for schemes I and II as close to each other as practicable. Similar exercise is carried out for scheme III yielding estimator e_{III} say for θ and variance estimator v_{III} based on SRSWOR's of BPHC's out of all BPHC's and independently chosen SRSWOR'S of PHC's out of all PHC's from each separate cross-sectional stratum in numbers comparable as practicable with those for schemes I, II. Similarly for the scheme IV also his replicated sampling is implemented yielding estimate e_{IV} for θ and v_{IV} for $V(e_{IV})$ on choosing SRSWOR's of BPHC's and PHC's out of all the BPHC's and the PHC's from each 'cross-sectional stratum' in independent manners' repeating the same independently across the strata. The formulae for e_{III}, e_{IV} and v_{III}, v_{IV} are too well-known to bear specifications here. The numerical findings appear in Table 2.

Table 2

Relative performance of ACP/ACV values for four schemes I-IV

Estimator, variance estimator	ACP/ACV	Estimator, variance estimator	ACP/ACV	Estimator, variance estimator	ACP/ACV
(e, v)		(e, v)		(e, v)	
(t_1, v_1)	94.3/45	(e_1, v_1)	100/115	(t_3, v_{13})	95.0/40
(t_1, v_1')	95.6/40				
$(t_2, v_{11}(2))$	93.9/29	$(t_2, v_{12}(2))$	94.7/29	(t_3, v_{23})	95.4/43
		$(t_2, v_{12}'(2))$	96.1/26	(t_3, v_{23}')	96.1/23
$(t_2, v_{21}(2))$	93.9/29	$(t_2, v_{22}(2))$	94.7/29	$(e_{\mathrm{III}}, v_{\mathrm{III}})$	94.0/41
		$(t_2, v_{22}'(2))$	96.2/26		
$(t_2, v_{31}(2))$	93.7/29	$(t_2, v_{32}(2))$	94.7/29	$(e_{\mathrm{IV}}, v_{\mathrm{IV}})$	92.0/38
		$(t_2, v_{32}'(2))$	96.2/26		
$(t_2, v_{41}(2))$	93.1/29	$(t_2, v_{42}(2))$	94.2/29	$(t_4, v_{12}(4))$	96.5/41
		$(t_2, v_{42}'(2))$	96.6/26	$(t_4, v_{12}'(4))$	94.3/39
$(t_4, v_{11}(4))$	96.3/40	$(t_4, v_{31}(4))$	96.3/40	$(t_4, v_{22}(4))$	96.5/41
				$(t_4, v_{22}'(4))$	94.3/39
$(t_4, v_{21}(4))$	96.3/40	$(t_4, v_{41}(4))$	96.2/40	$(t_4, v_{32}(4))$	96.5/41
				$(t_4, v_{32}'(4))$	94.3/36
$(t_2, v_{42}(4))$	96.3/41				
$(t_4, v_{42}'(4))$	97.3/38				

2.4 Concluding Remarks and Recommendations

From the illustrated empirical study reported in Table 2 it seems pretty clear that in terms of the twin criteria of ACP and ACV the Scheme II outperforms all other competitors if the estimator t_2 is employed no matter which variance estimator is used for it. The obvious traditional two stage procedure palpably takes a back seat. Schemes III and IV also are poorer. Even for the Scheme

II the rival estimators t_1, t_3 and t_4 do not fare competitively with t_2.The scheme II using v_1' fares better than when using v_1 instead vindicating the usefulness of Chaudhuri's(2000a) variance estimator for the HT estimator.

So, our Scheme II with the estimator t_2 is recommended to take care of a situation similar to the one presented here.

A Remark. The real data for the actual survey carried out in 1997-1998 based on our proposed cluster sampling scheme, providing estimated coefficients of variation are not quoted here because it is impossible for them to provide any insight about the performance of the strategy actually employed relative to other strategies which might have been implemented but actually not thereby producing no live data for comparison. Hence our falling back upon simulations alone for comparative studies.

A major justification for the choice of scheme II is that on choosing randomly the PHC's first and adding next the allied BPHC's we expect a wider territorial coverage and a greater information content compared to choosing the BPHC's first and the surrounding PHC's next.

Moreover the scheme II admits several alternative estimators and a few novel ones permitting comparison among themselves and thus it is quite flexible in applications.

Appendix

For the Scheme II we check here a few consistency conditions. To check that (i) $\nu = E(\nu(s))$ equals $\sum \pi_\alpha$ and (ii) $V(\nu(s)) + \nu(\nu - 1)$ equals $\sum\sum_{\alpha \neq \alpha'} \pi^*_{\alpha,\alpha'}$ let us illustrate with $k = 3, N_1 = 3, N_2 = N_3 = 2; \quad N = 7$ and $n = 4$.

Using proposition 1 we observe $m = 2, M = 3, 2 \leq d \leq 3$.

It follows that $p(2) = \frac{1}{\binom{7}{3}}[\binom{3}{2}\binom{2}{2}\binom{2}{0} + \binom{3}{2}\binom{2}{0}\binom{2}{2} + \binom{3}{0}\binom{2}{2}\binom{2}{0} + \binom{3}{3}\binom{2}{1}\binom{2}{0} + \binom{3}{3}\binom{2}{0}\binom{2}{1}] = \frac{11}{35}$; likewise $p(3) = \frac{24}{35}$. Then, $\nu = 6(\frac{11}{35}) + 7(\frac{24}{35}) = \frac{234}{35}; V(\nu(s)) = \frac{264}{1225}; V(\nu(s)) + \nu(\nu - 1) = \frac{1338}{35}$.

Also, $\pi_1 = \frac{34}{35}, \pi_2 = \frac{30}{35} = \pi_3, \pi_{11} = \pi_{12} = \pi_{13} = \pi_{21} = \pi_{22} = \pi_{31} = \pi_{32} =$

$\frac{4}{7}$.

Thus, (i) is verified. Furthermore,

$\pi_{1,2} = \frac{29}{35} = \pi_{1,3}; \pi_{2,3} = \frac{25}{35}; \pi'_{\alpha,\alpha'} = \frac{2}{7}$ for every $\alpha \neq \alpha'(= 11, 12, 13; 21, 22; 31, 32)$; $\pi''_{1,11} = \pi''_{1,12} = \pi''_{1,13} = \frac{4}{7} = \pi''_{2,21} = \pi''_{2,22} = \pi''_{3,31} = \pi''_{3,32}$; $\pi''_{1,21} = \pi''_{1,22} = \frac{19}{35} = \pi''_{1,31} = \pi''_{1,132}; \pi''_{2,11} = \pi''_{2,12} = \pi''_{2,13} = \pi''_{2,31} = \pi''_{2,32} = \frac{16}{35}; \pi''_{3,11} = \pi''_{3,12} = \pi''_{3,13} = \frac{16}{35} = \pi''_{3,21} = \pi''_{3,22}$.

Adding these $\pi^*_{\alpha,\alpha'}$'s the relation (ii) is also verified.

Chapter 3

Systematic sampling: 'Fixed' versus 'Random' Sampling Interval

Abstract

In the customary 'circular systematic sampling' (CSS) scheme with selection 'probabilities proportional to sizes' (PPS), with (i) a single random start, (ii) a pre-assigned number of draws and (iii) a preassigned 'sampling interval', the joint inclusion-probabilities for certain pairs of units may be zero leading to non-availability of an 'unbiased' variance estimator for a linear estimator of a finite population total. To get over this problem a well-known 'convention' is to apply the CSSPPS scheme with 2 'independent' random starts.

Here we consider an alternative approach with a 'single random start' but with a 'random', instead of a 'fixed' "sampling interval" for which 'modified CSSPPS' scheme every pair of units has a positive inclusion-probability admitting thereby unbiased variance estimation. Numerical evidences are presented to examine relative efficacies of some of the competing procedures

relevant to CSSPPS suggesting possibilities of improvements over the conventional one by certain alternatives.

3.1 Introduction

As, for example, with the Indian Annual National Sample Surveys (NSS), a very common practice of sample selection is Circular Systematic Sampling (CSS) with 'probabilities proportional' to known measures of 'sizes' (PPS) of the units in estimating a survey population total. This scheme, except in rare circumstances, fails to ensure a positive inclusion-probability of every pair of population units. Because of this, for a linear estimator of a population total a design unbiased variance estimator cannot exist. A standard way out is to repeat this 'CSSPPS' scheme independently twice. Following Das (1982) and Ray and Das (1997) we modify this CSSPPS scheme allowing the 'sampling interval' to be chosen at 'random' out of a specified set of positive integers rather than 'keeping it fixed', as described in Section 3.2. This modified scheme CSSPPS(M), say, ensures positive inclusion-probability of every pair of units and thereby admits unbiased variance estimation.

In the original CSSPPS and also in CSSPPS(M) schemes the number of distinct units realized may be less than the number of draws which is the intended effective sample-size. Also, the inclusion probabilities of the units need not be proportional to their size measures. Consequently, for both, to employ the usual Horvitz-Thompson's (HT,1952) estimator one has to calculate the inclusion-probabilities directly on counting the numbers of systematic samples containing the respective units. Even in this computer age this calculation is a nontrivial problem if the sum of the size-measures of all the population units is very large, as is commonly the case in practice. So, to judge the efficiency of CSSPPS(M) versus CSSPPS as a reasonable course, we resort to numerical exercises illustrated in section 3.3. The specimens presented suggest that in terms of accuracy in estimation CSSPPS(M) competes quite well against the currently common CSSPPS scheme independently applied twice.

3.2 Alternative Circular Systematic Sampling Schemes and Respective Estimators

Let $U = (1, \cdots, i, \cdots, N)$ denote a finite survey population. Let y be a variable of interest with values $y_i, i\epsilon U$ and x a correlated variable with values x_i which are positive integers, $i\epsilon U$. By Σ we shall denote summing over i in U and by $\underset{i \neq j}{\Sigma\Sigma}$ that over $i, j (i \neq j)$ in U. Our focus is on estimating $Y = \Sigma y_i$ using the values y_i for i in a sample s chosen with a probability $p(s)$ adopting a design P, say. In choosing a design or scheme of sample selection here we shall use the values $p_i = \frac{x_i}{X}$, called the 'normed size-measures' with X as Σx_i.

3.2.1 CSSPPS Scheme of Sample Selection

From Murthy (1967) we know that the standard "circular systematic sampling with probability proportional to size" (CSSPPS) scheme is applied in the following way. (i) The intended number of distinct units to be realized in a sample s to be selected, called the effective size of the sample, is fixed at an integer $n (1 \leq n < N)$, (ii) a positive integer k, called the 'Sampling interval' is suitably chosen - conventionally it is fixed either at $[\frac{X}{n}]$, the hightest integer not exceeding $\frac{X}{n}$ or at $[\frac{X}{n}] + 1$; (iii) an integer R is chosen at random from the interval $(1, X)$; (iv) the positive integers

$$a_r = (R + rk) \text{ mod } (X), \quad r = 0, 1, \cdots, (n-1)$$

are calculated ; (v) defining $C_0 = 0, C_i = \sum_{j=1}^{i} x_j, i = 1, \cdots, N$, one ascertains the units i in U for which $C_{i-1} < a_r \leq C_i$ are satisfied, $r = 0, 1, \cdots, n-1$. Then, s, the sample generated for this CSSPPS scheme, consists of the units ascertained in "(v)"applying the convention that if a_r equals 0, then the "unit N " is to be taken in s.

For any design $P, \pi_i = \sum\limits_s p(s) I_{si}$ and $\pi_{ij} = \sum\limits_s p(s) I_{sij}$ respectively denote the inclusion-probability of i and of (i, j) ; $I_{si} = 1$ if $i \epsilon s, 0$ if $i \not\epsilon s$; $I_{sij} = I_{si} I_{sj}$.

A design P is an IPPS ('inclusion probability proportional to size') design if $\pi_i \ \alpha \ x_i, \ i \epsilon U$. By $\nu(s)$ we shall denote the effective size of a sample s. It is well known that

$\Sigma \pi_i = \sum\limits_s \nu(s) p(s) = \nu$, say, the expected value of $\nu(s)$. If $\nu(s) = n$ for every s with $p(s) > 0$, then P is a 'fixed effective sample size' design.

If for the CSSPPS design described above (i) $\nu(s)$ equals n and (ii) $\pi_i = np_i$ for every i in U, then it is a truly 'IPPS' design. But "(ii)" obviously cannot hold if $np_i > 1$ for i in U. We shall presently illustrate $\underline{X} = (x_i, \cdots, x_i, \cdots, x_N)$ and n for which both (i) and (ii) may be violated.

Illustration 1. $N = 13, \underline{X} = (6, 5, 6, 7, 14, 5, 6, 6, 11, 9, 7, 13, 5), X = 100, n = 7, k = [\frac{X}{n}] + 1 = 15$. Here $np_i < 1 \quad \forall i \ \epsilon \ U = (1, \ldots, 13)$.

Table 1

Showing descrepancies in $(i)\nu(s), n$ and (ii) π_i, np_i

i	1	2	3	4	5	6	7	8	9	10	11	12	13
π_i	.42	.35	.42	.49	.94	.35	.42	.42	.76	.63	.49	.88	.35
$\delta_i = \pi_i - np_i$	0	0	0	0	-.04	0	0	0	-.01	0	0	-.03	0

Thus, $np_i < 1 \forall i, \Sigma \pi_i = 6.92, \nu(s)$ equals 6 for some and 7 for other samples.

Illustration 2. $N = 19, \underline{X} = (34, 1, 9, 3, 2, 1, 22, 2, 5, 2, 19, 5, 10, 2, 3, 19, 7, 21, 11)$. $X = 178, n = 8, k = [\frac{X}{n}] + 1 = 23; \quad np_1 = 1.5280 > 1, np_i < 1, i = 2, \cdots, 19$.

Table 2

Showing discrepancies in π_i, np_i.

i	π_i	$\delta_i = \pi_i - np_i$	i	π_i	δ_i	i	π_i	δ_i	i	π_i	δ_i
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	1.00	-.5280	4	.1348	0	7	.9606	-.0281	10	.0899	0
2	.0449	0	5	.0898	0	8	.0899	0	11	.8427	-.0112
3	.4044	0	6	.0449	0	9	.2447	0	12	.2247	0
13	.4494	0	14	.0899	0	15	.1348		16	.8426	-.0113
17	.3146	0	18	.9213	-.0225	19	.4943	0			

Here $\Sigma\pi_i = 7.3988$; $\nu(s)$ takes the values 6,7 and 8.

Remark 1. In case $np_i < 1 \quad \forall i \epsilon U$, Hartley and Rao (1962) have given a modified CSSPPS scheme for which the units of U are first randomly permuted and the CSSPPS scheme as described above is applied on the realized permutation of the elements of U. For this scheme, (a) $\nu(s) = n$ for every s, (b) $\pi_i = np_i \forall i$ and (c) $\pi_{ij} > 0$ for every $i, j (i \neq j)$ in U. For CSSPPS, "(c)" is known to be violated and (a), (b) also may not hold as illustrated above but it is applicable even if (d) "$np_i < 1 \quad \forall i$" is violated as is exemplified above too.

For a population total $Y = \Sigma y_i$ the well-known Horvitz and Thompson's (HT, 1952) estimator is

$t_H = \Sigma \frac{y_i}{\pi_i} I_{si} = \Sigma' \frac{y_i}{\pi_i}$, writing Σ' as sum over i in s.

Here it is assumed that "$\pi_i > 0 \ \forall \ i$"- a well-known 'necessary condition' for the existence of an unbiased estimator for Y. We have, however, the

Theorem 1. For CSSPPS, $\pi_i > 0 \quad \forall i \epsilon U$.

Proof : Since $x_i > 0$, Prob $[C_{i-1} < a_r \leq C_i] > 0 \ \forall \ i$ at least for one $r, r = 0, 1, \cdots, n-1$, if $n \geq 1$.

In practice, especially for example in Indian NSS, even without checking whether (a) $\nu(s) = n \; \forall \; s$ with $p(s) > 0$ and/or (b) $\pi_i = np_i < 1 \; \forall \; i$,

$$t = \frac{1}{n}\Sigma' \frac{y_i}{p_i} \tag{3.1}$$

is taken as an estimator for Y based on a CSSPPS, disregarding the possibility of the bias in t when $\pi_i \neq np_i \forall i$". Since for CSSPPS, π_{ij} may be zero for some $i, j (i \neq j)$ and hence an unbiased estimator may be unavailable for $V(t)$, the variance of t, a convention is to draw 2 samples s_1, s_2, say, from U independently applying the same CSSPPS scheme. Then, calculating t in (3.1) for these samples and denoting them by t_1, t_2,

(i) $\bar{t} = \frac{1}{2}(t_1 + t_2)$ is used to estimate Y and

(ii) $v = \frac{1}{4}(t_1 - t_2)^2$ is used to unbiasedly estimate $V(\bar{t})$.

For any two designs $P_k, k = 1, 2$ with inclusion-probabilities $\pi_i(k), \pi_{ij}(k), k = 1, 2$, if two samples s_1 and s_2 are 'independently' drawn, then the pooled sample $\bar{s} = (s_1, s_2)$ has the selection-probability, say,

$$p_0(\bar{s}) = p_1(s_1)p_2(s_2).$$

Writing $\pi_i(0), \pi_{ij}(0)$ as the inclusion-probabilities for this design P_0, say, giving $p_0(\bar{s})$, we have the

Theorem 2. If $0 < \pi_i(k) < 1 \quad \forall i, k = 1, 2$ then

(i) $\pi_i(0) = 1 - (1 - \pi_i(1))(1 - \pi_i(2)) > 0 \quad \forall i$

(ii) $\pi_{ij}(0) = 1 - [(1 - \pi_i(1))(1 - \pi_i(2)) + (1 - \pi_j(1))(1 - \pi_j(2)) - (1 - \pi_i(1) - \pi_j(1) + \pi_{ij}(1))(1 - \pi_i(2) - \pi_j(2)) + \pi_{ij}(2))] > 0 \quad \forall i, j.$ in $U(i \neq j)$

Proof: (i) is obvious. For (ii) note that
$\pi_{ij}(0) = \pi_{ij}(1)(1 - \pi_i(2)) + \pi_{ij}(2)(1 - \pi_i(1)) + \pi_j(1)(\pi_i(2) - \pi_{ij}(2)) + \pi_j(2)(\pi_i(1) - \pi_{ij}(1)) + \pi_{ij}(1)\pi_{ij}(2) > 0$

by the 'hypothesis' and on observing that $\pi_{ij}(k) \leq \pi_i(k)$ for every $j(\neq i)$ and every i in U, for $k = 1, 2$.

For the design P_0 corresponding to 2 independent drawings by the CSSPPS method we shall write θ_i for the inclusion-probability of i and θ_{ij} for $i, j(i \neq j)$ in U. Then, as an alternative to t in (3.1) it seems proper to use

$$t_H = \Sigma' \frac{y_i}{\theta_i} \tag{3.2}$$

as the 'unbiased' HT estimator for Y. Since $\theta_{ij} > 0 \; \forall \, i, j(i \neq j)$ an unbiased estimator exists for $V(t_H)$.

3.2.2 CSSPPS with 'random sampling interval'

In the context of Circular Systematic Sampling (CSS) with equal probabilities, namely the special case when $x_i = 1 \forall \, i, i \epsilon U$, Das (1982) and Ray and Das (1997) recommended the choice of a 'random sampling interval' k as a number to be chosen at random as an integer between 1 and $(N-1)$ in order to get over the problem of unbiased variance estimation inherent in the CSS with a fixed k.

Inspired by this, let us introduce the "Modified CSSPPS" scheme, denoted by CSSPPS(M), which is same as CSSPSS of section 3.2.1 except that in a_r, we take "k as an integer chosen at random between 1 and $(X - 1)$". Then, we have the

Theorem 3. For CSSPSS(M),

(i) $\pi_i > 0 \; \forall i$ and $\forall n \geq 1$

(ii) $\pi_{ij} > 0 \; \forall \; i, j(i \neq j)$ provided "$n \geq 2$"

Proof: Note first that

π_i=[Number of samples out of all possible X(X-1) samples for which $C_{i-1} < (R + rk) mod(X) \leq C_i, R = 1, .., X, k = 0, 1, .., (X - 1), r =$

$0, 1, .., (n-1)]/X(X-1)$
and similarly
π_{ij}=[Number of samples out of all possible X(X-1) samples for which $C_{i-1} < (R+rk)mod(X) \leq C_i, C_{j-1} < (R+rk)mod(X) \leq C_j$ with R, k, r as in π_i above]$/X(X-1)$.
(i)Then since $x_i > 0$, Prob $[C_{i-1} < a_r \leq C_i] > 0$ for every i and for every $r = 0, 1, \cdots, n-1$, since $0 \leq a_r \leq X-1$ and $1 \leq k \leq X-1$. So, $\pi_i > 0 \forall i$ if $n \geq 1$. (ii) Since $x_i > 0 \forall i, 0 \leq a_0 \leq X-1$ and $1 \leq k \leq X-1$ and $n \geq 2$,

Prob $[C_{i-1} < a_r \leq C_i, C_{j-1} < a_{r+1} \leq C_j] > 0 \forall i, j (i \neq j)$ in U and $\forall r = 0, 1 \cdots, (n-1)$. So $\pi_{ij} > 0 \forall i, j (i \neq j)$ in U if $n \geq 2$.

For the variance of t_H given by
$V(t_H) = \Sigma y_i^2 \frac{1-\pi_i}{\pi_i} + \underset{i \neq j}{\Sigma\Sigma} y_i y_j (\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j})$,
based on any design 'with variable effective sample-sizes' the following two unbiased estimators are available based on (I) "CSSPPS - repeated twice" for which π_i will be understood as θ_i and π_{ij} as θ_{ij} throughout and on (II) CSSPPS(M), namely

$$v_1 = \Sigma y_i^2 (\frac{1-\pi_i}{\pi_i}) \frac{I_{si}}{\pi_i} + \underset{i \neq j}{\Sigma\Sigma} y_i y_j (\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}) \frac{I_{sij}}{\pi_{ij}}$$

given by HT (1952), and

$$v_2 = \frac{1}{2} \underset{i \neq j}{\Sigma\Sigma} (\pi_i \pi_j - \pi_{ij}) (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 \frac{I_{sij}}{\pi_{ij}} + \Sigma \frac{y_i^2}{\pi_i} (1 + \frac{1}{\pi_i} \sum_{j \neq i} \pi_{ji} - \Sigma \pi_i) \frac{I_{si}}{\pi_i}$$

due to Chaudhuri (2000a), as discussed in Chapter 1.

One advantage of v_2 over v_1 is that it is easier to check for its 'uniform non-negativity' for a given design.

Remark 2. The first term of v_2 is the well-known Yates and Grundy's (YG, 1953) estimator for $V(t_H)$ but this is 'biased' for 'CSSPPS - repeated twice' and CSSPPS(M) because the effective sample-sizes for these two schemes are "variables".

For Y we propose another unbiased estimator as

$$e = \Sigma y_i \frac{f_{is}}{E(f_{is})}, \tag{3.3}$$

writing f_{is} for the number of times i occurs in s.

Writing $V_i = V(f_{is}), C_{ij} =$ Cov (f_{is}, f_{js}) and $F_i = E(f_{is})$, an unbiased estimator of the variance of e is

$$v_3 = \Sigma y_i^2 \frac{V_i}{F_i^2} \frac{I_{si}}{\pi_i} + \sum_{i \neq j}\Sigma y_i y_j \frac{C_{ij}}{F_i F_j} \frac{I_{sij}}{\pi_{ij}}$$

We give an example below to demonstrate that for a CSSPPS(M), $\nu(s)$ may vary and differ from n.

Illustration 3. $N = 19, \underline{X}$=(34,1,1,3,2,1,7,2,2,3,5,2,3,5,3,4,7,4,6). $X = 95, n = 6$. By s^* we shall denote the set of distinct units in s.

Table 3

Showing discrepancies in $\nu(s)$ vis-a-vis n in CSSPPS(M)

R, k	(3, 63)	(17, 48)
$(a_0, a_1, a_2, a_3, a_4, a_5)$	(3,66,34,2,65,33)	(17,65,18,66,19,67)
s^*	(1, 13)	(1,1,4,13)
$\nu(s)$	2	3

3.3. Numerical evaluation of comparative efficacies

With $\hat{Y}$ as an estimator for Y and $\hat{V}$ as an estimator of the variance of $\hat{Y}$, the pivotal

$$d = (\hat{Y} - Y)/\sqrt{\hat{V}}, \text{ provided } \hat{V} > 0,$$

is treated, in practice, for large samples, as a standardized normal deviate ignoring the resulting error. Then, $(\hat{Y} - 1.96\sqrt{\hat{V}}, \hat{Y} + 1.96\sqrt{\hat{V}})$ is treated as a 95% confidence interval (CI) for Y. Taking the pairs $(\bar{t}, v), (t_H, v_1), (t_H, v_2)$, (e, v_3) successively for $(\hat{Y}, \hat{V})$, we present some numerical evidences, through some simulations based on live data, about their relative performances. As performing criteria we consider $T = 1000$ replicated samples by a specified procedure, (A) the average length (AL) of the CI's, (B) the percent of replicates for which, the CI's cover Y – called the 'Actual Coverage percentage' (ACP) – the closer it is to 95 the better and (C) the average coefficient of variation (ACV) i.e. the average over the $T = 1000$ and also for the $T = 10,000$ replicates of the values of $100(\frac{\sqrt{\hat{V}}}{\hat{Y}})$ – the smaller it is the greater the accuracy of $\hat{Y}$ as well as the shorter the CI. Since it is possible that $v_k (k = 1, 2, 3)$ may turn out negative for a sample we also indicate the 'number of replicates for which they turn out negative' coded by 'NEG' in the tables – in calculating ACP, ACV the 'replicates with negative variance estimates' are omitted. The actual number of replicates is denoted by AT=T-NEG.

The live data we use in our calculations are (1) N=50 clusters along with their (2) Population figures in 1985 (y) and (3) the number of municipalities or $(3)'$ the population size in 1975 taken as x respectively coded as $x(M)$ or x (1975) as are reported in Särndal, Swensson, Wretman (SSW, 1992, pp. 660-661.)

In our numerical calculations we split the population into 4 strata of sizes $N_h (h = 1, \cdots, 4)$ and draw samples of sizes $n_h (h = 1, \cdots, 4)$ from the respective strata, calculate strata-wise estimates and variance estimates and adding the estimates across the strata derive estimates for the population total along with the variance estimates. The strata are formed by the 4 consecutively numbered clusters as are given in the Text by Särndal, Swensson and Wretman (SSW, 1992), pp. 660-661. In keeping, CSSPPS(M) closely comparable to 'CSSPPS repeated twice' we take n_h''s in the former as n_h's which are closest integers to $\Sigma\theta_i$ - strata-wise.

Table 4

Relative efficacies of competing procedures

(I) $N = 50; N_1 = 11, N_2 = 16, N_3 = 10, N_4 = 13; n_1 = 3, n_2 = 7, n_3 = 3, n_4 = 6; Y = 8339, x = x(M); n'_1 = 5, n'_2 = 10, n'_3 = 5, n'_4 = 9; X = 284.$

In this case $\pi_i = np_i \; \forall i$ in each stratum.
For $T = 1000$.

Sampling	$(\hat{Y}, \sqrt{\hat{V}})$ - values	Based on			
scheme	for 1 typical replicate	AT=T-NEG			
		replicates			
		AL	ACP	ACV	NEG
(1)	(2)	(3)	(4)	(5)	(6)
CSSPPS	$(\bar{t}, \sqrt{v}) = (10817.41, 2592.11)$	9857.36	89	22	0
repeated	$(t_H, \sqrt{v_1}) = (8320.73, 1031.38)$	5638.74	86	14	15
twice	$(t_H, \sqrt{v_2}) = (- - -, 948.77)$	5096.07	94	15	8
	$(e, \sqrt{v_3}) = (8733.31, 1543.78)$	4062.69	88	12	66
CSSPPS(M)	$(t_H, \sqrt{v_1}) = (8726.34, 1186.32)$	5984.22	91	18	0
	$(t_H, \sqrt{v_2}) = (- - -, 1267.08)$	5881.09	96	15	0
	$(e, \sqrt{v_3}) = (9995.82, 780.92)$	3782.03	90	10	6

(I′) For $T = 10,000$.

Sampling	$(\hat{Y}, \sqrt{\hat{V}})$ - values	Based on			
scheme	for 1 typical replicate	AT=T-NEG			
		replicates			
		AL	ACP	ACV	NEG
(1)	(2)	(3)	(4)	(5)	(6)
CSSPPS	$(\bar{t}, \sqrt{v}) = (9219.05, 2713.01)$	8803.25	92.8	22.82	0
repeated	$(t_H, \sqrt{v_1}) = (8252.27, 1039.70)$	4532.44	92.4	14.06	0
twice	$(t_H, \sqrt{v_2}) = (- - -, 1001.39)$	4201.59	93.5	14.27	21
	$(e, \sqrt{v_3}) = (8481.25, 1103.50)$	4301.03	95.9	10.78	34
CSSPPS(M)	$(t_H, \sqrt{v_1}) = (8575.11, 1091.88)$	5301.58	93.0	18.80	0
	$(t_H, \sqrt{v_2}) = (- - -, 1046.25)$	5305.51	94.4	13.20	0
	$(e, \sqrt{v_3}) = (8939.25, 803.50)$	3616.94	89.3	13.50	8

(II) $N = 50; N_1 = 11, N_2 = 16, N_3 = 9, N_4 = 14; n_1 = 3, n_2 = 7, n_3 = 3, n_4 = 6; Y = 8339, x = x(1975); n'_1 = 5, n'_2 = 9, n'_3 = 4, n'_4 = 8; X = 7980.$
For $T = 1000$.

Sampling scheme	$(\hat{Y},\sqrt{\hat{V}})$ - values for 1 typical replicate	Based on AT=T-NEG replicates			
		AL	ACP	ACV	NEG
(1)	(2)	(3)	(4)	(5)	(6)
CSSPPS repeated twice	$(\bar{t},\sqrt{v}) = (8749.24, 416.61)$	7204.48	99	18	0
	$(t_H,\sqrt{v_1}) = (7559.19, 978.03)$	2576.65	90	9	11
	$(t_H,\sqrt{v_2}) = (- - -, 656.24)$	2493.88	93	9	0
	$(e,\sqrt{v_3}) = (8195.99, 416.56)$	1466.49	90	7	95
CSSPPS(M)	$(t_H,\sqrt{v_1}) = (8853.26, 959.71)$	3394.45	93	11	0
	$(t_H,\sqrt{v_2}) = (- - -, 750.92)$	3356.09	94	11	0
	$(e,\sqrt{v_3}) = (8320.50, 630.11)$	2129.01	91	7	63

(II′) For $T = 10,000$.

Sampling scheme	$(\hat{Y},\sqrt{\hat{V}})$ - values for 1 typical replicate	Based on AT=T-NEG replicates			
		AL	ACP	ACV	NEG
(1)	(2)	(3)	(4)	(5)	(6)
CSSPPS repeated twice	$(\bar{t},\sqrt{v}) = (9378.24, 4120.74)$	6903.58	96.2	22.87	0
	$(t_H,\sqrt{v_1}) = (8375.10, 980.64)$	2623.06	94.9	11.37	10
	$(t_H,\sqrt{v_2}) = (- - -, 899.35)$	2215.39	95.5	11.59	5
	$(e,\sqrt{v_3}) = (9724.83, 603.50)$	1632.24	91.8	10.76	15
CSSPPS(M)	$(t_H,\sqrt{v_1}) = (8286.21, 1031.88)$	2301.58	92.0	18.70	0
	$(t_H,\sqrt{v_2}) = (- - -, 1021.37)$	2305.60	92.0	18.70	0
	$(e,\sqrt{v_3}) = (7667.11, 931.51)$	2616.94	84.5	10.10	15

(III) $N = 50; N_1 = 11, N_2 = 16, N_3 = 10, N_4 = 13$. $n_1 = 3, n_2 = 7, n_3 = 6, n_4 = 8, x = x(M); n'_1 = 5, n'_2 = 10, n'_3 = 8, n'_4 = 10; Y = 8339, X = 284$.
Here $\pi_i \neq np_i$ for every stratum.
For $T = 1000$.

Sampling scheme	$(\hat{Y}, \sqrt{\hat{V}})$ - values for 1 typical replicate	Based on AT=T-NEG replicates AL	ACP	ACV	NEG
(1)	(2)	(3)	(4)	(5)	(6)
CSSPPS	$(\bar{t}, \sqrt{v}) = (11486.73, 688.62)$	8046.19	83	19	0
repeated	$(t_H, \sqrt{v_1}) = (7912.92, 1733.61)$	4530.19	88	14	2
twice	$(t_H, \sqrt{v_2}) = (- - -, 387.52)$	4113.29	94	14	8
	$(e, \sqrt{v_3}) = (8794.72, 1326.64)$	3874.73	94	11	65
CSSPPS(M)	$(t_H, \sqrt{v_1}) = (9037.71, 2053.08)$	5579.18	93	19	0
	$(t_H, \sqrt{v_2}) = (- - -, 1819.31)$	5121.06	96	16	0
	$(e, \sqrt{v_3}) = (9208, 1035.57)$	3238.11	91	9	55

(III′) For $T = 10,000$.

Sampling scheme	$(\hat{Y}, \sqrt{\hat{V}})$ - values for 1 typical replicate	Based on AT=T-NEG replicates AL	ACP	ACV	NEG
(1)	(2)	(3)	(4)	(5)	(6)
CSSPPS	$(\bar{t}, \sqrt{v}) = (9764.94, 3185.07)$	8321.97	89.1	20.05	0
repeated	$(t_H, \sqrt{v_1}) = (8465.22, 1120.27)$	4364.30	89.5	13.56	22
twice	$(t_H, \sqrt{v_2}) = (- - -, 1147.70)$	4584.48	90.1	13.69	60
	$(e, \sqrt{v_3}) = (7651.51, 923.51)$	3581.39	93.9	10.49	126
CSSPPS(M)	$(t_H, \sqrt{v_1}) = (6481.28, 1109.04)$	6291.58	93.0	18.70	0
	$(t_H, \sqrt{v_2}) = (- - -, 1204.82)$	6293.46	93.0	18.80	0
	$(e, \sqrt{v_3}) = (7633.85, 1003.56)$	3522.01	86.1	9.50	73

The picture with $T = 10,000$ replicates does not show any marked deviations from the conclusions reachable from the Table 4 with $T = 1000$ replicates.

We add Table 5 below to present our simulation-based study of sampling errors relevant to the context.

Table 5

Illustrating relative performances of competing strategies in terms of an estimated standard error of Percentage error, CV $= 100.\sqrt{\frac{1}{T}\Sigma(e' - \bar{e'})^2}$, where $e' = \frac{(t-Y)}{Y}.100$ and B $= \bar{e'} = \frac{1}{T}\sum_{1}^{T}(\frac{t-Y}{Y}).100$, an average (over replicates) of relative sampling error $\bar{e'}$.

(I′) For the population (I) (as mentioned above), $T = 10,000$.

Sampling scheme	$(\hat{Y}, \sqrt{\hat{V}})$	NEG	CV	B
(1)	(2)	(3)	(4)	(5)
CSSPPS	$(\bar{t}, \sqrt{v})$	0	16.40	11.2
repeated	$(t_H, \sqrt{v_1})$	0	14.30	.65
twice	$(t_H, \sqrt{v_2})$	21	14.60	.68
	$(e, \sqrt{v_3})$	34	16.40	1.24
CSSPPS(M)	$(t_H, \sqrt{v_1})$	0	18.50	1.74
	$(t_H, \sqrt{v_2})$	0	18.50	1.74
	$(e, \sqrt{v_3})$	8	24.70	2.09

(II′) For the population (II) (as mentioned above), $T = 10,000$.

Sampling scheme	$(\hat{Y}, \sqrt{\hat{V}})$	NEG	CV	B
(1)	(2)	(3)	(4)	(5)
CSSPPS	$(\bar{t}, \sqrt{v}))$	0	16.95	2.1
repeated	$(t_H, \sqrt{v_1})$	0	14.78	.31
twice	$(t_H, \sqrt{v_2})$	5	15.02	.33
	$(e, \sqrt{v_3})$	15	16.76	.60
CSSPPS(M)	$(t_H, \sqrt{v_1})$	0	18.60	.02
	$(t_H, \sqrt{v_2})$	0	18.60	.02
	$(e, \sqrt{v_3})$	15	24.65	.03

(III′) For the population (III)(as mentioned above), $T = 10,000$.

Sampling scheme	$(\hat{Y}, \sqrt{\hat{V}})$	NEG	CV	B
(1)	(2)	(3)	(4)	(5)
CSSPPS	$(\bar{t}, \sqrt{v}))$	0	15.61	26.9
repeated	$(t_H, \sqrt{v_1})$	22	13.61	.16
twice	$(t_H, \sqrt{v_2})$	60	13.88	.16
	$(e, \sqrt{v_3})$	126	15.84	.60
CSSPPS(M)	$(t_H, \sqrt{v_1})$	0	18.50	1.73
	$(t_H, \sqrt{v_2})$	0	18.50	1.73
	$(e, \sqrt{v_3})$	73	24.77	1.96

This Table 5 gives some ideas about simulation-based sampling errors which appear to be mostly under control.

For the traditional CSSPPS scheme repeated twice, the pair (t_H, v_2) seems to beat the traditional $(\bar{t}, v)$ for all the three examples I–III; (e, v_3) with too many negative values of v_3 ‘replicate-wise’ is not quite a viable alternative. CSSPPS(M) can be treated just as a competitor against ‘CSSPPS – repeated twice’ – it cannot be labelled ‘superior’ through what is revealed above. However, compared to the current practice of using $(\bar{t}, v)$ all the other newly proposed pairs $(t_H, v_1), (t_H, v_2), (e, v_3)$ all based on ‘CSSPPS-repeated twice’ seem to perform better. Further, the same pairs based on CSSPPS(M) compete well against those based on ‘CSSPPS – repeated twice’ and are superior to $(\bar{t}, v)$.

Remark 3. A possible alternative to CSSPPS and CSSPPS(M) described above may be to “Continue drawing units’ till a pre-determined effective sample-size ν is realized. From the example below it may be checked that the number of draws may far exceed ν. By CSSPPS(R) and CSSPPS (M,R) we denote these revised schemes. However, we do not pursue with this approach further for the present.

Table 6

Showing 'Number of draws' versus 'Effective sample size'.

$N = 6, \underline{X} = (18, 1, 2, 15, 8, 6), X = 50, \nu = 4.$

	CSSPPS(R)	CSSPPS(M,R)
	$R = 1$	$R = 1, k = 28$
a_r $r = 0, 1, 2, \cdots$	(1,14,27,40,3,16,29, (42,5,18,31,44,7,20)	(1,29,7,35,41,19)
s	(1,1,4,5,1,1,4,5, 1,1,4,5,1,3)	(1,4,1,4,1,5,2)
s^*	(1,4,5,3)	(1,2,4,5)
$\nu(s)$	4	4

Finally, we illustrate that CSSPPS(M) is also really not an 'inclusion probability proportional to size' scheme of sampling so that detailed computations are needed to evaluate π_i by counting the samples out of the total number $X(X-1)$ of possible samples that contain the respectively specified units i in U.

Table 7

Showing deviations $D_i = \pi_i - np_i, i\epsilon U$ CSSPPS(M).

$N = 13; \underline{X} = (6, 5, 6, 5, 10, 5, 6, 12, 4, 6, 7, 9, 5), X = 86$

i	$D_i = \pi_i - np_i$	i	$D_i = \pi_i - np_i$	i	$D_i = \pi_i - np_i$	i	$D_i = \pi_i - np_i$
(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1	.0167	4	.0234	7	.0169	10	.0169
2	.0736	5	-.0520	8	-.1117	11	.0051
3	.0169	6	.0235	9	.0273	12	-.0299
13	.0235						

Remark 4. From the computational point of view it must be emphasized that for our proposed procedures only $\pi_i, \theta_i, \pi_{ij}, \theta_{ij}$ are needed for $i \epsilon s$ and $i \neq j$ in s. This is a great relief if $\nu(s)$ is modest.

3.3 Conclusion and Recommendation.

Kunte, Sudhakar (1978) is, to our knowledge, the first to point out that each draw may not yield always a distinct unit in systematic sampling. Though called a 'probability proportional to size' scheme of sampling, CSSPPS may not ensure 'inclusion-probability proportional to size'. With a single random start unbiased variance estimation with systematic sampling is 'always possible' only by 'CSSPPS(M)'; for Hartley and Rao's (1962) CSSPPS with a prior randomization it is possible only if "$np_i < 1 \quad \forall i$". Whenever "π_i differs from proportionality to p_i" computation of π_i is difficult in practice. With CSSPPS repeated twice the standard estimator is 'biased'. For the proposed CSSPPS(M) one may, as a third alternative, employ (e, v_3), but v_3 often turns out negative and its calculation is also not quite easy. So, research on systematic sampling promises to continue for better solutions.

Chapter 4

Estimating numbers/proportions of people with stigmatizing features from randomized responses by specific devices through complex survey sampling

Abstract

Warner(1965) gave a standard procedure for estimating the proportion of people bearing a sensitive characteristic, say, A like tax evasion, alcoholism etc. in a given community when a simple random sample(SRS) is chosen with replacement (WR) in a specified number of draws on eliciting a 'randomized response' (RR) rather than a 'direct response'(DR) of 'Yes' or 'No' truthfully about 'bearing A', on adopting a suitable randomization device. Many ramifications followed to effect improvements and/or to counter specific ex-

igencies. Chaudhuri (1987), Chaudhuri and Mukerjee (1988) among others gave general accounts of RR including situations needed to cover qualitative as well as quantitative characteristics when sampling may be 'simple' or 'complex' as well. We shall here follow Chaudhuri (1999, 2000b) to (1) show how Sarjinder Singh and Anwar H.Joarder's (1997) results given for SRSWR may extend to unequal probability sampling without replacement(WOR) and (2) develop similar and additional results concerning RR given by Franklin (1989a, 1989b) and Singh and Singh (1992, 1993) for SRSWR, when one extends to complex survey designs. Arnab (1996, 2000) also dealt with certain aspects of Franklin's and Singh and Singh's work and other related RR's but ours are different from his developments.

4.1 Introduction

In estimating the proportion θ of people bearing a stigmatizing characteristic A like habitual tax evasion, drunken driving, gambling etc it is well-known that Warner (1965) considered it useful to avoid seeking direct responses (DR) from respondents in a social survey. Instead he gave us a randomized response (RR) technique by way of protecting the respondent's privacy. According to this a sampled respondent is to implement a randomizing device by which with a pre-assigned probability p $(0 < p < 1)$ a truthful response is to be 'Yes' or 'No' about bearing A and with probability $(1 - p)$ about bearing the complementary characteristic $\bar{A}$ without divulging to the interviewer whether the response relates to A or $\bar{A}$.

Based on such RR's procured from an SRSWR chosen in n draws an unbiased estimator for θ and an unbiased estimator for its variance are given by Warner (1965). Singh and Joarder (1997) recommend a modification of Warner's RR procedure enjoining a (I) respondent bearing $\bar{A}$ to respond as in Warner's case but a (II) respondent bearing A to postpone the response to a second performance of Warner's randomizing device unless the first one induces a 'Yes' response.

With such responses from an SRSWR in n draws they prescribe a better unbiased estimator for θ along with an unbiased variance estimator.

Though the fact is not made explicit by these authors θ here is a 'finite survey population mean' of an 'Indicator' variable which is valued 1 for a population unit bearing A and 0 for one with $\bar{A}$. But in practice a finite population survey is implemented according to complex designs involving selection in multi-stages and through stratification with sampling in the early stages with unequal selection-probabilities. A sample survey in practice covers numerous, say, fifty items of enquiry of which only a few, say, five may relate to sensitive issues. From such a single survey one must derive good estimators based on 'direct responses' (DR) related to innocuous characteristics and those based on RR's related to the sensitive ones. So, we consider it important to present a theory to show how θ above may be estimated admitting variance estimates when RR's are obtained by Warner's (1965) and Singh and Joarder's (1997) techniques but the respondents are sampled by general sampling schemes with varying probabilities and without replacement.

After presenting revised methods of estimation we supplement Singh and Joarder's numerical findings with ours for the sake of comparison in sections 4.2 and 4.3.

Franklin (1989a, 1989b), Singh and Singh (1992, 1993) and Arnab (1996) consider repeated realizations of RR's from each sampled person to estimate numbers and proportions of people bearing a sensitive characteristic. We present a few further developments in sections 4.4 and 4.5.

4.2 Unbiased Estimators and Variance Estimators

According to Warner's RR device the probability for a 'Yes' response about the possession of the characteristic A or its complement $\bar{A}$ is

$$Y_W = p\theta + (1-p)(1-\theta) = (2p-1)\theta + (1-p) \tag{2.1}$$

The corresponding probability for Singh et al's scheme is

$$\begin{aligned} Y_{SJ} &= p\theta + p(1-p)\theta + (1-p)(1-\theta) \\ &= [(2p-1) + p(1-p)]\theta + (1-p) \\ &= Y_W + p(1-p)\theta. \end{aligned} \tag{2.2}$$

Writing n as the number of draws in SRSWR and m as the number of 'Yes' responses in either case we have:

Warner's well-known unbiased estimator for θ is

$$\hat{\theta}_W = (\tfrac{m}{n} - 1 + p)/(2p-1), \text{ taking } p \neq \tfrac{1}{2}. \tag{2.3}$$

Its variance and an unbiased estimator of the variance are:

$$V(\hat{\theta}_W) = \tfrac{Y_W(1-Y_W)}{n(2p-1)^2} = \tfrac{\theta(1-\theta)}{n} + \tfrac{p(1-p)}{n(2p-1)^2} \tag{2.4}$$

and

$$\begin{aligned} v_W &= \tfrac{m}{n}(1 - \tfrac{m}{n})/_{(n-1)(2p-1)^2} \\ &= \tfrac{\hat{\theta}_W(1-\hat{\theta}_W)}{(n-1)} + \tfrac{1}{4(n-1)}[\tfrac{1}{4(p-0.5)^2} - 1] \end{aligned} \tag{2.5}$$

Singh et al's unbiased estimator for θ is

$$\begin{aligned} \hat{\theta}_{SJ} &= [\frac{m}{n} - (1-p)]/[(2p-1) + p(1-p)], \\ &\quad \text{choosing its denominator non - zero.} \end{aligned} \tag{2.6}$$

Its variance and unbiased variance estimator are

$$\begin{aligned} V(\hat{\theta}_{SJ}) &= \tfrac{Y_{SJ}(1-Y_{SJ})}{n[(2p-1)+p(1-p)]^2} \\ &= \tfrac{\theta(1-\theta)}{n} + \tfrac{p(1-p)}{n[(2p-1)+p(1-p)]^2} \\ &\quad - \tfrac{\theta p(1-\theta)}{n[(2p-1)+p(1-p)]}, \end{aligned} \tag{2.7}$$

$$v_{SJ} \quad = \quad \frac{\frac{m}{n}(1-\frac{m}{n})}{(n-1)[(2p-1)+p(1-p)]^2}. \tag{2.8}$$

Singh et al's main theoretical result is:

$$\text{A.} \qquad V(\hat{\theta}_W) \geq V(\hat{\theta}_{SJ}) \text{ for every } p > 0.5. \tag{2.9}$$

Following Chaudhuri (1999, 2000b) we present below unbiased estimators for θ along with unbiased variance estimators based on RR's obtained by Warner's and Singh et al's devices when the respondents are sampled with unequal selection-probabilities.

Chaudhuri's (1999, 2000b) approach is the following. Let $U = (1, \cdots, i, \cdots, N)$ denote a finite survey population of a known number of N people labeled $i = 1, \cdots, N$. Let y be an indicator variable with its value y_i for i as

$$\begin{aligned} y_i &= 1 \text{ if } i \text{ bears A} \\ &= 0, \text{ otherwise.} \end{aligned}$$

Then, $\theta = \frac{1}{N}\Sigma y_i$, writing Σ as sum over $i \epsilon U$.

Let s be a sample from U chosen according to a design P with a selection-probability $p(s)$. By E_p, V_p we shall denote operators for expectation and variance with respect to P.

We suppose that y_i is not ascertainable for a person i in a sample but adopting a suitable RR device, from an i in a sample, an RR may be procured as r_i such that

$(i) E_R(r_i) = y_i$, $\quad (ii) V_R(r_i) = V_i(> 0)$, $\quad (iii) r_i$'s are independent over i in U and (iv) there exist v_i ascertainable from RR's such that $E_R(v_i) = V_i, i \epsilon U$.

Here E_R, V_R denote operators for expectation, variance with respect to RR devices. The over-all expectation and variance operators will be denoted by

$$E = E_p E_R = E_R E_p \text{ and } V = E_p V_R + V_p E_R = E_R V_p + V_R E_p.$$

Writing $I_{si} = 1$ if $i \epsilon s, 0$ if $i \not\epsilon s$, $I_{sij} = I_{si}I_{sj}$ let it be possible to choose b_{si}, d_{si} as constants free of $\underline{Y} = (y_1, \cdots, y_i, \cdots, y_N)$ and $\underline{R} = (r_1, \cdots, r_i, \cdots, r_N)$ such that

$$t_b = \frac{1}{N}\Sigma y_i b_{si} I_{si} \text{ subject to } E_p(b_{si}I_{si}) = 1 \ \forall i.$$

Then, $V_p(t_b) = \frac{1}{N^2}[\Sigma y_i^2 C_i + \sum\sum_{i \neq j} y_i y_j C_{ij}]$, where

$$C_i = E_p(b_{si}^2 I_{si}) - 1, C_{ij} = E_p(b_{si}b_{sj}I_{sij}) - 1.$$

Then $v_p(t_b) = \frac{1}{N^2}[\Sigma y_i^2 d_{si} I_{si} + \sum\sum_{i \neq j} y_i y_j d_{sij} I_{sij}]$ satisfies $E_p v_p(t_b) = V_p(t_b)$ provided d_{si}, d_{sij}'s are chosen subject to

$$E_p(d_{si}I_{si}) = C_i, E_p(d_{sij}I_{sij}) = C_{ij}.$$

The literature on 'Sample surveys' is full of numerous such possibilities of choices for $P, b_{si}, d_{si}, d_{sij}$'s. Since y_i's are not ascertainable, t_b is not available as an estimator for θ. So, Chaudhuri's (1999, 2000b) recommended unbiased estimator for θ based on RR is

$$e_b = \frac{1}{N}\Sigma r_i b_{si} I_{si} \text{ for which } E(e_b) = \theta.$$

Here e_b is just t_b with y_i's replaced by r_i's, $i \epsilon s$.

Similarly we should write $V_p(e_b)$ as $V_p(t_b)$ with y_i replaced by r_i for i in U and $v_p(e_b)$ as $v_p(t_b)$ with y_i replaced by $r_i, i \epsilon s$.

Two unbiased estimators for the variance $V(e_b)$, of e_b which is,

$$\begin{aligned} V(e_b) &= E_p V_R(e_b) + V_p E_R(e_b) \\ &= \tfrac{1}{N^2}(E_p[\Sigma V_i b_{si}^2 I_{si}]) + V_p(t_b) \end{aligned} \qquad (2.10)$$

$$\begin{aligned} &= E_R V_p(e_b) + V_R E_p(e_b) \\ &= E_R V_p(e_b) + \tfrac{1}{N^2}(V_R(\Sigma r_i)) \end{aligned} \qquad (2.11)$$

are

$$v(1) = v_p(e_b) + \frac{1}{N^2}(\Sigma v_i b_{si} I_{si}) \tag{2.12}$$

and

$$v(2) = v_p(e_b) + \frac{1}{N^2}[\Sigma v_i(b_{si}^2 - d_{si}) I_{si}]. \tag{2.13}$$

It is easy to check that

$$Ev(1) = V(e_b) = Ev(2). \tag{2.14}$$

In order to develop formulae corresponding to $e_b, v(1), v(2)$ for the specific RR devices by Warner (1965) and Singh et al (1997) based on a sample of r_i's for $i \epsilon s$ let us use the following notations.

$$\begin{aligned} \text{Let } I_i &= 1 \text{ if } i \text{ responds "Yes"} \\ &= 0, \text{ otherwise.} \end{aligned}$$

Then, for Warner's (1965) scheme r_i should be taken as

$$r_i = \frac{I_i - (1-p)}{(2p-1)} = r_i(W), \text{ say, for which } E_R(r_i(W)) = y_i, \tag{2.15}$$

with a variance, say, $V_i(W)$ as

$$\begin{aligned} V_i(W) &= V_R(r_i(W)) = \frac{1}{(2p-1)^2}[y_i(2p-1) + (1-p) \\ &\quad -(y_i(2p-1) + (1-p))^2] \\ &= \frac{p(1-p)}{(2p-1)^2}, \text{ noting } y_i = y_i^2. \end{aligned} \tag{2.16}$$

Since $V_i(W)$ does not involve any unknown parameters we need not seek any estimator $v_i(W)$, say, for it and use this $V_i(W)$ straightaway for v_i in (2.12) -(2.13).

For Singh et al's (1997) scheme, r_i should be taken as

$$r_i = \frac{I_i - (1-p)}{(2p-1)+p(1-p)} = r_i \text{ (SJ), say. Then, } E_R(r_i(SJ)) = y_i. \qquad (2.17)$$

Writing for simplicity, $\alpha = (2p-1)+p(1-p)$

we may work out the variance of $r_i(SJ)$ as, say,

$$\begin{aligned} V_i(SJ) &= V_R(r_i(SJ)) = \frac{1}{\alpha^2}[E_R(I_i)(1-E_R(I_i))] \\ &= \frac{1}{\alpha^2}[(\alpha y_i + (1-p) - (\alpha y_i + (1-p))^2] \\ &= \frac{1}{\alpha^2}[\beta y_i + p(1-p)], \text{ writing } \beta = \alpha(1-\alpha) - 2\alpha(1-p) \end{aligned}$$

Since β is thus known, this $V_i(SJ)$ may be estimated unbiasedly by

$v_i(SJ) = \frac{1}{\alpha^2}[\beta r_i + p(1-p)]$,

which may be used to replace v_i in (2.12),(2.13) in using $v(j), j = 1, 2$.

On simplifications we may check that

$$\begin{aligned} V_i(SJ) &= p(1-p)/\alpha^2 \text{ if } y_i = 0, \\ &= p(1-p)^2(2-p)/\alpha^2 \text{ if } y_i = 1. \end{aligned}$$

Writing $e_b(W), e_b(SJ)$ for e_b based respectively on Warner's (1965) and Singh et al's (1997) schemes and $V(e_b(W)), V(e_b(SJ))$ as their respective variances we have

Lemma 1. $V(e_b(W)) \geq V(e_b(SJ))$
if $V_i(W) \geq V_i(SJ)) \quad \forall i$.

Proof: Follows immediately from (2.10). Next we have

Lemma 2. $V_i(W) \geq V_i(SJ) \quad \forall i$ if $p \geq .4384$

Proof: $V_i(W) - V_i(SJ) = \frac{p(1-p)}{(2p-1)^2} - \frac{\beta y_i + p(1-p)}{\alpha^2}$

$= p(1-p)[\frac{1}{(2p-1)^2} - \frac{1}{((2p-1)+p(1-p))^2}]$ if $y_i = 0$

> 0 if $p > 0.4384$ as verifiable using Matlab , (i)

and $= p(1-p)[\frac{1}{(2p-1)^2} - \frac{(1-p)(2-p)}{((2p-1)+p(1-p))^2}]$ if $y_i = 1$

> 0 if $p > 0.4366$ as verifiable using Matlab. (ii)

Hence follows Lemma 2. Hence we have the

Theorem. $V(e_b(W)) \geq V(e_b(SJ)) \geq 0$ if $p > 0.4384$.

The next section presents a numerical study as a follow-up of Singh et al's exercise.

4.3 A Comparative Study with Numerical Illustrations

In order to maintain parity with Singh et al's (1997) numerical illustration let us make separately 9 alternative choices of y_i's in $\underline{Y} = (y_1, \cdots, y_i, \cdots, y_N)$ so as to get 9 alternative values for $\theta = \frac{1}{N}\Sigma y_i$ as 0.1 (0.1) 0.9 treating $y_i = 1$ for the i th person having a minimum monthly income C_j, say, with 9 choices of $j = 1, \cdots, 9$ with $y_i = 0$, else. Further, we associate with $\underline{Y}$ a vector $\underline{Z} = (Z_1, \cdots, Z_N)$ of positive numbers as size-measures to be used in drawing a sample with suitable unequal selection-probabilities. For illustration we take $N = 20, n = 7$ which in the case of (I) SRSWR is the number of draws and is the number of distinct units to be selected in employing two other sampling schemes, namely (II) Rao, Hartley and Cochran's (RHC, 1962) scheme and (III) Hartley and Rao's (HR,1962) scheme. We take

$\underline{Z} = (21.9, 20.1, 18.9, 18.3, 17.3, 17.2, 16.5, 16.4, 15.7, 11.6, 9.5, 9.3, 9.2, 9.2,$

8.4,
8.4, 7.6, 7.5, 7.2, 5.8).

Writing $Z = \Sigma z_i, p_i = \frac{z_i}{Z}$, which are the normed size-measures we may briefly describe the schemes (II), (III) as follows: In the RHC scheme the population is divided at random into n groups of sizes N_i each of which is closest to $\frac{N}{n}$ subject to $\Sigma_n N_i = N$, denoting by Σ_n the sum over the n groups. Writing Q_i as the sum of the p_i's of the N_i units in the ith group for the RHC scheme II, we have

$$\begin{aligned} t_b &= \frac{1}{N}\Sigma_n y_i \frac{Q_i}{p_i}, V_p(t_b) = \frac{1}{N^2}[\frac{\Sigma_n N_i^2 - N}{N(N-1)}\Sigma p_i(\frac{y_i}{p_i} - Y)^2], Y = \Sigma y_i \\ v_p(t_b) &= \frac{1}{N^2}(\frac{\Sigma_n N_i^2 - N}{N^2 - \Sigma_n N_i^2})\Sigma_n Q_i(\frac{y_i}{p_i} - t_b)^2; b_{si} = \frac{Q_i}{p_i}, \\ d_{si} &= (\frac{\Sigma_n N_i^2 - N}{N^2 - \Sigma_n N_i^2})(\frac{Q_i}{p_i^2} + (\Sigma_n Q_i)\frac{Q_i^2}{p_i^2} - 2\frac{Q_i^2}{p_i^2}) \\ N^2 V(e_b) &= B\Sigma\frac{V_i}{p_i} + (1-B)\Sigma V_i + B(\Sigma\frac{y_i}{p_i} - Y)^2 \end{aligned}$$

writing $B = \frac{\Sigma_n N_i^2 - N}{N(N-1)}$.

For the SRSWR scheme I, we have

$b_{si} = \frac{N f_{si}}{n}$, writing f_{si} = number of times i occurs in s;

$$d_{si} = \frac{N^2}{n(n-1)}(f_{si} - \frac{f_{si}^2}{n}), V(e_b) = \frac{\theta(1-\theta)}{n} + \frac{N+n-1}{nN^2}\Sigma V_i.$$

In the HR scheme III the units of U are permuted at random and then n units are chosen circular systematically with probabilities proportional to sizes. Further, for this

$$t_b = \tfrac{1}{N}\Sigma\tfrac{y_i}{\pi_i}I_{si}, \pi_i = np_i = \sum_{s \ni i} p(s), \pi_{ij} = \sum_{s \ni i,j} p(s), b_{si} = \frac{1}{\pi_i},$$

$$v_p(t_b) = \Sigma y_i^2 \tfrac{1-\pi_i}{\pi_i}\tfrac{I_{si}}{\pi_i} + \sum_{i \neq j}\sum y_i y_j(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j})\frac{I_{sij}}{\pi_{ij}},$$

$$d_{si} = \tfrac{1-\pi_i}{\pi_i^2}, b_{si}^2 - d_{si} = \tfrac{1}{\pi_i} = b_{si} \text{ implying } v(1) = v(2);$$

$$V(e_b) = \tfrac{1}{N^2}[\Sigma y_i^2\tfrac{1-\pi_i}{\pi_i} + \sum_{i \neq j}\sum y_i y_j\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} + \Sigma\frac{V_i}{\pi_i}].$$

Following Singh et al we consider the criteria for comparison, namely

$$PRE = 100\frac{V(\hat{\theta}_W)}{V(\hat{\theta}_{SJ})} \tag{3.1}$$

- the higher its magnitude the better is $\hat{\theta}_{SJ}$ relative to $\hat{\theta}_W$ and present these values based on each of the three schemes of sampling we employ as above.

Table

Showing the values of PRE for 3 schemes (I,II,III) given from top to bottom

p	0.45	0.6	0.7	0.8	0.9
θ					
0.1	219.85	477.60	226.23	154.48	119.18
	219.48	487.18	233.54	160.65	123.55
	219.23	487.62	234.06	161.21	124.06
0.2	222.40	477.31	224.63	153.06	118.64
	221.42	494.80	237.83	163.72	125.54
	220.83	494.75	238.29	164.39	126.31
0.3	225.21	482.88	226.87	154.40	119.55
	223.87	504.95	243.53	167.64	127.88
	223.06	504.02	234.55	168.07	128.51
0.4	228.33	494.65	232.88	158.14	121.54
	226.93	517.54	250.57	172.00	130.03
	226.09	516.31	250.26	172.39	130.63
0.5	231.76	513.58	243.29	164.66	124.77
	229.78	537.43	263.06	181.35	135.56
	228.70	535.38	262.73	181.91	136.61
0.6	235.54	541.48	259.69	175.18	129.92
	232.97	563.19	280.40	194.63	143.55
	231.71	560.47	280.04	195.75	145.60
0.7	239.73	581.54	285.44	192.69	138.68
	236.32	597.98	306.40	216.72	158.34
	234.84	594.84	306.62	219.88	163.90
0.8	244.37	639.46	327.98	225.04	155.99
	240.14	642.84	343.45	252.07	185.06
	238.56	641.44	347.12	262.89	204.73
0.9	249.50	726.07	407.08	301.04	204.80
	245.23	694.49	386.29	291.64	210.43
	244.08	702.17	401.99	323.65	262.03

Comments on the results in the Table :

(a) Compared to I, the other two schemes yield higher efficiency except when p=0.45;

(b) The schemes II and III are quite competetive with each other but with increasing p the scheme III tends to fare better than II but values of p close to 0.5 are of real consequenses.

(c) Even though they utilize additional data, namely the size-measures for sample selection it is surprising that II and III do not uniformly outperform I.

Remark : The entries in the first rows of the above table corresponding to $p = 0.6(0.1)0.9$ for each θ "equal to 0.1 (0.1) 0.9" match the PRE values given by Singh et al (with a few slight discrepancies possibly because of misprints in Singh et al) calculated by them using the formula

$$PRE = 100\frac{V(\hat{\theta}_W)}{V(\hat{\theta}_{SJ})}$$

as they obviously should.

4.4 Repeated randomized response techniques of Franklin (1989a,b) and of Singh and Singh (1992, 1993):

On taking a simple random sample (SRS) with replacement (WR) and on eliciting repeated "Randomized Responses"(RR) in suitably devised ways from each person sampled, methods have been given by Franklin (1989a, 1989b) and Singh and Singh (1992, 1993) to estimate the proportion of people bearing a sensitive characteristic in a specified community. But in large scale sample surveys a complex design with unequal selection probabilities 'Without replacement' (WOR) is commonly employed yielding sample observations on numerous variables of which only a few may relate to stigmatizing issues. Moreover, sponsorship for a comprehensive exercise to cover sensitive features alone is hard to come by. So, we present 'modified' procedures to extend the above estimation methods to apply not only to SRSWR but also to complex sampling designs. Arnab's (1996,2000) works deal with the extension of Franklin's and Singh and Singh's procedures to cover complex designs but follow a somewhat different line of approach.

4.4.1 Estimation using SRSWR-based repeated RR's

First we discuss below the procedures given by Franklin (1989a, 1989b). A person labeled i, if selected in an SRSWR, taken in n draws, is to report k numbers x_{ij} if he/she bears A or k numbers y_{ij} if he/she bears $\bar{A}$; here for each respondent labeled i, the numbers x_{ij} are 'independently' drawn from a population with pre-assigned means μ_{1j} and variances σ^2_{1j} and y_{ij}'s are 'independently' drawn from populations with means μ_{2j}, and variances $\sigma^2_{2j}, (j = 1, \ldots, k)$ 'independently' across every i in U.

Then, these RR's may be written as z_{ij} such that

$$z_{ij} = I_i x_{ij} + (1 - I_i) y_{ij}, j = 1, \ldots, k; i \in U. \qquad (2.1)'$$

Here we write

$$\begin{aligned} I_i &= 1 \text{ if } i \text{ bears } A \\ &= 0 \text{ if } i \text{ bears } \bar{A}, \text{ the complement of } A. \end{aligned}$$

Of course, $\theta = \frac{1}{N}\Sigma I_i$

Writing expectation, variance and covariance operators generically as E_R, V_R, C_r for the RR's we have

$$E_R(z_{ij}) = \theta\mu_{1j} + (1 - \theta)\mu_{2j}$$

$$\begin{aligned}
V_R(z_{ij}) &= \theta\sigma_{1j}^2 + (1-\theta)\sigma_{2j}^2 + \theta(1-\theta)(\mu_{1j}-\mu_{2j})^2 \\
&\quad j = 1,\ldots,k; i = 1,\ldots,N \\
C_R(z_{ij}, z_{ij'}) &= \theta(1-\theta)(\mu_{1j}-\mu_{2j})(\mu_{1j'}-\mu_{2j'}), \\
&\quad j, j'(j \neq j') = 1,\ldots,k; i = 1,\ldots,N.
\end{aligned}$$

Letting

$$Z_{i0} = \sum_{j=1}^{k} z_{ij}, m_r = \sum_{j=1}^{k} \mu_{rj}, r = 1, 2; m_1 \neq m_2$$

it follows that

$$\begin{aligned}
E_R(Z_{i0}) &= \theta(m_1 - m_2) + m_2 \\
\hat{\theta}_i &= \tfrac{Z_{i0}-m_2}{m_1-m_2} \text{ satisfies } E_R(\hat{\theta}_i) = \theta \text{ and} \\
V_R(\hat{\theta}_i) &= \theta(1-\theta) + \frac{\theta \sum_{j=1}^{k}(\sigma_{1j}^2 - \sigma_{2j}^2) + \sum_{j=1}^{k}\sigma_{2j}^2}{(m_1-m_2)^2}, \text{ and for} \\
\hat{\theta} &= \tfrac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i
\end{aligned} \tag*{(2.2)$'$}$$

one has $E(\hat{\theta}) = \theta$ and on writing $\sigma_r^2 = \sum_{j=1}^{k}\sigma_{rj}^2, r = 1, 2,$

$$V_R(\hat{\theta}) = \frac{V_R(\hat{\theta}_i)}{n} = \frac{\theta(1-\theta)}{n} + \frac{\theta(\sigma_1^2 - \sigma_2^2) + \sigma_2^2}{n(m_1-m_2)^2} \tag*{(2.3)$'$}$$

Further, writing

$$Z_{0j} = \sum_{i=1}^{n} Z_{ij}, \bar{Z}_{0j} = \frac{Z_{0j}}{n} \text{ one has}$$

$$E_R(\bar{Z}_{0j}) = \theta(\mu_{1j} - \mu_{2j}) + \mu_{2j}.$$

Then, choosing $\mu_{1j} \neq \mu_{2j}$ it follows that for

$$\theta_w^* = \sum_{j=1}^{k} W_j(\bar{Z}_{0j} - \mu_{2j})/(\mu_{1j} - \mu_{2j}) \tag*{(2.4)$'$}$$

with assignable weights $W_j (0 < W_j < 1, \sum_{j=1}^{k} W_j = 1)$, one has $E_R(\theta_w^*) = \theta$. Further, observing

$$\begin{aligned}
V_R(\bar{Z}_{0j}) &= \tfrac{1}{n}[\theta(\sigma_{1j}^2 - \sigma_{2j}^2) + \sigma_{2j}^2 + \theta(1-\theta)(\mu_{1j} - \mu_{2j})^2] \text{ and} \\
C_R(\bar{Z}_{0j}, \bar{Z}_{0j'}) &= \tfrac{\theta(1-\theta)}{n}(\mu_{1j} - \mu_{1j'})(\mu_{2j} - \mu_{2j'}), \text{ one has} \\
V_R(\theta_w^*) &= \tfrac{\theta(1-\theta)}{n} + \tfrac{1}{n}\sum_{j=1}^{k} \frac{W_j^2}{(\mu_{1j} - \mu_{2j})^2}[\theta(\sigma_{1j}^2 - \sigma_{2j}^2) + \sigma_{2j}^2]
\end{aligned} \tag{2.5$'$}$$

Hence Franklin (1989a, 1989b) recommends the choice

$$W_j = \frac{|\mu_{1j} - \mu_{2j}|}{\sum_{j=1}^{k} |\mu_{1j} - \mu_{2j}|} = \frac{D_j}{D},$$

say, and gets θ_w^* as, say,

$$\theta_D^* = \frac{1}{D}\Sigma \frac{D_j}{(\mu_{1j} - \mu_{2j})}(\bar{Z}_{0j} - \mu_{2j}) \tag{2.6$'$}$$

for which

$$V_R(\theta_D^*) = \frac{\theta(1-\theta)}{n} + \frac{1}{nD^2}[\theta(\sigma_1^2 - \sigma_2^2) + \sigma_2^2] \tag{2.7$'$}$$

Franklin's (1989a, 1989b) proposed estimators for θ are $\hat{\theta}$ in (2.2)$'$ and θ_D^* in (2.6)$'$ with the

Theorem 1 (Franklin).

$$V_R(\theta_D^*) \leq V_R(\hat{\theta})$$

with an equality only if the signs of all the pairs $(\mu_{1j} - \mu_{2j}), (\mu_{1j'} - \mu_{2j'})$ are the same for $j \neq j'(= 1, \ldots, k)$ implying $(m_1 - m_2)^2 = D^2$.

Singh and Singh (1992) modify Franklin's (1989a, 1989b) procedure. first, to require a sampled person i to choose with a probability $P_r (0 < P_r < 1, r = 1, 2, 3, P_1 + P_2 + P_3 = 1)$ a number x_{rij} from a distribution with mean μ_{1j} and variance σ^2_{r1j} if i bears A but a number y_{rij} from a distribution with mean μ_{2j} and variance σ^2_{r2j} if i bears $\bar{A}$. For each i, the x_{rij}'s are 'independent' for $j = 1, 2, \ldots, k$ and $r = 1, 2, 3$ and similarly for y_{rij}'s. Also, across i in U, x_{rij}'s are 'independent' and so are y_{rij}'s. For the observable RR's denoted by L_{ij} for $j = 1, \ldots, k$ and $i = 1, \ldots, N$,

$$E_R(L_{ij}) = \theta(\mu_{1j} - \mu_{2j}) + \mu_{2j} \qquad (2.8)'$$

$$\begin{aligned} V_R(L_{ij}) &= \theta\left(\sum_{r=1}^{3} P_r(\sigma^2_{r1j} - \sigma^2_{r2j}) + \sum_{r=1}^{3} P_r\sigma^2_{r2j}\right) \\ &\quad +\theta(1-\theta)(\mu_{1j} - \mu_{2j})^2 \end{aligned} \qquad (2.9)'$$

$$C_R(L_{ij}, L_{ij'}) = \theta(1-\theta)(\mu_{1j} - \mu_{2j})(\mu_{1j'} - \mu_{2j'}), \qquad (2.10)'$$

$i \in U; j = 1, \ldots, k;\ j, j'(j \neq j') = 1, \ldots, k.$

Writing $\bar{L}_{i0} = \frac{1}{k}\sum_{j=1}^{k} L_{ij}$, noting $E_R(\bar{L}_{i0}) = \frac{1}{k}[\theta(m_1 - m_2) + m_2]$, Singh and Singh (1992) propose the estimator for θ as

$$\tilde{\theta} = \frac{1}{n(m_1 - m_2)}[k\sum_{1}^{n} \bar{L}_{i0} - m_2] \text{ for which} \qquad (2.11)'$$

$$E_R(\tilde{\theta}) = \theta \text{ and}$$

$$V_R(\tilde{\theta}) = \frac{\theta(1-\theta)}{n} + \frac{[\theta\sum_{r=1}^{3} P_r(\sum_{j=1}^{k}(\sigma^2_{r1j} - \sigma^2_{r2j})) + \sum_{r=1}^{3} P_r\sum_{j=1}^{k}\sigma^2_{r2j})]}{n(m_1 - m_2)^2} \qquad (2.12)'$$

Also, noting that

$$\tilde{\theta}_j = \frac{1}{(\mu_{1j} - \mu_{2j})}[\frac{1}{n}\sum_{i=1}^{n} L_{ij} - \mu_{2j}]$$

satisfies $E_R(\tilde{\theta}_j) = \theta$, they propose another estimator for θ as $\tilde{\theta}_w = \sum_{j=1}^{k} W_j\tilde{\theta}_j$ with W_j's as weights ($0 < W_j < 1$, $\sum_{j=1}^{k} W_j = 1$) so that $E_R(\tilde{\theta}_w) = \theta$ and

$$V_R(\tilde{\theta}_w) = \tfrac{\theta(1-\theta)}{n} + \tfrac{1}{n}\sum_1^k W_j^2 \frac{[\theta \sum_r P_r(\sigma^2_{r1j} - \sigma^2_{r2j}) + \sum_r P_r(\sigma^2_{r2j})]}{(\mu_{1j} - \mu_{2j})^2} \qquad (2.13)'$$

They further recommend the choice $W_j = \frac{|\mu_{1j}-\mu_{2j}|}{\sum_j |\mu_{1j} - \mu_{2j}|} = \frac{D_j}{D}$, say, so as to propose finally $\tilde{\theta}_\mu = \frac{1}{D}\,\Sigma\, D_j\ \tilde{\theta}_j$ as their estimator for θ and observe that

$$V_R(\tilde{\theta}_\mu) = \tfrac{\theta(1-\theta)}{n} + \tfrac{1}{nD^2}[\theta \sum_r P_r \sum_j(\sigma^2_{r1j} - \sigma^2_{r2j}) + \sum_r P_r \sum_j \sigma^2_{r2j}] \qquad (2.14)'$$

They have then the

Theorem 2 (Singh and Singh).

$$V_R(\tilde{\theta}_\mu) \leq V_R(\tilde{\theta})$$

with an equality only if $(\mu_{1j} - \mu_{2j})$ and $(\mu_{1j'} - \mu_{2j'})\forall j \neq j'$ have a common sign in which case $(m_1 - m_2)^2 = D^2$.

Singh and Singh's (1993) subsequent work is only a simple modification of their previous work which only allows r to take on only 2 values, with $P_1 = T$ and $P_2 = 1 - T$ keeping everything else in tact. Denoting $\tilde{\theta}$ by $\tilde{\theta}_T$ and $\tilde{\theta}_\mu$ by $\tilde{\theta}_{\mu T}$ to cover this case they have

$$\begin{aligned} V(\tilde{\theta}_T) \quad &= \quad \tfrac{\theta(1-\theta)}{n} + \frac{(\theta \sum_j \sigma^2_{21j} + (1-\theta)\sum_j \sigma^2_{22j})}{n(m_1-m_2)^2} \\ &\quad + \tfrac{T}{n(m_1-m_2)^2}[\theta(\sum_j \sigma^2_{11j} - \sum_j \sigma^2_{21j}) + (1-\theta)\sum_j(\sigma^2_{12j} - \sigma^2_{22j})] \end{aligned}$$

and

$$\begin{aligned} V(\tilde{\theta}_{\mu T}) \quad &= \quad \tfrac{\theta(1-\theta)}{n} + \tfrac{1}{nD^2}[(\theta \sum_j \sigma^2_{21j} + (1-\theta)\sum_j \sigma^2_{22j}] \\ &\quad + \tfrac{T}{nD^2}[\theta(\sum_j \sigma^2_{11j} - \sigma^2_{21j}) + (1-\theta)\sum_j(\sigma^2_{12j} - \sigma^2_{22j})] \end{aligned} \qquad (2.15)'$$

They also have a corresponding

Theorem 3. (Singh and Singh).

$$V(\widetilde{\theta}_{\mu T}) \leq V(\widetilde{\theta}_{T})$$

with equality only if $(\mu_{1j} - \mu_{2j})$ and $(\mu_{1j'} - \mu_{2j'}) \forall j \neq j'$ have a common sign rendering $(m_1 - m_2)^2 = D^2$.

4.4.2 Estimation using Repeated RR's in Complex Surveys

Let s be a sample drawn from U according to any sampling design P with a probability $p(s)$. We shall write E_p, V_p, C_p to denote operators for expectation, variance, covariance in respect of P. Let b_{si} be freely assignable constants not involving the elements of $\underline{I} = (I_1, \ldots, I_i, \ldots, I_N)$ but subject to

$$E_p(b_{si} I_{si}) = 1 \text{ for every } i \text{ in } U. \qquad (3.1)'$$

Here

$$\begin{aligned} I_{si} &= 1 \text{ if } i \in s \\ &= 0 \text{ if } i \notin s. \end{aligned}$$

Later we shall write $I_{sij} = I_{si} I_{sj}, i, j \in U$. If I_i's were ascertainable for i in s we could employ

$$t = \sum_{i=1}^{N} I_i b_{si} I_{si} \qquad (3.2)$$

as an unbiased estimator for I because

$$E_p(t) = \sum_{1}^{N} I_i = I$$

by virtue of (3.1)$'$.

Since DR's are not available on I_i's for i in s we need suitable estimators $\hat{I}_i$ for I_i for $i \in s$ to be substituted into t in (3.2) to derive estimators

$$e = \sum_{i=1}^{N} \hat{I}_i b_{si} I_{si} \tag{3.3}$$

for I. We shall write $E = E_p E_R = E_R E_p, C = C_p(E_r(.), E_r(.)) + E_p C_R(.,.)$ and $V = E_p V_R + V_p E_R = E_R V_p + V_R E_p$ to denote the over-all expectation, covariance, variance operators covering both sampling design and RR-based generation of data.

Then, e is an unbiased estimator for I because

$$E_p(e) = \Sigma \hat{I}_i \text{ and so } E(e) = E_R(\Sigma \hat{I}_i) = \Sigma I_i = I$$

and also

$$E_R(e) = t \text{ and } E(e) = E_p(t) = \Sigma I_i = I.$$

Noting $I_i^2 = I_i$ and writing

$$V_p(t) = \sum_i I_i d_i + \sum\sum_{i \neq j} I_i I_j d_{ij},$$

where

$$d_i = E_p(b_{si}^2 I_{si}) - 1, d_{ij} = E_p(b_{si} b_{sj} I_{sij}) - 1,$$

let it be possible to find constants C_{si}, C_{sij}, both free of $\underline{I}$ satisfying the conditions

$$E_p(C_{si} I_{si}) = d_i, E_p(C_{sij} I_{sij}) = d_{ij}.$$

Then

$$v_p(t) = \sum_i I_i C_{si} I_{si} + \sum\sum_{i \neq j} I_i I_j C_{sij} I_{sij}$$

satisfies $E_p v_p(t) = V_p(t)$. We may also note that

$$V_p(e) = \Sigma \hat{I}_i^2 d_i + \underset{i \neq j}{\Sigma\Sigma} \hat{I}_i \hat{I}_j d_{ij}$$

since $\hat{I}_i^2$ may not equal $\hat{I}_i$.

The literature on Survey sampling, in particular the texts by Cochran (1977), Chaudhuri and Stenger (1992), Särndal, Swensson and Wretman (1992), among many others, give numerous accounts of choices concerning P, b_{si}, C_{si} and C_{sij}'s . A choice of b_{si} equal to $\frac{1}{\pi_i}$, where $\pi_i = \underset{s \ni i}{\Sigma} p(s)$, is very common leading to $C_{si} = \frac{1-\pi_i}{\pi_i^2}$, $C_{sij} = \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}$, where $\pi_{ij} = \underset{s \ni i,j}{\Sigma} p(s)$ - it is assumed that $\pi_i > 0 \forall i, \pi_{ij} > 0 \forall i, j (i \neq j)$, a design P ensuring this is of course possible. This choice of b_{si}, C_{si}, C_{sij}'s is due to Horvitz and Thompson (HT, 1952).

If it is possible to find $\hat{I}_i, i \in U$, such that

$$(i) E_R(\hat{I}_i) = I_i \quad (ii) V_R(\hat{I}_i) = V_i, \text{ say},$$

admitting v_i such that $E_R(v_i) = V_i$ and (iii) $\hat{I}_i$'s are 'independently' distributed, then one may easily check the following:

Let

$$v_p(e) = \Sigma \hat{I}_i^2 C_{si} I_{si} + \underset{i \neq j}{\Sigma\Sigma} \hat{I}_i \hat{I}_j C_{sij} I_{sij}$$

writing $\underset{i \neq j}{\Sigma\Sigma}$ for sum over $i, j (i \neq j)$ in U.

Then, for

$$v_1(e) = v_p(e) + \Sigma v_i b_{si} I_{si} \text{ and} \tag{3.4}$$

$$v_2(e) = v_p(e) + \Sigma v_i (b_{si}^2 - C_{si}) I_{si} \tag{3.5}$$

one has

$$\begin{aligned}
(iv) Ev_1(e) &= E_R E_p v_1(e) \\
&= E_R[\Sigma \hat{I}_i^2 d_i + \underset{i \neq j}{\Sigma\Sigma} \hat{I}_i \hat{I}_j d_{ij}] + E_R \Sigma v_i \\
&= \Sigma I_i d_i + \underset{i \neq j}{\Sigma\Sigma} I_i I_j d_{ij} + \Sigma V_i d_i + \Sigma V_i, \\
\text{noting } E_R(\hat{I}_i^2) &= V_i + I_i; \\
\text{so, } Ev_1(e) &= V_p(t) + \Sigma V_i(1 + d_i); \\
\text{and } (v) Ev_2(e) &= E_p E_R v_2(e) \\
&= E_p[\Sigma I_i C_{si} I_{si} + \underset{i \neq j}{\Sigma\Sigma} I_i I_j C_{sij} I_{sij}] \\
&\quad + E_p(\Sigma V_i C_{si} I_{si}) + E_p \Sigma V_i (b_{si}^2 - C_{si}) I_{si} \\
&= \Sigma I_i d_i + \underset{i \neq j}{\Sigma\Sigma} I_i I_j d_{ij} + \Sigma V_i d_i + \Sigma V_i(1 + d_i - d_i) \\
&= \Sigma I_i d_i + \underset{i \neq j}{\Sigma\Sigma} I_i I_j d_{ij} + \Sigma V_i(1 + d_i) \\
&= V_p(t) + \Sigma V_i(1 + d_i)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
V(e) &= E_p V_R(e) + V_p E_R(e) = E_p[\Sigma V_i b_{si}^2 I_{si}] + V_p(t) \\
&= V_p(t) + \Sigma V_i(1 + d_i) \\
\text{Also, } V(e) &= E_R V_p(e) + V_R E_p(e) \qquad (3.6) \\
&= \sum_i I_i d_i + \underset{i \neq j}{\Sigma\Sigma} I_i I_j d_{ij} + \Sigma V_i d_i + V_R(\Sigma \hat{I}_i) \\
&= V_p(t) + \Sigma V_i(1 + d_i)
\end{aligned}$$

So, $v_1(e)$ and $v_2(e)$ may be taken as unbiased estimators for $V(e)$. Further, noting that for

$$w_p = v_p(t)\Big|_{I_i = \hat{I}_i} = \Sigma \hat{I}_i C_{si} I_{si} + \underset{i \neq j}{\Sigma\Sigma} \hat{I}_i \hat{I}_j c_{sij} I_{sij}$$

we have

$E(w_p) = E_p E_R(w_p) = E_p v_p(t) = V_p(t)$, and so for $v_3(e) = w_p + \Sigma v_i b_{si}^2 I_{si}$ we have $Ev_3(e) = V_p(t) + \Sigma V_i(1 + d_i)$.

So, $v_3(e)$ is a third unbiased estimator of $V(e)$.

With these preliminaries let us extend the methods of Franklin (1989a, 1989b) and of Singh and Singh (1992, 1993) to general sampling designs. Chaudhuri (1999, 2000b) examined certain aspects of the approach we are going to present here. An interested reader may consult Chaudhuri and Mukerjee's (1992) text.

We shall now calculate RR-based expectation, variance, covariance "conditionally throughout on $I = (I_1, \ldots, I_i, \ldots, I_N)$ as given" but continue to use the same notations E_R, V_R, C_R. Then, for the RR's z_{ij} based on Franklin's (1989a, 1989b) scheme we have

$$\begin{aligned} E_R(z_{ij}) &= I_i\mu_{1j} + (1 - I_i)\mu_{2j}, \\ V_R(z_{ij}) &= I_i\sigma_{1j}^2 + (1 - I_i)\sigma_{2j}^2, \\ C_R(z_{ij}, z_{ij'}) &= 0 \quad \forall j \neq j', i \in U. \end{aligned}$$

Then, for $\hat{I}_i = \frac{z_{i0} - m_2}{(m_1 - m_2)}$, we have $E_R(\hat{I}_i) = I_i$ and for

$$\begin{aligned} e &= \Sigma\hat{I}_i b_{si} I_{si} \text{ we have} \\ V(e) &= V_p(t) + \frac{1}{(m_1 - m_2)^2}[\sum_{j=1}^{k}(\sigma_{1j}^2 - \sigma_{2j}^2)E_p\Sigma I_i b_{si}^2 I_{si} \\ &\quad +(\sum_{j=1}^{k} \sigma_{2j}^2)E_p(\Sigma b_{si}^2 I_{si})] \end{aligned}$$

$$= V_p(t) + \frac{1}{(m_1 - m_2)^2}[(\sigma_1^2 - \sigma_2^2)\Sigma I_i(1 + d_i) + \sigma_2^2\Sigma(1 + d_i)] \tag{3.7}$$

$$\text{Then, } v_3(e) = \sum_i \hat{I}_i C_{si} I_{si} + \sum\sum_{i \neq j} \hat{I}_i\hat{I}_j C_{sij} I_{sij}$$

$$+\frac{1}{(m_1 - m_2)^2}[(\sigma_1^2 - \sigma_2^2)\Sigma\hat{I}_i b_{si}^2 I_{si} + \sigma_2^2\Sigma b_{si}^2 I_{si}] \tag{3.8}$$

is an unbiased estimator for $V(e)$. Similarly, $v_1(e), v_2(e)$ may also be written down explicitly. Let

$$Z_{bj} = \Sigma z_{ij} b_{si} I_{si};\ \delta_j = \mu_{1j} - \mu_{2j};\ \triangle_j = \sigma^2_{1j} - \sigma^2_{2j}.$$

Then,

$$E_R(Z_{bj}) = \delta_j \Sigma I_i b_{si} I_{si} + \mu_{2j} \Sigma b_{si} I_{si} \tag{3.9}$$

$$V_R(Z_{ij}) = \triangle_j I_i + \sigma^2_{2j}.$$

Let

$$e_j = \frac{1}{\delta_j}(Z_{bj} - \mu_{2j} \Sigma b_{si} I_{si}).$$

Then,

$$\begin{aligned} E_R(e_j) &= \Sigma I_i b_{si} I_{si} = t; \\ V_R(e_j) &= \frac{1}{\delta_j^2}[\triangle_j \Sigma I_i b^2_{si} I_{si} + \sigma^2_{2j} \Sigma b^2_{si} I_{si}] \\ V(e_j) &= E_p V_R(e_j) + V_p(E_R(e_j)) \\ &= V_p(t) + \frac{1}{\delta_j^2}[\triangle_j \Sigma I_i (1 + d_i) + \sigma^2_{2j} \Sigma (1 + d_i)] \end{aligned}$$

Let W_j's be assignable weights such that $0 < W_j < 1$, $\sum\limits_{j=1}^{k} W_j = 1$ and $e^*_w = \sum\limits_{j=1}^{k} W_j e_j$. Then,

$$V(e^*_w) = \Sigma W_j^2 V(e_j) + \mathop{\Sigma\Sigma}_{j \neq j'} W_j W_{j'} C(e_j, e_{j'})$$

writing $\mathop{\Sigma\Sigma}_{j \neq j'}$ for sum over $j, j'(j \neq j') = (1, \ldots, k)$.

Now

$$\begin{aligned} C(e_j, e_{j'}) &= E_p[C_R(e_j, e_{j'})] + C_p[E_R(e_j), E_R(e_{j'})] \\ &= C_p(t,t) = V_p(t) \end{aligned}$$

because z_{ij}'s being independent across $j = 1, \ldots, k, C_R(e_j, e_{j'}) = 0 \forall j \neq j'$.

So,

$$\begin{aligned} V(e^*_w) &= V_p(t)(1 - \sum_{j=1}^{k} W_j^2) + V_p(t) \sum_{j=1}^{k} W_j^2 \\ &+ \sum_{j=1}^{k} \frac{W_j^2}{\delta_j^2}[\triangle_j \sum_i I_i(1+d_i) + \sigma_{2j}^2 \sum_i (1+d_i)] \\ &= V_p(t) + (\sum_i I_i(1+d_i))[\sum_{j=1}^{k} \frac{W_j^2 \triangle_j}{\delta_j^2}] \\ &+ \sum_i (1+d_i) \left(\sum_{j=1}^{k} \frac{W_j^2 \sigma_{2j}^2}{\delta_j^2} \right). \end{aligned}$$

Then,

$$\begin{aligned} v_3(e^*_w) &= \sum_i \hat{I}_i C_{si} I_{si} + \sum\sum_{i \neq j} \hat{I}_i \hat{I}_j C_{sij} I_{sij} \\ &+ (\sum_i \hat{I}_i b_{si}^2 I_{si}) \sum_{j=1}^{k} \frac{W_j^2 \triangle_j}{\delta_j^2} \\ &+ (\sum_i b_{si}^2 I_{si}) \left(\sum_{j=1}^{k} \frac{W_j^2 \sigma_{2j}^2}{\delta_j^2} \right) \end{aligned}$$

is an unbiased estimator for $V(e^*)$. Similarly, $v_1(e^*_w), v_2(e^*_w)$ may also be written down. But we omit them.

Remark 1. To facilitate a comparison between e and e^*_w we need to have

$$\sigma_{1j}^2 = \sigma_{2j}^2 = \sigma_j^2 \text{ for every } j = 1, \ldots, k. \qquad (3.10)$$

For simplicity we shall write $\sigma^2 = \sum_{j=1}^{k} \sigma_j^2$.

Then,

$$V(e) = V_p(t) + \frac{\sigma^2}{(m_1 - m_2)^2} \sum_i (1 + d_i).$$

$$\text{and } V(e_w^*) = V_p(t) + \left(\sum_{j=1}^{k} W_j^2 \frac{\sigma_j^2}{\delta_j^2} \right) \sum_i (1 + d_i) \tag{3.11}$$

Then, we have the

Lemma 1. $V(e_w^*)$ is minimized for the choice

$$W_j = \frac{\delta_j^2}{\sigma_j^2}, j = 1, \ldots, k. \tag{3.12}$$

Proof: $V(e_w^*)$ is minimized if W_j's subject to $\sum_{j=1}^{k} W_j = 1$ are chosen to minimize $\sum_{j=1}^{k} W_j^2 \frac{\sigma_j^2}{\delta_j^2}$.

This is achieved on solving

$$0 = \frac{\partial}{\partial W_j} \left[\sum_j W_j^2 \frac{\sigma_j^2}{\delta_j^2} + \lambda (\sum_i W_j - 1) \right],$$

with λ as the Lagrangian undetermined multiplier, leading to (3.12). The optimal e_w^* will be written as e^* and it follows that

$$V(e^*) = V_p(t) + (\sum_i (1 + d_i)) \frac{1}{\sum_{j=1}^{k} \frac{\delta_j^2}{\sigma_j^2}} \tag{3.13}$$

Then, we have the

Theorem A. $V(e) \geq V(e^*)$.

Proof.

$$\frac{\sigma^2}{(m_1-m_2)^2}-\frac{1}{\sum\limits_{j=1}^{k}\frac{\delta_j^2}{\sigma_j^2}}=\frac{\sum\limits_{1}^{k}\sigma_j^2}{(\sum\limits_{j=1}^{k}\delta_j)^2}-\frac{1}{(\sum\limits_{j=1}^{k}\frac{\delta_j^2}{\sigma_j^2})}\geq 0$$

using the Cauchy inequality.

Remark 2. Franklin's (1989a, 1989b) choice

$$W_j=\frac{|\mu_{1j}-\mu_{2j}|}{\sum\limits_{j=1}^{k}|\mu_{1j}-\mu_{2j}|} \text{ leads to} \tag{3.14}$$

$$\Sigma W_j^2\frac{\sigma_j^2}{\delta_j^2}=\frac{\sum\limits_{i}\sigma_j^2}{\sum\limits_{j}|\mu_{1j}-\mu_{2j}|]^2}.$$

Writing e_F^* for e_w^* with this choice of W_j as in (3.14) we have the

Theorem B.

$$V(e)\geq V(e_F^*) \tag{3.15}$$

Proof.

$$(m_1-m_2)^2-[\sum_j|\mu_{1j}-\mu_{2j}|]^2$$

$$=\sum_j\sum_{j'}(\mu_{1j}-\mu_{2j})(\mu_{1j'}-\mu_{2j'})-\sum_j\sum_{j'}|\mu_{1j}-\mu_{2j}|\,|\mu_{1j'}-\mu_{2j'}|\leq 0.$$

Hence (3.15) follows.

For the Singh and Singh's (1992) scheme, corresponding to (2.8)′ - (2.12)′ we respectively have in the present case:

$$E_R(L_{ij}) = I_i(\mu_{1j} - \mu_{2j}) + \mu_{2j} \tag{3.16}$$

$$V_R(L_{ij}) = I_i \sum_{r=1}^{3} P_r(\sigma^2_{r1j} - \sigma^2_{r2j}) + \sum_{r=1}^{3} P_r\sigma^2_{r2j}$$

$$C_R(L_{ij}, L_{ij'}) = 0 \forall\ j \neq j'$$

$$\widetilde{e} = \Sigma \hat{I}_i b_{si} I_{si}$$

where

$$\hat{I}_i = \frac{L_{i0} - m_2}{m_1 - m_2}, \text{ giving } E_r(\widetilde{e}) = t$$

$$\begin{aligned} V(\widetilde{e}) &= V_p(t) + \frac{1}{(m_1 - m_2)^2}[(\sum_r P_r \sum_{r=1}^{3}(\sigma^2_{r1j} - \sigma^2_{r2j})) \\ &\quad \Sigma I_i(1 + d_i) + (\sum_r P_r \sum_{r=1}^{3} \sigma^2_{r2j})\Sigma(1 + d_i)] \end{aligned}$$

for which an unbiased estimator is

$$\begin{aligned} v_3(\widetilde{e}) &= \Sigma \hat{I}_i C_{si} I_{si} + \sum\sum_{i \neq j} \hat{I}_i \hat{I}_j C_{sij} I_{sij} \\ &\quad + \frac{1}{(m_1 - m_2)^2}[(\sum_r P_r \sum_{j=1}^{k}(\sigma^2_{r1j} - \sigma^2_{r2j})) \\ &\quad \Sigma \hat{I}_i b^2_{si} I_{si} + (\sum_r P_r \sum_{j=1}^{k} \sigma^2_{r2j})\Sigma b^2_{si} I_{si}] \end{aligned}$$

Similarly, $v_1(\widetilde{e}), v_2(\widetilde{e})$ may be written down but we omit. Next, let

$$L_{bj} = \Sigma L_{ij} b_{si} I_{si}, \psi_j = \sum_r P_r(\sigma^2_{r1j} - \sigma^2_{r2j}), \phi_j = \sum_r \sigma^2_{r2j}$$

Then,

$$\begin{aligned} E_R(L_{bj}) &= \delta_j \Sigma I_i b_{si} I_{si} + \mu_{2j} \Sigma b_{si} I_{si} \\ \widetilde{e}_j &= \frac{1}{\delta_j}(L_{bj} - \mu_{2j} \Sigma b_{si} I_{si}) \text{ giving } E_R(\widetilde{e}) = t, \\ V_R(\widetilde{e}_j) &= \frac{1}{\delta_j^2}[\psi_j \Sigma I_i b_{si}^2 I_{si} + \phi_j \Sigma b_{si}^2 I_{si}] \\ V(\widetilde{e}_j) &= V_p(t) + \frac{1}{\delta_j^2}[\psi_j \Sigma I_i (1 + d_i) + \phi_j \Sigma (1 + d_i)] \end{aligned}$$

With W_j's, as before, as assignable constants, let $\widetilde{e}_w = \sum_{j=1}^{k} W_j \widetilde{e}_j$ be a convex linear combination of $\widetilde{e}_j$'s. Since L_{ij} and $L_{ij'}$ are 'uncorrelated' for $j \neq j'$,

$$V(\widetilde{e}_w) = \Sigma W_j^2 V(\widetilde{e}_j) + (\sum\sum_{j \neq j'} W_j W_{j'}) V_p(t)$$

because

$$C(\widetilde{e}_j, \widetilde{e}'_j) = C_p(E_R(\widetilde{e}_j), E_R(\widetilde{e}_{j\prime})) = V_p(t).$$

So,

$$V(\widetilde{e}_w) = V_p(t) + \sum_j \frac{W_j^2}{\phi_j^2}[\phi_j \Sigma I_i (1 + d_i) + \phi_j \Sigma (1 + d_i)]$$

and

$$\begin{aligned} v_3(\widetilde{e}_w) &= \Sigma \hat{I}_i C_{si} I_{si} + \sum\sum_{i \neq j} \hat{I}_i \hat{I}_j C_{si} I_{sij} \\ &\quad + \sum_j \frac{W_j^2}{\delta_j^2}[\psi_j \Sigma \hat{I}_i b_{si}^2 + \phi_j \Sigma b_{si}^2 I_{si}] \end{aligned}$$

is an unbiased estimator for $V(\widetilde{e}_w)$. Formulae for $v_1(\widetilde{e}_w), v_2(\widetilde{e}_w)$ are similar but omitted.

Remark 3. To compare $\tilde{e}$ with $\tilde{e}_w$ let us incorporate the simplifying assumption that

$$\sigma^2_{r1j} = \sigma^2_{r2j} = \sigma^2_j$$

for every $r = 1, 2, 3$ and $j (= 1, \ldots, k)$ and write $\Sigma\sigma^2_j = \sigma^2$.

Then,

$$V(\tilde{e}) = V_p(t) + \frac{\sigma^2}{(m_1 - m_2)^2} \sum_i (1 + d_i)$$

and

$$V(\tilde{e}_w) = V_p(t) + \left(\sum_{j=1}^{k} W_j^2 \frac{\sigma_j^2}{\delta_j^2} \right) \Sigma(1 + d_i)$$

Since $V(\tilde{e})$ equals $V(e)$ and $V(\tilde{e}_w)$ equals $V(e^*_w)$, as in (3.10) and (3.11) respectively, the results concerning "e Vs e^*_w" apply to "$\tilde{e}$ Vs $\tilde{e}_w$" under the assumptions in Remark 3.

Since Singh and Singh's (1993) work is a special case of Singh and Singh's (1992) work, extension of the former to cover the general sampling designs is straightforward and hence we omit the details.

Remark 4. Incidentally, Franklin (1989a, 1989b) and Singh and Singh (1992, 1993) have not presented estimators for the variances of their estimators.

Remark 5. Franklin (1989a, 1989b) gave as follows also a maximum likelihood estimator (MLE) for θ and not only the estimators by the 'Method of Moments' (MM) as discussed in Section 4.2. Taking f_{ij} as the 'probability density function' (pdf) of x_{ij} or the 'probability mass function' (pmf) in the discrete case and g_{ij} as that of y_{ij} the likelihood of θ given the

$$RR \text{ as } \underset{\sim}{z} = (z_{11}, \ldots, z_{1k}, \ldots, z_{i1}, \ldots, z_{ik}, \ldots, z_{n1}, , \ldots, z_{nk}) \text{ is}$$

$$\begin{aligned} L(\theta|\underset{\sim}{z}) &= \underset{i=1}{\overset{n}{\pi}} [\theta \underset{j=1}{\overset{k}{\pi}} f_{ij} + (1-\theta) \underset{j=1}{\overset{k}{\pi}} g_{ij}] \\ &= \sum_{i=1}^{n} [\theta(\gamma_i - \eta_i) + \eta_i], \text{ say, the MLE of } \theta \end{aligned}$$

is $\hat{\theta}_M$ which is the solution of the equation

$$0 = \frac{\partial \log L(\theta|\underset{\sim}{z})}{\partial\theta} = \sum_{i=1}^{n} \left[\frac{(\gamma_i - \eta_i)}{\theta(\gamma_i - \eta_i) + \eta_i}\right]$$

as obtainable by the 'grid search' method. Further properties of $\hat{\theta}_M$ are not reported. But a major shortcoming is that its value may often go beyond the possible values of θ which are only $\frac{i}{N}, i = 0, 1, \ldots, N-1, N$.

With our formulation we may write down the 'Likelihood' of I_i given $\underset{\sim}{z}_{ij} = (z_{i1}, \ldots, z_{ij}, \ldots, \ldots z_{ik})$ as

$$L(I_i(\underset{\sim}{z}_{ij})) = \left(\underset{j=1}{\overset{k}{\pi}} f_{ij}\right)^{I_i} \left(\underset{j=1}{\overset{k}{\pi}} g_{ij}\right)^{1-I_i} = \gamma_i^{I_i} \eta_i^{1-I_i}.$$

Since $I_i = 1$ or 0, the parametric space is composed of only two elements, namely 1 and 0. So, the MLE of I_i is, say, m_i given by

$$\begin{aligned} m_i &= 1 \text{ if } \gamma_i > \eta_i \\ &= 0 \text{ if } \gamma_i < \eta_i; \end{aligned}$$

if $\gamma_i = \eta_i$, no MLE of m_i exists.

Using this m_i as an estimator for I_i we may proceed to estimate I by

$$e_m = \Sigma m_i b_{si} I_{si} \tag{3.17}$$

For Singh and Singh's (1992, 1993) schemes, in our formulation, the 'Likelihoods' of I_i given the RR as L_{ij} and R_{ij}, say, in these two respective cases, are

$$\begin{aligned} L(I_i|\underline{L}_{ij}) &= \left((\sum_{j=1}^{k} P_r \mathop{\pi}_{j=1}^{k} f_{rij})^{I_i} \right) \left(\sum_{r=1}^{3} P_r \left(\mathop{\pi}_{j=1}^{k} g_{rij} \right) \right)^{1-I_i} \\ &= \alpha_i^{I_i} \beta_i^{1-I_i} \text{ say} \\ \text{and } L(I_i|\underline{R}_{ij}) &= (T \mathop{\pi}_{j=1}^{k} f_{1ij})^{I_i} ((1-T) \mathop{\pi}_{j=1}^{k} f_{2ij})^{1-I_i} \end{aligned}$$

The MLE's of I_i in both cases are immediately derived. But subsequent investigation of the properties of the estimators of I of the form (3.17) is not easy and is not taken up here.

Chaudhuri (2001) presented estimators by the maximum likelihood approach to cover RR models given by Warner (1965), Kuk (1990), Mangat (1992), Mangat and Singh (1990) and Mangat, Singh and Singh (1992).But we do not persue further with this approach here.

4.5 Conclusion and recommendation

Repeated RR's elicited from each sampled person allow alternative estimators for proportions of people bearing sensitive characteristics and an appropriate choice among them is possible when samples are suitably drawn not necessarily with equal probabilities and with replacement. Unbiased estimation of the variances of the estimators is also easy to implement. Our recommendation is that for complex large scale surveys a few sensitive items should be covered and inference concerning them may be implemented employing some of the procedures presented here.

Chapter 5

Bootstrap Procedures for Generalized Regression Estimators

Abstract

From the works of Särndal (1996), Deville (1999), Brewer (1999, 2000) and Brewer and Gregoire (2000) among others we gather that "Variance and Mean Square Error" - estimation for Horvitz and Thompson's (HT, 1952) and generalized regression (greg) estimators for a finite population total needs to be simplified preferably by omitting the cross-product terms that involve 'second order inclusion-probabilities' which are often hard to compute. We present two 'Bootstrap' sampling procedures as a simpler alternative to cover situations left beyond the applicability of Rao and Wu's (1988) bootstrap method for the greg estimator. Through simulated numerical exercises we illustrate how the procedure may work vis-a-vis the available traditional estimation procedures concerning the greg estimator. We illustrate three sampling schemes, each with variable 'effective sample sizes'.

5.1 Introduction

We consider estimating the total Y of a real-valued variable y defined on a finite survey population. We suppose that positive values of a related variable z are available for all the units of the population for their utilization as the size-measures to assist the drawing of a suitable sample with unequal selection probabilities. We also suppose that for a sample the values of another correlated positive-valued variable x with a known population total X may be ascertained along with those on the first variable of interest. In such a situation an appropriate estimator of the total is known to be the generalized regression (greg) estimator introduced by Cassel, Särndal and Wretman (CWS, 1976) motivated by a postulated line of regression of y on x through the origin as an improvement upon the classical Horvitz and Thompson's (HT, 1952) estimator HTE. In large -scale surveys utilization of both is of late being considered jeopardized because (1) one has to accurately evaluate a large number of cross-product terms with widely variable coefficients in the formulae for the estimators of the variances of HTE and of the Mean Square Errors (MSE) for the greg estimator and (2) in both cases computation of the Second order inclusion-probabilities of the pairs of units becomes discouragingly hard for many sampling schemes. To tackle these problems several prescriptions are emerging especially through the recent works of Särndal (1996), Deville (1999), Brewer (1999, 2000) and Brewer and Gregoire (2000) among others. Without detailing them we may only suggest here that the use of the 'Bootstrap' technique may be convenient to produce simple variance estimators for the greg estimator avoiding one of the two difficulties (1) and (2) noted above. This may be accomplished using Rao and Wu's (1988) bootstrap technique, provided (A) every sample s with a positive selection probability $p(s)$ has the number of distinct units $\nu(s)$ in it which is constant across the samples and (B) the design P employed with the probabilities $p(s)$ is such that "$\pi_i \pi_j \geq \pi_{ij} \forall i, j$ in the population $U = (1, \cdots, N)$ of the units, $i \neq j$", writing $\pi_i = \sum\limits_s p(s) I_{si}, \pi_{ij} = \sum\limits_s p(s) I_{sij}$, assumed throughout positive

$\forall i \neq j = 1, \cdots, N$ and $i = 1, \cdots, N$. Here $\sum\limits_{s}$ is sum over all the samples, $I_{si} = 1$ if $i \epsilon s, 0$ otherwise and $I_{sij} = I_{si} I_{sj}$.

We shall first present 'methods of drawing the bootstrap samples' needed to evaluate estimates of MSE of the greg estimator when (I) (A) is violated but not (B) and (II) both (A) and (B) are violated so that Rao and Wu's (1988) method does not apply.

Next we illustrate 3 sampling schemes for which (I) holds, generally, or at least for certain 'real life' data. Next we present simulated exercises to show 'How the bootstrap procedures' proposed may fare vis-a-vis the traditional procedures of using the MSE-estimator $m(t_g)$ for the greg estimator t_g. For this we apply the usual criteria of

(i) ACP, the actual coverage Percentage for a 95 percent confidence interval (CI), namely $(t_g - 1.96\sqrt{m(t_g)}, t_g + 1.96\sqrt{m(t_g)})$ treating $(t_g - Y)/\sqrt{m(t_g)}$ as a standard normal deviate calculated for $R = 1000$ replicates of samples drawn by the same method when all population values y, z, x are given - the closer ACP to 95 the better; and

(ii) ACV, the Average coefficient of variation, which is the average, over the same $R = 1000$ replicates, of the values of $100\frac{\sqrt{m(t_g)}}{t_g}$ - the smaller it is the narrower the CI and the more accurate the point estimator t_g.

The procedures are given in section 5.2, the simulated results in section 5.3 and a few concluding remarks in section 5.4.

5.2 Bootstrap Procedures

With $Q_i(> 0)$ as suitably assignable, for example, as $\frac{1}{x_i}, \frac{1}{x_i^2}, \frac{1}{\pi_i x_i}, \frac{1-\pi_i}{\pi_i x_i}$ etc we have $t_g = \Sigma \frac{y_i}{\pi_i} g_{si} I_{si}$ with $g_{si} = 1 + (X - \Sigma \frac{x_i}{\pi_i}) \frac{x_i Q_i \pi_i}{\Sigma x_i^2 Q_i I_{si}}$. Writing t_w for the HTE for the total of a variable w we may write $t_g = f(t_y, t_x, t_{yxQ\pi}, t_{x^2Q\pi}) = t_y + (X - t_x)\frac{t_{yxQ\pi}}{t_{x^2Q\pi}}$ which is thus a non-linear function of 4 HTE's of the population totals of 4 specified variables.

In case (A) and (B) of section 5.1 hold, Yates and Grundy's (1953) estimator

$$v_{YGS} = \sum\sum_{i<j}(\pi_i\pi_j - \pi_{ij})(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2\frac{I_{sij}}{\pi_{ij}}$$

is a 'uniformly non-negative' (UNN) unbiased estimator for the variance of the HTE $= \Sigma\frac{y_i}{\pi_i}I_{si} = t_H$ namely $V(t_H)$. In this case, Rao and Wu (1988) have given a procedure of drawing a bootstrap sample from the original sample so that the 'Bootstrap'-sampling based expectation E_* and variance V_*, generically denoted, of two functions defined by them may respectively be equated to t_H and v_{YGS}. Taking the cue from them we propose the following 'bootstrap' sampling schemes in the two cases (I) when (A) fails but (B) holds and (II) when both (A) and (B) fail.

Case I. Out of $\nu(s)(\nu(s) - 1)$ 'ordered' pairs of units $i, j (i \neq j)$ in s let a 'Bootstrap' sample s_1^* of pairs (i^*, j^*) in 'm' draws 'with replacement' be chosen with probabilities (to be specified in what follows)

$$q_{i^*j^*}, (i^* \neq j^*), \quad q_{i^*j^*} = q_{j^*i^*}.$$

With numbers $k_{i^*j^*}$ to be presently specified, let

$$t_1 = \frac{1}{m}\sum\sum_{(i^*,j^*\epsilon s_1^*)} k_{i^*j^*}(\frac{y_{i^*}}{\pi_{i^*}} - \frac{y_{j^*}}{\pi_{j^*}}).$$

Then, $E_*(t_1) = \sum\sum_{i\neq j\epsilon s} q_{ij}k_{ij}(\frac{y_i}{\pi_i} - \frac{y_i}{\pi_j}) = 0$ and $V_*(t_1) = \frac{1}{m}\sum\sum_{i\neq j\epsilon s} q_{ij}k_{ij}^2(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2$ which equals v_{YGS} on choosing (a) $m = \nu(s)(\nu(s) - 1)$, (b) $q_{ij} = \frac{1}{m}$ and (c) $k_{ij} = m(\frac{\pi_i\pi_j - \pi_{ij}}{2\pi_{ij}})^{1/2}, i \neq j\epsilon s$. Chaudhuri (2000a) has shown that even if (A) fails, $v_c = v_{YGS} + \Sigma\frac{y_i^2}{\pi_i}\alpha_i\frac{I_{si}}{\pi_i}$, with $\alpha_i = 1 + \frac{1}{\pi_i}(\sum_{j\neq i}\pi_{ij}) - \Sigma\pi_i = \frac{1}{\pi_i}\sum_s \nu(s)p(s)I_{si} - \sum_s \nu(s)p(s)$ is an unbiased estimator for $V(t_H)$.

In order to see the useful role of v_c let us now draw from s a second bootstrap sample s_2^* 'independently' of the selection of s_1^* following the Poisson's sampling scheme, as described by Hájek (1981) with r_i as the probability of

'Success' associated with i in s implementing a Bernoullian trial. Let

$$t_2 = \sum_{i^* \epsilon s_2^*} \frac{y_{i^*}}{\pi_{i^*}} \frac{I_{s_2^* i^*}}{r_{i^*}}.$$

Then, $E_*(t_2) = t_H$ and $V_*(t_2) = \Sigma(\frac{1}{r_i} - 1)(\frac{y_i}{\pi_i})^2 I_{si}$. Letting $t = t_1 + t_2$ and choosing $r_i = \frac{1}{1+\alpha_i}$ one has $E_*(t) = t_H$ and $V_*(t) = v_c$, provided of course $\alpha_i \geq 0$ ensuring $0 \leq r_i \leq 1 \quad \forall i$.

This is just a mimicry of Rao and Wu (1988) when "because (A) fails v_{YGS} is not an unbiased estimator for $V(t_H)$ so that a modification is needed".

Just as t above is calculated using y_i one has to calculate the same t with y_i replaced by $x_i, y_i x_i Q_i \pi_i$ and $x_i^2 Q \pi_i$ so as to be able to calculate t_g based on the bootstrap sample $s^* = (s_1^*, s_2^*)$ as above. Calling such a sample s^* as a bth bootstrap sample s_b^*, one has now to replicate the same, a large number of times, say, $B = 1000$ and calculate t_g for these B replicated bootstrap samples $s_b^*, b = 1, \cdots, B$. Then, $\bar{t}_g = \frac{1}{B} \sum_{b=1}^{B} t_g(s_b^*)$ gives us the 'bootstrap greg estimate' and we take

$$v_g = \frac{1}{B-1} \sum_{b=1}^{B} (t_g(s_b^*) - \bar{t}_g)^2$$

as the Bootstrap estimate of the MSE of the original greg estimator t_g about Y.

Obviously in this MSE estimation though computation of π_{ij} is not avoided there is no problem of correctly computing too many cross-product terms with $(\pi_i \pi_j - \pi_{ij})/\pi_{ij}$ which are usually volatile rendering v_{YG} unstable. A 95% CI for Y is then calculated as $(L_{2.5}, U_{97.5})$ using the lower 2.5% tail point $L_{2.5}$ and the upper 2.5% tail point $U_{97.5}$ of the 'Histogram' of the $t_g(s_b^*)$ values, $b = 1, \cdots, B$. This is by the 'well-known' percentile method. An alternative 'Double Bootstrap' CI may be calculated as follows: From $s_b^*, B = 1000$ more bootstrap samples are independently drawn using the same scheme and they are used to calculate v_g as above – to be denoted by $v_g(b)$. This is repeated for every initial $b = 1, \cdots, B = 1000$. For the histogram of

$$\frac{t_g(s_b^*) - t_g}{\sqrt{v_g(b)}}, b = 1, \cdots, B$$

the lower 2.5% point $l_{2.5}$ and the upper 2.5% point $u_{97.5}$ are then calculated and $(t_g - u_{97.5}\sqrt{v_g}, t_g - l_{2.5}\sqrt{v_g})$ is taken as the 95% 'double bootstrap' CI for Y.

In these computations of CI's the traditional MSE estimators for t_g using the cross-product terms are avoided. With powerful computers the above computations are easy. Next we present our proposed 'Bootstrap' sampling procedures to cover the

Case II. First we note that though in case $\nu(s)$ for every s is a constant it is impossible to have $\pi_{ij} \geq \pi_i\pi_j \quad \forall i, j\epsilon U, (i \neq j)$, in a contrary case it may yet hold especially if for the 'largest sample-size n, say', one has $n \geq 1 + E(\nu(s)) - \pi_i \quad \forall i$ leading to

$$\text{Var } (\nu(s)) \geq \Sigma\pi_i(1 - \pi_i).$$

So, in case (II) holds, let from s a bootstrap sample s_1^* be drawn by Poisson scheme with k_{i*} as the 'probability for success' for a unit i^* in s. Again (2) 'independently' of the draw of s_1^*, let a 'bootstrap' sample s_2^* be drawn from the $\nu(s)(\nu(s) - 1))$ 'ordered' pairs of distinct units of s again by Poisson scheme with λ_{i*j*} as the 'probability of success' for the (i^*, j^*)-paired unit, $(i^* \neq j^*$ in $s)$. Let us construct the bootstrap statistic

$$\begin{aligned} t &= \Sigma\frac{y_{i*}}{\pi_{i*}}(\frac{I_{s_1^*i^*}}{k_{i*}}) + (\sum_{i^*\neq j^*}\Sigma \frac{\sqrt{\frac{y_{i*}}{\pi_{i*}}\frac{y_{j*}}{\pi_{j*}}}}{\lambda_{i*j*}} I_{s_2^*i^*j^*} \\ &\quad - \sum_{i\neq j}\Sigma \sqrt{\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}} I_{sij}) \end{aligned}$$

Then, $E_*(t) = \Sigma\frac{y_i}{\pi_i}I_{si}$ and

$$V_*(t) = \Sigma\frac{y_i^2}{\pi_i^2}(\frac{1}{k_i} - 1)I_{si} + \sum_{i\neq j}\Sigma(\frac{1}{\lambda_{ij}} - 1)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}.$$

Choosing (1) $k_i = \frac{1}{2-\pi_i}$ and (2) $\lambda_{ij} = \frac{1}{2-(\frac{\pi_i\pi_j}{\pi_{ij}})}$ this $V_*(t)$ is equated to the HT form of the estimator of $V(t_H)$, which is $v_{HT} = \Sigma y_i^2(\frac{1-\pi_i}{\pi_i})\frac{I_{si}}{\pi_i} + \underset{i\neq j}{\Sigma\Sigma} y_i y_j (\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j})\frac{I_{sij}}{\pi_{ij}}$. The condition (Π) is needed to ensure $0 < \lambda_{ij} \leq 1 \forall i \neq j$. For the applicability of this 'bootstrap' sampling scheme it is a 'prime requirement that' v_{HT} for s must be 'non-negative' because to it is equated a variance. If $y_i \geq 0 \forall i$, of course v_{HT} is non-negative in case (Π) holds and moreover $\pi_{ij} \geq \pi_i\pi_j \ \forall \ i \neq j$.

5.3 Simulation for a numerical study of efficacies

We have not encountered yet any real-life data for which $\nu(s)$ varies and $\pi_{ij} \geq \pi_i\pi_j \forall i \neq j$. But we shall illustrate 3 specific sampling schemes below for which (I) applies for every vector $\underline{Y} = (y_1, \cdots, y_i, \cdots, y_N)$ of real y_i's, $i\epsilon U$ or for the values encountered for $\underline{Y}$ at hand together with related $\underline{Z}$.

Scheme 1. As discussed in Chapter1 in details, Chaudhuri and Pal (2002a) showed that v_c is uniformly non-negative for the scheme of sampling in which in the first 2 draws units are chosen using given normed size-measures $p_i (0 < p_i < 1 \forall i, \Sigma p_i = 1)$ employing Brewer's (1963) scheme followed by either $(n-2)$ additional draws by simple random sampling (SRS) without replacement (WOR) or $(n-1)$ additional draws by SRSWOR from U leaving aside the 2 distinct units already drawn - here n is pre-specified but the choice of $(n-2)$ is made with an assignable probability $w(0 < w < 1)$ and $(n-1)$ is chosen with probability $1 - w$.

Scheme 2. Following Ray and Das (1997), as discussed in Chapter 3, we consider a 'circular systematic sampling' (CSS) with 'probabilities proportional to sizes' (PPS) which are certain known numbers with the modification on a standard CSSPPS that 'the sampling interval', instead of being a pre-assigned positive integer is chosen as a random integer between 1 and $Z-1$, where $Z = \Sigma z_i$, where z_i's are positive integers chosen as size-measures of i in U. The number of draws is taken as a pre-assigned number n.

Scheme 3. This is the standard PPS scheme with replacement (WR) in a pre-determined number of n draws.

We shall illustrate $N, \underline{Y}, n, z_i, p_i$ etc. for which at least one of Rao's requirements (A), (B) is violated for all the above schemes 1-3. From Särndal, Swensson and Wretman's (SSW, 1992) book, p.660, we consider $N = 50$ clusters and $n = 17$ for schemes 1 and 3, with y as the 'cluster population in 1985', x as the 'cluster population in 1975' and z as the 'number of municipalities' in a cluster, the size-measure, $p_i = \frac{z_i}{Z}, Z = \sum_1^N z_i$. For scheme 2 we take $n = 7$ and $N = 29$, leaving out the last 21 clusters above. For schemes 1 and 3, $Y = 8339, Z = 284$ and for scheme 2, $Y = 5816$ and $Z = 165$. For scheme 1, $w = 0.4$.

Following Chaudhuri and Pal (2002a) we consider for $t_g = \Sigma \frac{y_i}{\pi_i} I_{si} + (X - \Sigma \frac{x_i}{\pi_i} I_{si}) \frac{\Sigma y_i x_i Q_i I_{si}}{\Sigma x_i^2 Q_i I_{si}}$, $Q_i = \frac{1-\pi_i}{\pi_i x_i}$ and the 2 MSE-estimators as, for $k = 1, 2$, $m_k(t_g) = \underset{i<j}{\Sigma\Sigma} (\pi_i \pi_j - \pi_{ij}) \frac{I_{sij}}{\pi_{ij}} (\frac{a_{ki} e_i}{\pi_i} - \frac{a_{kj} e_j}{\pi_j})^2 + \Sigma \frac{(a_{ki} e_i)^2}{\pi_i} \alpha_i \frac{I_{si}}{\pi_i}$ with $a_{1i} = 1, a_{2i} = g_{si}, i \epsilon s$; $e_i = y_i - b_Q x_i, b_Q = \frac{\Sigma y_i x_i Q_i I_{si}}{\Sigma x_i^2 Q_i I_{si}}$. Next $\sqrt{m_k(t_g)}$ is taken as the Standard Error (SE) of t_g. In Tables 1-3 below we present, for the respective schemes 1-3, (a) for a specific replicate of a sample the values of $t_g, \bar{t}_g, SE$ separately as (1) $\sqrt{v_g}$, (2) $\sqrt{m_1(t_g)}$ and (3) $\sqrt{m_2(t_g)}$, CI's as (1)$'$ by percentile method, (2)$'$ by Double bootstrap method, (3)$'(t_g \pm 1.96\sqrt{m_1(t_g)})$ and (4)$'(t_g \pm 1.96\sqrt{m_2(t_g)}$, and lengths of CI's and (b) for $R = 1000$ replicates of the original samples (i) the average lengths (AL) of CI's, (ii) ACP's, and (iii) ACV's for the pertinent procedures that we could actually implement with the given data described above.

The nature of the distribution of the statistic $e = \frac{t_g - Y}{s.e(t_g)}$ may be examined through the skewness and kurtosis coefficients γ_1 and γ_2 (defined below) for the T=1000 replicates drawn.

Here $\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}}$ and $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$

where $\mu_r = \frac{1}{T}\Sigma(e - \bar{e})^r, \bar{e} = \frac{1}{T}\Sigma e, r = 2, 3, 4$ over the replicates.

Table-0

Showing values of γ_1, γ_2 for different sampling schemes

Sampling scheme (1)	γ_1 (2)	γ_2 (3)
Scheme 1	.02	.76
Scheme 2	-.03	-.68
Scheme 3	-.54	.93

From the calculated values given above, departure from normality seems to be evident. So it is desirable to employ an alternative procedure avoiding normality. So we use bootstrap samples to construct CI's by 1) Percentile method and also 2) by the Double bootstrap method.

Table 1

Performance characteristics based on Scheme 1

t_g	$\bar{t}_g$	SE	CI	Length of CI	AL	ACP	ACV
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
8436.0	8439.4	(1) 269.0	(1)′ (7978.7, 8780.0)	801.3	958.6	98.1	(1) 3.4
		(2) 120.3	(2)′ (7979.0, 8647.0)	668.0	549.0	93.4	(2) 1.0
		(3) 112.0	(3)′ (8059.9, 8514.1)	454.2	338.8	90.7	(3) 2.1
			(4)′ (8117.6, 8456.3)	338.7	351.0	92.1	-

Table 2

Performance characteristics based on Scheme 2

t_g	$\bar{t}_g$	SE	CI	Length of CI	AL	ACP	ACV
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
5598.5	5723.2	(1) 204.6	(1)′ (5530.7, 6116.0)	585.3	652.1	94.5	(1) 6.1
		(2) 180.1	(2)′ -	-	-		(2) 3.5
		(3) 106.9	(3)′ (5343.4, 5853.5)	510.1	549.1	91.1	(3) 2.4
			(4)′ (5485.3, 5711.6)	226.3	473.2	90.3	

Table 3

Performance characteristics based on Scheme 3

t_g	$\bar{t}_g$	SE	CI	Length of CI	AL	ACP	ACV
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
8462.1	8440.8	(1) 408.0	(1)′ (7583.9, 9395.9)	1776.0	1503.4	99.2	(1) 7.3
		(2) 141.6	(2)′ (8157.4, 8620.7)	463.3	672.8	97.1	(2) 2.2
		(3) 83.7	(3)′ (8258.7, 8813.9)	555.2	493.6	89.1	(3) 1.3
			(4)′ (8372.2, 8700.4)	328.2	446.1	87.1	

5.4 A Discussion on the Simulated Results and Concluding Remarks

Our proposed bootstrap method in "Case (I)" gives a bootstrap estimate quite close to the population total and gives the best ACP-values among its competitors. But the bootstrap standard errors and ACV-values come out much poorer, with too wide lengths of the CI's. The sophisticated standard errors for the greg estimator especially the ones using g_{si}-values give better SE's, ACV's and lengths of CI's but the 'coverage-probabilities' turn out less than the 'desirable' magnitudes.

Even though for the use of HTE for Y the PPSWR method may not be a good sampling scheme it competes well in respect of the greg estimator against the more reasonable sampling schemes 1 and 2.

We may conclude however that though there is a clamour against the use of traditional MSE-estimators for the greg estimator because of too many cross-product terms a 'bootstrap' technique though available need not be good enough to dislodge them in terms of performances. It may be useful to study how our bootstrap procedures may compete against the emerging approximate MSE-estimators that bypass the involvement of the 'cross-product' terms as are recommended by Särndal (1996).

A final remark. Brewer and Gregoire's (2000) estimator

$$t_{BG} = \frac{\Sigma\pi_i}{\nu(s)}\Sigma\frac{y_i}{\pi_i}I_{si},$$

though not originated through them, based on Poisson's scheme of sampling with $\pi_i(0 < \pi_i < 1, \forall i \epsilon U)$ as the 'probability of success for i' is really a greg estimator, as was noted by Van Deusen (1987) for the choices: $Q_i = \frac{1}{\pi_i x_i}$ and $x_i = \pi_i$. Here $\nu(s)$ varies over $(0, 1, 2, \cdots, N-1, N)$. So, for this our procedures discussed above may be applied with no difficulty, as in this case $\pi_{ij} = \pi_i\pi_j \quad \forall i \neq j$ in U. We do not pursue with this here.

Chapter 6

Estimating Domain-wise Distribution of Scarce Objects by Adaptive Sampling and Model-based Borrowing of Strength

Abstract

Utilizing the known (1) geographical areas (z) of the districts in India and (2) those of the total wastelands (x) therein we consider estimating (3) the total unknown areas (y), under 'Mining and Industrial Wastelands' for the groups of districts together but separately in the Northern, Southern, Eastern and Western regions in India, restricting to districts each possessing at least 5 per cent of its total area as a wasteland. The total numbers of such districts in these respective 4 regions of separate interest are 48,48,42 and 91 giving a total of 229 out of which we consider sampling a total of 73 districts employing Rao, Hartley and Cochran's (RHC, 1962) scheme of sampling

using the known values of z above as the size-measures. Treating the above 4 regions of districts as the 4 domains of interest we consider utilizing known x above as a regressor in estimating the 'domain total' values of y above to form an idea of the distribution of these district-wise scarce objects in these regions.

For this we employ (a) non-synthetic as well as (b) synthetic versions of generalized regression (greg) estimators motivated respectively by postulated regression lines of y on x through the origin, for simplicity, with (i) domain-specific and alternatively with (ii) domain-invariant 'slope-parameters'.

Next we employ empirical Bayes estimators (EBE) with these greg estimators as the 'initials' with further specifications in the models.

Finally, in order to capture more districts beyond the 'initial sample' accommodating the rare commodities namely the 'mining and industrial wastelands' we employ the technique of Adaptive sampling defining appropriate (1) 'neighbourhoods' and (2) 'networks'. One may refer to Thompson (1992), Thompson and Seber (1996) and Chaudhuri (2000a) for a discussion on adaptive sampling technique. The resulting relative performances of the alternative estimators noted above based on 'initial' and 'adaptive' samples are numerically examined through a simulation exercise utilizing known values of all the 3 variables noted above based on a given set of 'Remote sensed' observations.

The synthetic greg estimates based on adaptive samples turn out to be the most promising ones in terms of the standard twin criteria of (A) actual coverage percentage (ACP) of confidence intervals (CI) based on assumed normality of a standardized 'pivotal' derived from a 'domain-specific' estimator and of (B) average coefficient of variation (ACV) of an estimator both calculated from 'replicated samples'.

6.1 Introduction

From the website "envfor.nic.in/naeb/naeb.html" entitled "The National Wasteland Identification Project" (NWIP) we gather certain data relating to 48, 48, 42 and 91 districts, each with at least 5 per cent of its total area as 'a Wasteland' area respectively in the northern region of UP, Haryana,Himachal Pradesh, Punjab and Jammu & Kashmir states, the southern region of Karnataka, Andhra Pradesh, Tamil Nadu and Kerala states, the eastern region composed of Arunachal Pradesh, Nagaland, Manipur, Assam, West Bengal, Orissa and Bihar and the western region consisting of the states of Maharashtra, Gujarat, Goa, Rajasthan and Madhya Pradesh.

For each of these 229 districts are separately known the total (1) geographical area (z), (2) the total 'wasteland area' (x) and (3) the total 'mining and industrial wasteland area' (y). Since the value of y for many of the districts is zero while when it is positive its magnitude is substantial and 'how far the remote-sensed data on y matches the ground realities' is unknown, we consider it useful to prescribe, through a prior investigation, a fruitful method of (A) sampling of these 229 districts and of (B) estimating the total values of y for all the districts together but separately within the above-noted 4 regions of interest.

Using the known values of z as size-measures it seems plausible to adopt a suitable 'unequal probability sampling' scheme to start with and since x-values are known, a generalized regression estimator seems worthy of application. Further, since even with as high as a 25% sample of districts we may not find enough 'region-wise' sample-sizes it may be useful to apply the 'principle of borrowing strength' as in small area estimation by appropriate modelling. Finally, since y is positive only for a very few districts region-wise, in order to capture more districts with positive y's we may contemplate employing adaptive sampling to extend the original sample to hope for improved estimation.

In section 6.2 we describe the procedures of sample selection and the estimation methods along with the motivating models. In section 6.3 we present a numerical evaluation of the competing procedures by a simulation exercise. We give our recommendations in the section 6.4 with which we conclude.

6.2 Sampling and Estimation Methods

For a simple presentation we need the following notations. Let $U = (1, \cdots, i, \cdots, N)$ denote a population of units labelled $i = 1, \cdots, N$ and let this be a union of D non-overlapping sets of units U_d, called 'domains', with known sizes $N_d, d = 1, \cdots, D$. Let $y_i, x_i, z_i, i \epsilon U$ be the values of the variables respectively y, x, z with (1) totals Y, X, Z and (2) domain totals $Y_d, X_d, Z_d, d = 1, \cdots, D$. By $p_i = \frac{z_i}{Z}$, we shall denote the 'normed size-measures' of the units.

From U let a sample of n units be chosen employing the Rao-Hartley-Cochran (RHC, 1962) scheme. For this, U is randomly divided into n groups of $M_1, \cdots, M_i, \cdots, M_n$ units with M_i's as integers closest to $\frac{N}{n}$ with their sum $\Sigma_n M_i$ over the n groups equal to N. From the ith group so formed one unit, say, ij is chosen with a probability $\frac{p_{ij}}{r_i}$, writing $r_i = p_{i1} + \cdots + p_{iM_i}$; this is repeated independently over all these n groups.

Let $I_{di} = 1$ if $i \epsilon U_d$; 0 else and (p_i, y_i) be the normed size-measure and the y-value for the unit chosen from the ith group. Let Σ denote summing over i in U. Then,

$Y_d = \Sigma y_i I_{di}$ and RHC's unbiased-estimator for Y_d is

$$\hat{Y}_d = \Sigma_n \frac{r_i}{p_i} y_i I_{di}.$$

Writing $B = \frac{\Sigma_n M_i^2 - N}{N^2 - \Sigma_n M_i^2}$, RHC's unbiased estimator of $V(\hat{Y}_d)$, the variance of $\hat{Y}_d$ is $v(\hat{Y}_d) = B \Sigma_n \Sigma_n r_i r_j (\frac{y_i I_{di}}{p_i} - \frac{y_j I_{dj}}{p_j})^2$, writing $\Sigma_n \Sigma_n$ as sum over pairs of distinct groups with no overlaps. Let us postulate a model so that we may write

$$y_i = \beta_d x_i + \epsilon_i, i\epsilon U_d, d = 1, \cdots, D$$

with β_d as constants and ϵ_i's as random variables. Following Chaudhuri et al (1995) we may employ the following version of a possible improvement upon $\hat{Y}_d$, namely

$$\begin{aligned} t_{gd} &= \Sigma_n \frac{r_i}{p_i} y_i I_{di} + b_{Qd}(X_d - \Sigma_n \frac{r_i}{p_i} x_i I_{di}) \\ &= \Sigma_n \frac{r_i}{p_i} g_{di} y_i I_{di}. \end{aligned}$$

Here $b_{Qd} = \Sigma_n y_i x_i Q_i I_{di} / \Sigma_n x_i^2 Q_i I_{di}$ with Q_i as a suitably assignable positive constant, for example, as $\frac{1}{x_i}, \frac{1}{x_i^2}, \frac{r_i}{p_i x_i}, \frac{1-\frac{p_i}{r_i}}{\frac{p_i}{r_i} x_i}$, etc. and $g_{di} = 1 + (X_d - \Sigma_n \frac{r_i}{p_i} x_i I_{di}) \frac{x_i Q_i \frac{p_i}{r_i}}{\Sigma_n x_i^2 Q_i I_{di}}$

In this presentation we shall mostly take Q_i as $(\frac{1-\frac{p_i}{r_i}}{p_i x_i / r_i}) = \frac{r_i - p_i}{p_i x_i}$. Letting $e_{di} = y_i - b_{Qd} x_i$, following Särndal (1982) the mean square error (MSE) of t_{gd} about Y_d may be estimated by

$m_{kgd} = B \Sigma_n \Sigma_n r_i r_j (\frac{e_{di} a_{kdi} I_{di}}{p_i} - \frac{e_{dj} a_{kdj} I_{dj}}{p_j})^2, k = 1, 2$ on writing $a_{1di} = 1$ and $a_{2di} = g_{di}$. In order to improve upon t_{gd} by 'borrowing strength' from outside the 'intersection of the sample with U_d' but within the initially chosen sample, s let us postulate an alternative model for which β_d above is replaced by β for every d but keeping everything else in tact. This revised model motivates the 'synthetic' greg predictor for Y_d as t_{gSd} which is t_{gd} with b_{Qd} replaced by $b_Q = \Sigma_n y_i x_i Q_i / \Sigma_n x_i^2 Q_i$. Then we may write

$$\begin{aligned} t_{gSd} &= \Sigma_n \frac{r_i}{p_i} y_i [I_{di} + (X_d - \Sigma_n \frac{r_i}{p_i} x_i I_{di}) \frac{x_i Q_i \frac{p_i}{r_i}}{\Sigma_n x_i^2 Q_i}] \\ &= \Sigma_n \frac{r_i}{p_i} y_i g_{sdi}, \text{ say,} \end{aligned}$$

with g_{sdi} as 'within the square brackets'. Then, following Särndal (1982), MSE-estimators for t_{gSd} are

$$m_{kgSd} = B \Sigma_n \Sigma_n r_i r_j (\frac{e_i b_{kdi}}{p_i} - \frac{e_j b_{kdj}}{p_j})^2, k = 1, 2$$

on writing

$$e_i = y_i - b_Q x_i, b_{1di} = I_{di}, b_{2di} = g_{sdi}.$$

In contrast with t_{gSd}, the t_{gd} is a 'non-synthetic' greg predictor.

Writing t_d for an initial estimator/predictor for Y_d let us now postulate the more sophisticated model permitting us to write:

(i) $t_d | Y_d \overset{ind}{\cap} N(Y_d, m_d)$, with m_d as a known MSE-estimator for t_d,

(ii) $Y_d \overset{ind}{\cap} N(\theta X_d, A)$, θ, A as unknown constants

(iii) $\epsilon_d = (t_d - Y_d)$ "independent" of $\eta_d = Y_d - \theta X_d$ for $d = 1, \cdots, D$.

Then, from Fay and Herriot (1979) we have

$$t_{Bd} = (\frac{A}{A + m_d}) t_d + (\frac{m_d}{A + m_d})(\theta X_d)$$

as the Bayes estimator of $Y_d, d = 1, \cdots, D$.

$$\text{Writing } \tilde{\theta} = \frac{\sum_{d=1}^{D} t_d X_d / (A + m_d)}{\sum_{d=1}^{D} X_d^2 / (A + m_d)}$$

and solving by iteration for θ and A starting with a 'zero value for A', the equation

$$\sum_{d=1}^{D} (t_d - \tilde{\theta} X_d)^2 / (A + m_d) = D - 1,$$

we may derive moment estimates $\hat{\theta}, \hat{A}$ respectively for θ, A. Then,

$$t_{EBd} = (\frac{\hat{A}}{\hat{A} + m_d}) t_d + (\frac{m_d}{\hat{A} + m_d})(\hat{\theta} X_d)$$

gives the EBE for Y_d.

From Prasad and Rao (1990) we get the MSE-estimator for t_{EBd} as

$$\begin{aligned} m_{EBd} &= g_{1d}(\hat{A}) + g_{2d}(\hat{A}) + 2g_{3d}(\hat{A}) \\ \text{where} \quad g_{1d}(\hat{A}) &= \gamma_d m_d \\ \gamma_d &= \frac{\hat{A}}{\hat{A} + m_d} \\ g_{2d}(\hat{A}) &= (1 - \gamma_d^2)\frac{X_d^2}{\sum\limits_{d=1}^{D} \frac{X_d^2}{(\hat{A} + m_d)}} \\ g_{3d}(\hat{A}) &= \frac{m_d^2}{(\hat{A} + m_d)^3} V(\bar{A}) \\ \text{where} \quad V(\bar{A}) &= \frac{2}{D^2} \sum_{d=1}^{D} (\hat{A} + m_d)^2 \end{aligned}$$

For the validity of m_{EBd}, D is required to be large. But in the present case we employ this even though D is only four hoping that this may still work.

Suspecting that the initial sample s drawn as above may not yield enough units with positive values of y 'respective domain-wise', we may apply in the following way the technique of adaptive sampling to enhance the capture of more sampled units with positive and possibly high positive y-values.

For every unit, namely district in the present investigation, let a 'neighbourhood' be defined as the collection of districts including this unit itself and those with a common boundary with it as is determined from the map of the 229 districts we are considering.

Any unit, rather district with a zero value for y is called an 'edge' unit or a singleton network. For any unit with a positive y-value one should check for the positive/zero-value of y for each of its neighbouring units and proceed with this checking until every neighbouring unit has a zero value. The 'set of units thus checked starting with the positive y-valued unit' constitutes a 'cluster' for the unit including itself. Those units with positive y-values in the cluster constitute a 'network' for the initial unit. Writing $A(i)$ for the 'network' to which the unit i belongs and m_i for its cardinality, let

$$t_i = \frac{1}{m_i} \sum_{k \epsilon A(i)} y_k, l_i = \frac{1}{m_i} \sum_{k \epsilon A(i)} x_k.$$

Then, as is recorded by Chaudhuri (2000a), one may check that $T = \sum_{i \epsilon U} t_i$ equals Y and $L = \sum_{i \epsilon U} l_i$ equals X.

Similarly, letting

$$\begin{aligned} t_{id} &= \frac{\sum_{j I_{dj} \epsilon A(i)} y_j I_{dj}}{(\sum_{j I_{dj} \epsilon A(i)} 1)} \text{ and } l_{id} = \frac{\sum_{j I_{dj} \epsilon A(i)} x_j I_{dj}}{(\sum_{j I_{dj} \epsilon A(i)} 1)}, \text{ it follows that} \\ Y_d &= \sum_{i \epsilon U} t_{id} = T_d, \text{ say, } X_d = \sum_{i \epsilon U} l_{id} = L_d, \text{ say .} \end{aligned}$$

The collection of the units in the original sample s together with those in their respective clusters constitutes an adaptive sample.

Corresponding to t_{gd} the non-synthetic greg predictor for Y_d based on the adaptive sample is
$t_{gd}(A) = (\Sigma_n \frac{r_i}{p_i} t_{id} I_{di}) + b_{Qd}(A)(X_d - \Sigma_n \frac{r_i}{p_i} l_{id} I_{di})$ writing $b_{Qd}(A) = \frac{\Sigma_n t_{id} l_{id} Q_i I_{di}}{\Sigma_n l_{id}^2 Q_i I_{di}}$.

Since l_{id} is often zero, we shall take Q_i as $(1 - \frac{p_i}{r_i})/\frac{p_i}{r_i}$ omitting l_{id} in the denominator which we might use as equivalent to x_i.

The MSE estimators for $t_{gd}(A)$ are $m_{kgd}(A)$ obtained from m_{kgd} replacing therein y_i by t_{id}, x_i by l_{id}, b_{Qd} by $b_{Qd}(A)$.

Instead of t_{gSd} we shall employ $t_{gSd}(A)$ for the adaptive sample obtained on replacing y_i, x_i by t_{id}, l_{id} in the former. The MSE estimator for $t_{gSd}(A)$ will be taken as $m_{gSd}(A)$ obtained from m_{1gd} on replacing y_i, x_i in the latter by t_{id} and l_{id} respectively in the terms involving I_{di} and by t_i, l_i for the terms free of I_{di}. Because of the form of $m_{gSd}(A)$ it is not possible to use a second MSE-estimator corresponding to m_{gSd} because

$$t_{gSd}(A) = \Sigma_n \frac{r_i}{p_i} t_{id} I_{di} + (L_d - \Sigma_n \frac{r_i}{p_i} l_{id} I_{di})(\frac{\Sigma_n t_i l_i Q_i}{\Sigma_n l_i^2 Q_i})$$

cannot be expressed as a weighted sum of $(t_{id}I_{di})$-values. Corresponding to $(t_{EBd}, m_{kEBd}), (t_{EBSd}, m_{kEBSd})$ we obviously have $(t_{EBd}(A), m_{kEBd}(A)), k = 1, 2$ and $(t_{EBSd}(A), m_{EBSd}(A))$ with obvious notations for the EB estimators based on adaptive sampling and the MSE-estimators corresponding to $m_{kEBd}, k = 1, 2$ and m_{1EBSd}.

6.3 Simulation - based Numerical Evaluation of Relative Efficacies

Given an estimator/predictor f_d for Y_d with an MSE-estimator v_d we shall treat $s_d = (f_d - Y_d)/\sqrt{v_d}$ as a standard normal deviate and take $(f_d - 1.96\sqrt{v_d}, f_d + 1.96\sqrt{v_d})$ as the 95% confidence interval (CI) for Y_d. To compare alternative choices of (f_d, v_d) we shall calculate, based on $R = 1000$ replicates of the samples, the criteria measures (I), ACP, the actual coverage percentage which is the percent of the replicated samples with CI's covering Y_d - the closer it is to 95 the better and (II) ACV, the average coefficient of variation namely the average over the R replicates of the values of $100 \frac{\sqrt{v_d}}{f_d}$ the less it is the less the width of CI and the more accurate is the point estimator f_d for Y_d.

For the NWIP data mentioned earlier our numerical observations are as in the Table below.

Table

Relative Efficacies of Alternative Procedures

ACP/ACV Values

Serial	Domain	North	South	East	West
Numbers	Specifications				
of items	numbered (d)	(1)	(2)	(3)	(4)
	Domain sizes				
	N_d	48	48	42	91
	f_d/v_d				
1.	t_{gd}/m_{1gd}	80.2/20.0	83.4/24.1	80.6/23.9	87.1/14.1
2.	t_{gd}/m_{2gd}	82.1/21.3	84.4/27.2	83.5/22.2	88.0/17.2
3.	$t_{gd}(A)/m_{1gd}(A)$	87.1/15.6	89.4/19.3	86.3/21.4	93.0/9.8
4.	$t_{gd}(A)/m_{2gd}(A)$	89.4/19.1	92.7/26.3	88.1/25.9	93.6/10.3
5.	t_{EBd}/m_{1EBd}	90.3/23.5	91.1/32.0	84.1/39.1	92.1/22.2
6.	t_{EBd}/m_{2EBd}	92.5/31.4	93.7/34.1	83.1/32.3	97.1/31.6
7.	$t_{EBd}(A)/m_{1EBd}(A)$	94.6/41.3	90.1/29.2	86.3/41.4	96.3/29.5
8.	$t_{EBd}(A)/m_{2EBd}(A)$	94.4/37.4	92.1/32.2	88.1/43.5	95.1/24.7
9.	t_{gSd}/m_{1gSd}	88.8/27.1	82.1/29.1	85.6/25.1	89.6/18.2
10.	t_{gSd}/m_{2gSd}	91.3/28.5	84.9/29.3	83.6/24.9	90.0/19.1
11.	$t_{gSd}(A)/m_{gSd}(A)$	93.9/18.5	92.4/20.1	89.9/23.8	96.1/10.5
12.	t_{EBSd}/m_{1EBSd}	95.4/24.3	90.3/34.9	89.1/40.2	95.1/24.1
13.	t_{EBSd}/m_{2EBSd}	95.8/32.9	90.6/29.6	94.8/37.1	96.1/29.5
14.	$t_{EBSd}(A)/m_{EBSd}(A)$	95.3/42.1	94.9/30.1	91.2/43.2	95.1/26.3

It may be noted that the number of districts to be covered by adaptive sampling varies between 117 and 146 with an average of 134 while the initial sample size is only 73. Adaptive sampling always involves additional costs. The question is whether and how much it pays in terms of gain in accuracy in estimation.

6.4 Concluding Remarks and Recommendations

If guided by the criterion of ACP, one may be convinced that empirical Bayes estimators for adaptive samples as well as the original samples fare better than t_{gd} and t_{gSd} and more so if coupled with m_{2gd}, m_{2gSd} respectively rather than with m_{1gd}, m_{1gSd}.

Moreover, adaptive sampling coupled with non-synthetic, synthetic greg estimators and the empirical Bayes estimators based thereupon seems to have an edge over the original one.

In terms of the ACV criterion empirical Bayes methods perform poorer than the initial ones on which they are based. But adaptive sampling achieves improvements when combined with t_{gd} with both m_{1gd}, m_{2gd} and also with t_{gSd} but the ACV increases when it is used with empirical Bayes versions of these greg estimators. Taking both the criteria together, the synthetic greg estimator $t_{gSd}(A)$ based on adaptive sampling seems to be the most promising one. So, if the resources permit, our recommendation is in favour of adaptive sampling even at an additional Cost. Compared to $(t_{gSd}, m_{kgSd}), k = 1, 2$, the pair $(t_{gSd}(A), m_{gSd}(A))$ is a better choice - this vindicates the efficacy of adaptive sampling. Keeping in mind simultaneously the width of the confidence interval and the accuracy in point estimation, empirical Bayes procedure does not seem to be a right option in the present exercise. But adaptive sampling coupled with synthetic greg estimator is a promising choice.

A possible reason for a partial failure of the empirical Bayes estimation approach in the present exercise may be the inadequacy of m_{EBd} as an MSE-estimator in view of the number of domains here being too small-only four.

A repulsive feature of adaptive sampling is its lack of control on the ultimate sample-size. Salehi and Seber (1997, 2002) have invented certain safeguards against excessive inflation in sample-sizes by some ingeneous devices. One easy way available for the type of work presented in this chapter is to (1) first set an upper limit on the total of the cardinality of all the networks that one may come across on the basis of the initial sample and keeping that inview suitably sub-sample, by SRSWOR method, each network covered during the actual survey and keep the cost under more control than for the uncontrolled Adaptive sampling. In our simulation illustrated in this Chapter, there was no excessive increase over the initial sample size and so we did not apply any precautionary measures.

Chapter 7

Simplified variance and mean square error estimation avoiding inclusion-probabilities of paired units

Abstract

Särndal (1996) followed by Deville (1999), Brewer (1999,2000), Brewer and Gregoire (2000) among others recommend avoiding the terms involving the inclusion-probabilities of pairs of units in the estimators of variance of Horvitz and Thompson's (HT, 1952) estimator and of the mean square errors (MSE's) of the generalized regression (greg) estimators derived from the HT estimator by certain ingeneous ways. For Hájek's (1964, 1981) Poisson sampling scheme and its special case called Bernoullian sampling scheme these probabilities are not even needed in variance or MSE estimation. Certain developments concerning these topics are available in the literature. We attempt here at adding to them a few by analytical as well as numerical exercises.

7.1 Introduction

Hájek's (1964, 1981) Poisson scheme of sampling associates numbers $\theta_i (0 < \theta_i < 1)$ with the respective units i in a survey population $U = (1, \cdots, i, \cdots, N)$ and includes the units in a sample s from U for which the N Bernoullian trials independently implemented yield 'success'es with θ_i as the probability of 'success' for the ith unit of U, omitting the units for which there are 'failures' with probabilities $(1 - \theta_i)$ for i in U. In case θ_i is taken as a common number for every i in U, then the scheme is called Bernoullian sampling scheme. The following consequences are of interest for this scheme.

(i) $\nu(s)$, the number of distinct units in the sample s is a random variable with possible values $0, 1, \ldots, N-1, N$;

(ii) θ_i equals π_i, the inclusion probability - hence we shall write π_i for θ_i throughout for this scheme;

(iii) $E(\nu(s)) = \Sigma\pi_i = \nu$, say; in practice, a number ν is first fixed keeping in view the cost of a survey and π_i's are chosen as numbers in $(0, 1)$ subject to $\Sigma\pi_i = \nu$,

(iv) The HT for Σy_i namely

$$t_H = \Sigma\frac{y_i}{\pi_i}I_{si}, \quad \begin{aligned} I_{si} &= 1 \text{ if } i\epsilon s. \\ &= 0, \text{ else} \end{aligned}$$

has the variance $V(t_H) = \Sigma y_i^2 \frac{1-\pi_i}{\pi_i}$ because, for this scheme $\pi_{ij} = \pi_i\pi_j$ and hence the cross product terms vanish

and (v) $v(t_H) = \Sigma y_i^2 (\frac{1-\pi_i}{\pi_i}) \frac{I_{si}}{\pi_i}$ is an unbiased estimator of $V(t_H)$.

Brewer, Early and Joyce (1972) and Brewer, Early and Hanif (1984) have considered a 'modified Poisson' scheme introduced by Ogus and Clark (1971) where the selection process is repeated in case $\nu(s)$ turns out zero and stopped as soon as $\nu(s)$ turns out 'positive' and then the Poisson scheme is applied with revised selection-probabilities to retain the prescribed π_i's. Some details are given in Chaudhuri and Vos (1988, p.198). Grosenbaugh's (1965) 3P-sampling is a precursor to Ogus and Clark's (1971) above-mentioned introduction as is lately gathered through a private communication.

For "modified Poisson" scheme (MPS) $\pi_{ij} = \pi_i \pi_j (1 - P_0)$ where P_0 is the probability of an empty sample. Thus P_0 is the solution of the equation $\prod_{i=1}^{N} [1 - \pi_i(1 - P_0)] - P_0 = 0$ because $\pi_i(1 - P_0)$ is the revised selection probability of i for this MPS. Then the variance of the HTE (t_H) based on MPS reduces to

$$V'(t_H) = \Sigma(1 - \pi_i)\frac{y_i^2}{\pi_i} - P_0(Y^2 - \Sigma y_i^2)$$

with an unbiased estimator of $V'(t_H)$ as

$$v'(t_H) = \Sigma(1 - \pi_i)\frac{y_i^2}{\pi_i^2} I_{si} - \frac{P_0}{1 - P_0}(t_H^2 - \Sigma\frac{y_i^2}{\pi_i^2} I_{si}).$$

When employing the original Poisson scheme of sampling, an approach is to use, instead of t_H, the ratio estimator, namely,

$$t_{RH} = \frac{\nu}{\nu(s)} t_H = \frac{\nu}{\nu(s)} \Sigma \frac{y_i}{\pi_i} I_{si}$$

for Y, assuming $\nu(s) > 0$.

Grosenbaugh (1965), however introduced this t_{RH}.

Writing $x_i = \pi_i, X = \Sigma x_i = \Sigma \pi_i = \nu, Q_i = \frac{1}{\pi_i x_i} = \frac{1}{\pi_i^2}$ it follows that t_{RH} equals the generalized regression (greg) estimator for Y, namely

$$t_g = \Sigma\frac{y_i}{\pi_i} I_{si} + b_Q(X - \Sigma\frac{x_i}{\pi_i} I_{si}),\ b_Q = \frac{\Sigma y_i x_i Q_i I_{si}}{\Sigma x_i^2 Q_i I_{si}}.$$

Such a fact was earlier recognized by Deusen (1987).

Since, $e_i = y_i - b_Q x_i$ equals $y_i - \left(\frac{\Sigma \frac{y_i}{\pi_i} I_{si}}{\nu(s)}\right)\pi_i$, two usual MSE estimators for $t_g = t_{RH}$ are

$$m_1 = \Sigma\frac{1 - \pi_i}{\pi_i}\left(y_i - \frac{\Sigma\frac{y_i}{\pi_i} I_{si}}{\nu(s)}\pi_i\right)^2 \frac{I_{si}}{\pi_i}$$

and

$$m_2 = \left(\frac{\nu}{\nu(s)}\right)^2 m_1$$

because

$$g_{si} = 1 + \left(X - \Sigma\frac{x_i}{\pi_i}I_{si}\right)\frac{x_iQ_i\pi_i}{\Sigma x_i^2Q_iI_{si}}$$

equals $\frac{\nu}{\nu(s)}$ for every i in s.

For Poisson sampling scheme an alternative MSE estimator for t_{RH} is, according to Brewer (2000),

$$m_{BRH} = \Sigma(\frac{1}{C} - \pi_i^*)(\frac{y_i}{\pi_i^*} - \frac{t_{RH}}{\nu(s)})^2 I_{si}$$

where $\pi_i^* = \frac{\pi_i\nu(s)}{\nu}$, the adjusted inclusion probability and $C = \dfrac{\nu(s) - 1}{\nu(s) - \frac{1}{\nu(s)}\sum\limits_{i=1}^{N}\pi_i^2}$.

From Särndal (1980) we know that Cassel, Särndal and Wretman's (CSW, 1976) greg estimator t_g above for Y as derived from the HT estimator $t_H = \Sigma\frac{y_i}{\pi_i}I_{si}$ being motivated by a modelled regression of y on x which is linear through the origin, has the property of being asymptotically design unbiased (ADU) and also asymptotically design consistent (ADC) for Y. A generalized version of t_g may be taken as

$$t_{gb} = \Sigma y_ib_{si}I_{si} + b_Q\left(X - \Sigma x_ib_{si}I_{si}\right)$$

with b_{si}'s as constants free of $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$ but subject to the constraint

$$E_p(b_{si}I_{si}) = 1 \; \forall i \in U,$$

writing E_p as the operator for expectation with respect to a sampling design P.

This t_{gb} shares with t_g the ADU and ADC properties. Chaudhuri and Stenger (1992) have discussed these properties in depth.

If we relax the requirement of design unbiasedness of $\Sigma y_i b_{si} I_{si}$, then, for Poisson sampling scheme we may employ for Y the estimator

$$t_{gc} = \frac{\nu}{\nu(s)} \Sigma \frac{y_i}{\pi_i} I_{si} + b_Q \left(X - \frac{\nu}{\nu(s)} \Sigma \frac{x_i}{\pi_i} I_{si} \right)$$

with

$$b_Q = \frac{\Sigma y_i x_i Q_i I_{si}}{\Sigma x_i^2 Q_i I_{si}}$$

with

$$Q_i = \frac{1}{x_i}, \frac{1}{x_i^2}, \frac{1}{\pi_i x_i}, \frac{1 - \pi_i}{\pi_i x_i}$$

etc.

Since $\frac{\nu}{\nu(s)} \Sigma \frac{y_i}{\pi_i} I_{si}$ is already shown to be a greg estimator for Y with $x_i = \pi_i, Q_i = \frac{1}{\pi_i x_i} = \frac{1}{\pi_i^2}$, it is ADU and ADC for Y. Hence t_{gc} is also ADU and ADC for Y. This is our justification for the use of t_{gc}.

Since $t_g = \Sigma \frac{y_i - b_Q x_i}{\pi_i} I_{si} + b_Q X = \Sigma \frac{e_i}{\pi_i} I_{si} + b_Q X$ and by Taylor Series expansion, neglecting suitable terms it is well-known that the MSE of t_g about Y is approximated by

$$M(t_g) = MSE \left(\Sigma \frac{e_i}{\pi_i} I_{si} \right) \simeq MSE \left(\Sigma \frac{e_i}{\pi_i} g_{si} I_{si} \right)$$

it follows that for

$$t_{gc} = \frac{\nu}{\nu(s)} \left[\Sigma \frac{e_i}{\pi_i} I_{si} \right] + b_Q X$$

one may approximate the MSE of t_{gc} about Y by

$$M(t_{gc}) = MSE \left[\frac{\nu}{\nu(s)} \Sigma \frac{e_i}{\pi_i} I_{si} \right].$$

So, just as $M(t_g)$ may be estimated by

$$m_k(t_g) = \Sigma \left(\frac{1 - \pi_i}{\pi_i} \right) (a_{ki} e_i)^2 \frac{I_{si}}{\pi_i}, \ k = 1, 2$$

where

$$a_{1i} = 1, a_{2i} = g_{si} = 1 + \left(X - \Sigma \frac{x_i}{\pi_i} I_{si} \right) \frac{x_i Q_i \pi_i}{\Sigma x_i^2 Q_i I_{si}},$$

$M(t_{gc})$ may be estimated by

$$m_k(t_{gc}) = \left(\frac{\nu}{\nu(s)}\right)^2 m_k(t_g),\ k = 1, 2.$$

Remark.

A major problem with Poisson sampling scheme is that $\nu(s)$ varies across $0, 1, \ldots, N$ and in using the estimator t_H, this variability is likely to create excesses in the latter's variance. The use of t_{RH} is an attempt at a possible check on this excess. But if $\nu(s)$ happens to be zero, then t_H and t_{RH} are not usable. Brewer, Early and Joyce (1972) and Brewer, Early and Hanif (1984) considered alternative estimators allowing a course to follow in case $\nu(s)$ equals zero and they also treated a 'modified Poisson sampling' scheme, in which 'Poisson sampling' is continued till one sample with "$\nu(s) > 0$" is realized throwing away the previous ones with "$\nu(s) = 0$".

For a sampling design admitting a fixed size n for every sample s with a positive selection-probability, Brewer (2000) gives us the identity

$$\begin{aligned} V(t_H) &= \sum_i \pi_i(1-\pi_i)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2 \\ &\quad + \sum\sum_{i \neq j} (\pi_{ij} - \pi_i\pi_j)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)\left(\frac{y_j}{\pi_j} - \frac{Y}{n}\right) \qquad (7.1) \\ &= \sum_i y_i^2\left(\frac{1-\pi_i}{\pi_i}\right) + \sum\sum_{i \neq j} y_i y_j(\pi_{ij} - \pi_i\pi_j)/\pi_i\pi_j \end{aligned}$$

To avoid π_{ij} in estimating this $V(t_H)$, Brewer (2000) approximates π_{ij} by

$$\tilde{\pi}_{ij} = \pi_i\pi_j\left(\frac{c_i + c_j}{2}\right),$$

with suitably chosen positive numbers c_i in $(0, 1)$ and using $\tilde{\pi}_{ij}$ in lieu of π_{ij} in $V(t_H)$ he approximates the latter by

$$\tilde{V}(t_H) = \Sigma\pi_i(1 - c_i\pi_i)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2 \qquad (7.2)$$

and recommends a biased estimator for $V(t_H)$ given by

$$v_B(t_H) = \Sigma(\frac{1}{c_i} - \pi_i)\left(\frac{y_i}{\pi_i} - \frac{t_H}{n}\right)^2 I_{si}$$

as derived from (7.2) by multiplying by $\frac{I_{si}}{c_i\pi_i}$.

Following Brewer (2000), in case a sampling design admits varying number of distinct units $\nu(s)$ with an expected value $\nu = \Sigma\pi_i$, we may present the following alternative identity

$$\begin{aligned} V(t_H) &= \Sigma\pi_i(1-\pi_i)\left(\frac{y_i}{\pi_i} - \frac{Y}{\nu}\right)^2 + \underset{i\neq j}{\Sigma\Sigma}(\pi_{ij} - \pi_i\pi_j)\left(\frac{y_i}{\pi_i} - \frac{Y}{\nu}\right)\left(\frac{y_j}{\pi_j} - \frac{Y}{\nu}\right) \\ &\quad -Y^2\left(1 - \frac{1}{\nu} + \frac{1}{\nu^2}\underset{i\neq j}{\Sigma\Sigma}\pi_{ij}\right) + \frac{2Y}{\nu}\Sigma\frac{y_i}{\pi_i}\left(\Sigma_{j\neq i}\pi_{ij}\right) \qquad (7.3) \\ &= \Sigma y_i^2\frac{1-\pi_i}{\pi_i} + \underset{i\neq j}{\Sigma\Sigma}y_iy_j\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_i\pi_j}. \end{aligned}$$

Incidentally we may note that for Poisson's sampling scheme this reduces to

$$V(t_H)\Big|_{\text{Poisson}} = \Sigma y_i^2\frac{1-\pi_i}{\pi_i}$$

as it should be because $\pi_{ij} = \pi_i\pi_j$ for Poisson's scheme.

With Brewer's (2000) choice of π_{ij} as

$$\tilde{\pi}_{ij} = \pi_i\pi_j\left(\frac{c_i + c_j}{2}\right),$$

(7.3) reduces to

$$V^*(t_H) = \Sigma y_i^2\frac{1-\pi_i}{\pi_i} + \Sigma\pi_i^2(c_i - 1)\left(\frac{y_i}{\pi_i} - \frac{Y}{\nu}\right)^2 \qquad (7.4)$$

Following Brewer (2000), we may employ a biased estimator for $V(t_H)$ as

$$v_{MB}(t_H) = \Sigma y_i^2\frac{1-\pi_i}{\pi_i}\frac{I_{si}}{\pi_i} + \Sigma\pi_i\left(1 - \frac{1}{c_i}\right)\left(\frac{y_i}{\pi_i} - \frac{t_H}{\nu}\right)^2 I_{si} \qquad (7.5)$$

as derived from (7.4).

7.2 Numerical comparison of efficacies of a few alternative estimators of a population total

We consider from Särndal, Swensson and Wretman's (SSW, 1992) text, pp. 660-661, a collection of $N = 50$ clusters of municipalities for which the values $y_i(i = 1, \ldots, N)$ are taken as the populations in 1985 for the respective clusters and x_i's as the 1975 population figures to be used in regression modelling with known totals $Y = 8339$ and $X = 8182$.

We consider the first $N_1 = 23$ clusters as the first stratum and the last $N_2 = 27$ clusters as the second stratum. From these two strata respectively we draw samples of sizes $n_1 = 9$ and $n_2 = 8$ by alternative schemes using the size-measures as z_i^2's where z_i's are the numbers of municipalities in the respective clusters, the total sample-size being $n = n_1 + n_2 = 17$.

We consider $R = 1000$ replicates of the stratified samples drawn by alternative schemes mentioned below. For just one of the replicates we present in tables below the estimates of Y along with the estimated standard errors (SE) which are the positive square roots of the estimated variances or estimated MSE's. Also, treating the pivotal $\delta = \frac{t-\theta}{\sqrt{v}}$, where t is an estimator for a parameter θ with v as the estimator of its variance or MSE, as a standard normal deviate, we treat

$$(t - 1.96\sqrt{v}, t + 1.96\sqrt{v}) = (t \pm 1.96\sqrt{v})$$

in brief, as a 95 percent confidence interval (CI) for θ. We present in tables below the values of

(I) ACP = the actual coverage percentage, which is the percent of replicates for which a CI covers Y - the closer it is to 95 the better the CI,

(II) ACV = the average coefficient of variation, which is the average over the replicates of the values of $100\frac{\sqrt{v}}{t}$ - the less it is the less is the width of CI and the more accurate the point estimator t for θ and finally

(III) AL = average length of the confidence interval over the replicates. We consider the following alternatives :

(i) Poisson sampling scheme with $\pi_i = \frac{nz_i^2}{\Sigma z_i^2}$, with n as the pre-assigned sample-size intended; for this scheme we consider

$$(t_H, v(t_H)), (t_{RH}, m_1), (t_{RH}, m_2), (t_{RH}, m_{BRH}), (t_g, m_k(t_g)),\ k = 1, 2,$$

$(t_{gc}, m_k(t_{gc})), k = 1, 2$ as alternative choices of (t, v) when $\theta = Y$,

(ii) and for a "modifed Poisson sampling" scheme (MPS), $(t_H, v'(t_H))$, (t_{RH}, m_1'), (t_{RH}, m_2'), $(t_g, m_k'(t_g)), k = 1, 2$ are the alternative choices of (t, v) when $\theta = Y$, namely $v'(t_H)$ as discussed earlier,and

$$m_k'(t_g) = \Sigma(1 - \pi_i)\frac{a_{ki}^2 e_i^2}{\pi_i^2} I_{si} - \frac{P_0}{1 - P_0}[(\Sigma\frac{a_{ki}e_i}{\pi_i} I_{si})^2 - \Sigma\frac{a_{ki}^2 e_i^2}{\pi_i^2} I_{si}]$$

where

$$a_{1i} = 1, a_{2i} = g_{si} = 1 + \left(X - \Sigma\frac{x_i}{\pi_i} I_{si}\right)\frac{x_i Q_i \pi_i}{\Sigma x_i^2 Q_i I_{si}},$$

and m_k' will be obtained from $m_k'(t_g)$ for which $x_i = \pi_i, X = \Sigma x_i = \Sigma\pi_i = \nu, Q_i = \frac{1}{\pi_i x_i} = \frac{1}{\pi_i^2}$.and $g_{si} = \frac{\nu}{\nu(s)}$.
With Brewer's (2000) choice of π_{ij} as

$$\tilde{\pi}_{ij} = \pi_i\pi_j\left(\frac{c_i + c_j}{2}\right),$$

for the above MPS scheme, $(t_H, v_{MB}(t_H)), (t_{RH}, m_{1MB}'), (t_{RH}, m_{2MB}')$, $(t_g, m_{kMB}'(t_g)), (t_{gc}, m_{kMB}'(t_{gc})), k = 1, 2$ are the further alternative choices of (t, v). The above $v_{MB}(t_H)$ is described earlier. For $m_{1MB}'(t_g)$ and $m_{2MB}'(t_g)$, y_i in $v_{MB}(t_H)$ is to be simply replaced by e_i and $e_i g_{si}$ respectively. For t_{RH}, m_{1MB}' and m_{2MB}' are obtained from $m_{1MB}'(t_g)$ and $m_{2MB}'(t_g)$ with $x_i = \pi_i$ and $Q_i = \frac{1}{\pi_i x_i}$ respectively. Also

$$m_{kMB}'(t_{gc}) = \left(\frac{\nu}{\nu(s)}\right)^2 m_{kMB}'(t_g),\ k = 1, 2.$$

(iii) Rao-Hartley-Cochran (RHC, 1962) sampling scheme. Here to draw a sample of size n from a population of size N we first divide the population at random into n groups taking in the ith group $N_i = \left[\frac{N}{n}\right]$ or $\left[\frac{N}{n}\right] + 1$ units subject to $\Sigma_n N_i = N$, writing Σ_n as sum over the n groups. From the ith group so formed one unit is chosen with a probability proportional to z_i^2 for the units in the ith group and this is repeated independently over the n groups. Writing (y_i, p_i, r_i) for the y-values, normed size-measure-values and the summed p_i-values over the N_i units in the ith group, RHC's unbiased estimator for Y is

$$t_{RHC} = \Sigma_n y_i \frac{r_i}{p_i}.$$

The RHC-unbiased variance-estimator (RHC, 1962) is

$$v_{RHC} = \left(\frac{\Sigma_n N_i^2 - N}{N^2 - \Sigma_n N_i^2}\right) \Sigma_n \Sigma_n r_i r_{i'} \left(\frac{y_i}{p_i} - \frac{y_{i'}}{p_{i'}}\right)^2,$$

writing $\Sigma_n \Sigma_n$ as the sum over the pairs of the groups already formed with no duplication.

A greg version of t_{RHC} is

$$t_{gR} = \Sigma_n y_i \frac{r_i}{p_i} + b_R(X - \Sigma_n x_i \frac{r_i}{p_i}) = \Sigma_n y_i \frac{r_i}{p_i} h_{si},$$

$$b_R = \frac{\Sigma_n y_i x_i R_i}{\Sigma_n x_i^2 R_i}, h_{si} = 1 + \left(X - \Sigma_n x_i \frac{r_i}{p_i}\right) \frac{x_i R_i \frac{p_i}{r_i}}{\Sigma_n x_i^2 R_i}$$

with $R_i (> 0)$ to be suitably chosen, as for example,

$$R_i = \frac{r_i}{p_i x_i}, \quad \frac{1 - \frac{p_i}{r_i}}{\frac{p_i}{r_i} x_i} = \frac{r_i - p_i}{p_i x_i}$$

etc.

MSE estimators of t_{gR} are

$$v_k = \left(\frac{\Sigma_n N_i^2 - N}{N^2 - \Sigma_n N_i^2}\right) \Sigma_n \Sigma_n r_i r_{i'} \left(\frac{b_{ki} e_{ri}}{p_i} - \frac{b_{k_{i'}} e_{r_{i'}}}{p_{i'}}\right)^2, \quad k = 1, 2,$$

where $b_{1i} = 1, b_{2i} = h_{si}, \ e_{ri} = y_i - b_R x_i,$

(iv) Hartley and Rao's (HR, 1962) sampling scheme. Here the units in a population are first randomly permuted. Then, from the permuted vector of labeled units a PPSCSS sample (as described in Chapter 3) is chosen in n draws, with n as the intended sample size, provided

$$np_i < 1 \ \forall i \in U, \ p_i = \frac{z_i^2}{\Sigma z_i^2}.$$

For this scheme values of π_{ij} as given approximately by HR will be used which are as follows :

$$\begin{aligned}
\pi_{ij}(HR) \quad = \quad & \left(\frac{n-1}{n}\right) \pi_i \pi_j + \left(\frac{n-1}{n^2}\right) \left(\pi_i^2 \pi_j + \pi_i \pi_j^2\right) \\
& - \left(\frac{n-1}{n^3}\right) \pi_i \pi_j \Sigma \pi_i^2 + \frac{2(n-1)}{n^3} \left(\pi_i^3 \pi_j + \pi_i \pi_j^3 + \pi_i^2 \pi_j^2\right) \\
& - \frac{3(n-1)}{n^4} \left(\pi_i^2 \pi_j + \pi_i \pi_j^2\right) \Sigma \pi_i^2 + \frac{3(n-1)}{n^5} \pi_i \pi_j (\Sigma \pi_i^2)^2 \\
& - \frac{2(n-1)}{n^4} \pi_i \pi_j \Sigma \pi_i^3
\end{aligned}$$

and of course $\pi_i = np_i$.

For the t_H based on the HR scheme, $V(t_H)$ will be estimated by

$$v_{YG}(t_H) = \mathop{\Sigma\Sigma}_{i<j} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\right) \left(\frac{y_i}{\pi_i} - \frac{y_i}{\pi_j}\right)^2 I_{sij}$$

with $\pi_{ij} = \pi_{ij}$ (HR) and $\pi_i = np_i$.

We shall also use, following Brewer (2000), for the HR scheme, the approximate values of π_{ij}'s as

$$\pi_{ij}(B) = \pi_i \pi_j \left(\frac{c_i + c_j}{2}\right), \ c_i = \frac{n-1}{n - \pi_i},$$

and $V(t_H)$ will be estimated by

$$v_B(t_H) = \Sigma(\frac{1}{c_i} - \pi_i) \left(\frac{y_i}{\pi_i} - \frac{t_H}{n}\right)^2 I_{si}$$

Also, for the HR scheme we shall use the v_{YG} to estimate $V(t_H)$ by using the following two alternative formulae for π_{ij} as given by Stehman and Overton (SO, 1994) namely

$$\pi_{ij}(So1) = \frac{(n-1)\pi_i\pi_j}{n-(\pi_i+\pi_j)/2}$$

and

$$\pi_{ij}(So2) = \frac{(n-1)\pi_i\pi_j}{n-\pi_i-\pi_j+\frac{1}{n}\Sigma\pi_i^2};$$

of course $\pi_i = np_i(< 1)$.

Also, we shall use t_g based on HR scheme and employ $m_k(t_g), k = 1,2$, using the formulae as $\pi_{ij}(HR), \pi_{ij}(B), \pi_{ij}(So1)$ and $\pi_{ij}(So2)$. The corresponding estimates for variance of t_H will be denoted respectively by $v_{HR}()$, $m_{HRk}()$, $v_B(), m_{Bk}()$, $v_{So1}()$, $v_{So2}()$, $m_{(So1)k}()$, $m_{(So2)k}()$ for variance and MSE estimators respectively for t_H and t_g using the corresponding formulae for π_{ij} as above.

(v) Also for the modified sampling scheme of Seth as discussed in Chapter 1 we use $(t_H, v_{MB}(t_H))$ and also $(t_H, v_{MS}(t_H))$ with $v_{MS}(t_H)$ given by

$$v_{MS}(t_H) = \sum_{i<j}\sum\frac{I_{sij}}{\pi_{ij}}(\pi_i\pi_j-\pi_{ij})(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j})^2+\Sigma\frac{I_{si}}{\pi_i}\frac{y_i^2}{\pi_i}\beta_i$$

where

$$\beta_i = (1+\frac{1}{\pi_i}\sum_{j\neq i}\pi_{ij}-\Sigma\pi_i), i\epsilon U$$

for which π_i, π_{ij}'s are as given in Chapter 1; the expression for $v_{MB}(t_H)$ is given in equation (7.5) in this Chapter in which the π_i should be as in modified Seth scheme.

Table 1(a)

Performances of alternative procedures based on Poisson sampling scheme

(t, v)	t (for one replicate)	S.E. $= \sqrt{v}$ (for one replicate)	ACP	ACV	AL
$t_H, v(t_H)$	9885.56	2912.16	86.9	22.20	7876.59
t_{RH}, m_1	8860.16	2178.56	84.10	12.47	5566.04
t_{RH}, m_2		2183.50	85.10	12.25	5521.22
t_{RH}, m_{BRH}		2023.60	83.60	13.36	4921.53
with $Q_i = \frac{1}{\pi_i x_i}$					
$t_g, m_1(t_g)$	8322.40	127.89	84.30	1.80	508.10
$t_g, m_2(t_g)$		100.39	87.70	1.20	491.10
$t_{gc}, m_1(t_{gc})$	8860.16	1936.49	84.80	13.40	5888.70
$t_{gc}, m_2(t_{gc})$		1741.64	85.80	13.94	6003.20
with $Q_i = \frac{1-\pi_i}{\pi_i x_i}$					
$t_g, m_1(t_g)$	8329.20	103.22	71.30	.91	297.93
$t_g, m_2(t_g)$		179.89	74.90	.85	280.30

Table 1(b)

Comparative performances of alternative procedures based on Poisson sampling scheme

(t, v)	t (for one replicate)	S.E. $= \sqrt{v}$ (for one replicate)	ACP	ACV	AL
$t_H, v'(t_H)$	9885.56	2911.91	86.9	22.20	7875.56
t_{RH}, m'_1	8860.16	2243.30	85.40	13.26	5892.94
t_{RH}, m'_2		2158.84	85.60	13.05	5815.06
t_{RH}, m'_{1MB}		2236.29	89.30	12.32	4938.10
t_{RH}, m'_{2MB}		2158.36	86.10	13.01	4900.36
with $Q_i = \frac{1}{\pi_i x_i}$					
$t_g, m'_1(t_g)$	8322.40	188.60	89.60	1.32	511.30
$t_g, m'_2(t_g)$		185.44	91.30	1.63	522.63
$t_g, m'_{1MB}(t_g)$		179.38	90.10	1.29	493.29
$t_g, m'_{2MB}(t_g)$		180.36	89.30	1.9	485.06
$t_{gc}, m'_1(t_{gc})$	8860.16	1994.04	86.50	14.41	5260.41
$t_{gc}, m'_2(t_{gc})$		1918.97	87.00	14.56	5312.43
$t_{gc}, m'_{1MB}(t_{gc})$		1936.59	88.30	14.90	5006.30
$t_{gc}, m'_{2MB}(t_{gc})$		1921.06	87.10	14.52	5323.90
with $Q_i = \frac{1-\pi_i}{\pi_i x_i}$					
$t_g, m'_1(t_g)$	8329.20	88.55	79.70	.95	312.39
$t_g, m'_2(t_g)$		110.36	83.40	.98	323.02
t_{RHC}, v_{RHC}	10185.10	2275.48	93.90	18.02	4554.01
with $Q_i = \frac{r_i}{p_i x_i}$					
t_{gR}, v_1	8228.09	140.49	88.50	1.98	521.97
t_{gR}, v_2		638.96	98.10	3.88	1273.66
with $Q_i = \frac{r_i - p_i}{p_i x_i}$					
t_{gR}, v_1	8234.42	134.62	82.00	.98	320.31
t_{gR}, v_2		524.62	98.30	4.35	1425.13

Table 2
Comparative performances of alternative procedures using HT and greg estimators based on Hartley-Rao scheme and modified Seth's scheme

(t, v)	t	SE	ACP	ACV	AL
			(Harlety-Rao Scheme)		
$t_H, v_{HR}(t_H)$	9238.17	1700.39	96.70	17.47	5873.99
$t_H, v_B(t_H)$		2037.00	93.50	14.69	4980.48
$t_H, v_{S01}(t_H)$		1608.00	91.70	13.91	4729.24
$t_H, v_{S02}(t_H)$		1604.81	91.60	13.69	4654.44
with $Q_i = \frac{1}{\pi_i x_i}$					
$t_g, m_{HR1}(t_g)$	8272.50	124.12	86.00	1.07	351.13
$t_g, m_{HR2}(t_g)$		98.41	93.10	1.13	370.71
$t_g, m_{B1}(t_g)$		117.45	83.60	.98	320.21
$t_g, m_{B2}(t_g)$		93.11	88.50	1.00	328.43
$t_g, m_{(S01)1}(t_g)$		117.38	84.80	1.02	332.85
$t_g, m_{(S01)2}(t_g)$		93.18	90.70	1.06	346.81
$t_g, m_{(S02)1}(t_g)$		117.07	84.50	.99	324.69
$t_g, m_{(S02)2}(t_g)$		92.98	90.50	1.03	338.88
with $Q_i = \frac{1-\pi_i}{\pi_i x_i}$					
$t_g, m_{HR1}(t_g)$	8272.68	123.49	85.60	1.07	350.36
$t_g, m_{HR2}(t_g)$		94.90	92.40	1.12	367.79
$t_g, m_{B1}(t_g)$		116.86	82.60	.97	317.23
$t_g, m_{B2}(t_g)$		89.73	88.00	.99	325.51
$t_g, m_{(S01)1}(t_g)$		116.78	84.40	1.02	332.20
$t_g, m_{(S01)2}(t_g)$		89.86	89.50	1.04	342.28
$t_g, m_{(S02)1}(t_g)$		116.48	84.10	.99	323.94
$t_g, m_{(S02)2}(t_g)$		89.71	89.80	1.02	336.02
$t_H, v_{MS}(t_H)$	8736.21	7488.54	97.00	32.02	9665.10
$t_H, v_{MB}(t_H)$		2297.58	96.30	23.76	9415.93

Comments :

Särndal's (1996) work motivated us to investigate whether there is any advantage in using Poisson scheme coupled with Horvitz and Thompson's estimator to avoid computing π_{ij}'s with complicated formulae rather than employing the simple alternative procedure given by Rao, Hartley and Cochran (RHC, 1962) for sample selection and estimation. The Table 1(a), Table 1(b) show that neither t_H nor t_{RH} really beats t_{RHC}. Schabenberger and Gregoire (1994) also noted the good performance of t_{RHC}. However t_g competes quite well and closely with t_{gR} with various choices of Q_i, R_i and MSE estimators. The far-fetched version of 'greg' approach starting with another greg rather than with an unbiased estimator does not yield any advantage as is evident from the numerical values concerning t_{gc} from the Tables 1(a), 1(b) above. The modified Poisson sampling scheme fares competitively with the original Poisson scheme with slightly less efficacies. The estimators given by Brewer (2000) and Brewer and Gregoire (2000) now included in this revision for comparison fare quite well.

Also, as expected, greg estimator fares much better than the HT estimator for every sampling scheme. There is not much to choose between Hartley Rao's, Brewer's and Stehman and Overton's schemes but modified Seth's scheme is inferior to them. Between the two MSE-estimators of the greg estimator the one that uses a 'mulplier for the residual' fares better than the other which uses the residual term alone in each case.

Bibliography

Ajgaonkar, A. (1967). "Unbiased estimator of the variance of the Narain, Horvitz and Thompson estimator", *Sankhyá* A., **29**, 55-60.

Arnab, R. (1996). "Randomized response trials: A unified approach for qualitative data", *Comm. Stat.-Theo. Meth.*, **25**(6), 1173-1183.

———— (2000). "Analysis of randomized response survey data". *In 'Perspectives in Statistical sciences' ed. Basu, S.K., Ghosh, J.K., Sen, P.K. and Sinha, Bimal K.*, Oxford Univ. Press, New Delhi, 18-26.

Basu, D. (1958) "On sampling with and without replacemenet", *Sankhyá*, **20**, 287-294.

———— (1969). "Role of the sufficiency and likelihood principles in sample survey theory", *Sankhyá*, A, **31**, 441-454.

———— (1971). "An essay on the logical foundations of survey sampling", Part I, *Foundations of Statistical Inference* , 203-242.

Basu, D. & Ghosh, J.K. (1967). "Sufficient statistics in sampling from a finite universe", *Bull. Int. Stat. Inst.*

Bolfarine, H. & Zacks, S. (1992). *"Prediction theory for finite populations"*, Springer Verlag, N.Y.

Brewer, K.R.W. (1963). "A model of systematic sampling with unequal probabilities", *Aust. Jour. Stat.*, **5**, 5-13.

————(1979). "A class of robust sampling designs for large scale surveys ", *Jour. Amer. Stat. Assoc.*, **74**, 911-915.

———————(1999). "Cosmetic calibration with unequal probability sampling", *Survey Meth.*, **25** (2), 205-212.

——————— (2000). "Deriving and estimating an approximate variance for the Horvitz-Thompson estimator using only first order inclusion-probabilities", *Contributed to second international conference on Establishment Surveys*, Buffalo, N.Y., June, 17-21, 2000.

Brewer, K.R.W., Early, L.J. & Hanif, M. (1984). "Poisson, Modified Poisson and Collocated sampling", *Jour. Stat. Plan. Inf.*, **10**, 15-30.

Brewer, K.R.W., Early, L.J. & Joyce, S.F. (1972). "Selecting several samples from a single population", *Aust. Jour. Stat.*, **14**, 231-239.

Brewer, K.R.W. & Gregoire, T.G. (2000). "Estimators for use with Poisson Sampling and related selection procedures", *Invited paper in Second International conference on Establishment Surveys*, Buffalo, N.Y.. June 17-21, 2000.

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

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

——————— (1999). "Towards a unified theory of randomized response surveys for dichotomous finite populations", Tech. Rep. ASD/99/36, ISI, Calcutta.

——————— (2000a). "Network and Adaptive Sampling with unequal probabilities", *Cal. Stat. Assoc. Bull.*, **50**, 237-253.

——————— (2000b). "Using a randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population", *Jour. Stat. Plan. Inf.*, **94** (1), 37-42.

——————— (2001). "Estimating sensitive proportions from unequal probability samples using randomized responses", *Pak. Jour. Stat.*, **17**(3), 259-270.

Chaudhuri, A., Adhikary, Arun. K. & Dihidar, S. (2000). "Mean square error estimation in multi-stage sampling", *Metrika*, **52**(2), 115-131.

Chaudhuri, A., Bose, M. & Ghosh, J.K. (2002). "An application of adaptive sampling to estimate highly localized population segments", *To appear in Jour. Stat. Plan. Inf.*

Chaudhuri, A. & Maiti, T. (1995). "On the regression adjustments to Rao-Hartley-Cochran estimator", *Jour. Stat. Res.*, **29**, 71-78.

Chaudhuri, A. & Mukerjee, R. (1988). *"Randomized response: Theory and Techniques"*, Marcel Dekker, Inc. N.Y.

Chaudhuri, A. & Pal, S. (2002a). "On certain alternative mean square error estimators in complex survey sampling", *Jour. Stat. Plan. Inf.*, **104**(2), 363-375.

——————— (2002b). "Estimating proportions from unequal probability samples using randomized responses by Warner's and other devices", *Jour. Ind. Soc. Agri. Stat.*, **55**(2), 174-183.

——————— (2003a). "On a version of cluster sampling and its practical use", *Jour. Stat. Plan. Inf.*,**113**(1), 25-34.

———————(2003b) "Systematic sampling: 'Fixed' versus 'Random' sampling interval", *Pak. Jour. Stat.*, 2003, **19**(2), 259-271.

Chaudhuri, A. & Stenger, H. (1992). *"Survey Sampling"*, Marcel Dekker, Inc. N.Y.

Chaudhuri, A. & Vos, J.W.E. (1988). *"Unified Theory and Strategies of Survey Sampling"*, North-Holland, Amsterdam.

Cochran, W.G. (1939). "The use of analysis of variance in enumerating by sampling,", *Jour. Amer. Stat. Assoc.*, **34**, 492-510.

———— (1946). "Relative accuracy of systematic and stratified random samples for a certain class of populations", *Ann. Math. Stat.*, **17**, 164-177.

———— (1977). *"Sampling Techniques"*, John Wiley and Sons, N.Y.

Das, M.N. (1982). *"Systematic sampling without drawback"*, Tech. Rep. 8206, ISI, Delhi.

Deusen, P.C. Van. (1987). *"3-p sampling and design versus model-based estimates"*, *Canadian Jour. Forest Research*, **17**(2), 115-117.

Deville, Jean-Claude (1999). "Variance estimation for complex statistics and estimators : linearization and residual techniques", *Survey Meth.*, **25**(2), 193-203.

Fay, R.E. & Herriot, R.A. (1979). "Estimates of income for small places: an application of James Stein procedures to census data", *Jour. Amer. Stat. Assoc.*, **74**, 269-277.

Franklin, L.A. (1989a). "Randomized response sampling from dichotomous populations with continuous randomization", *Survey Meth.*, **15**(2), 225-235.

———— (1989b). "A comparison of estimators for randomized response sampling with continuous distributions from a dichotomous population", *Comm. Stat.-Theo. Meth.*, **18**(2), 489-505.

Fuller, W.A. (1987). *"Measurement error models"*, Wiley, N.Y.

Fuller, W.A.& Isaki, C.T. (1981). "Survey design under super-population models,", *Current Topics in Survey Sampling*, Acad. Press, New York , 199-226.

Godambe, V.P. (1955). "A unified theory of sampling from finite populations", *Jour. Roy. Stat. Soc.* B, **17**, 269-278.

———— (1960)."An admissible estimate for any sampling design", *Sankhyá*, **22**, 285-288.

———— (1998)."A new look at confidence intervals in survey sampling", Working paper 1998-2002,Dept. stat. and Actuarial Sc., Univ. Waterloo, Canada.

Godambe, V. P., Joshi, V.M. (1965). "Admissibility and Bayes estimation in sampling finite populations", I, Ann. Math. Stat., **36**, 1707-1722.

Grosenbaugh, L.R. (1965). "Three pee sampling theory and program THRP for computer generation of selection criteria", *USDA Forest Service Research Paper, PSW*, **21**, 53.

Hájek, J. (1958). "Some contributions to the theory of probability sampling", *Bull. Int. Stat. Inst.*, **36**(3), 127-134.

———— (1964). "Asymptotic theory of rejective sampling with varying probability from a single population", *Aust. Jour. Stat.*, **14**, 231-239.

———— (1981), "Sampling from a finite population", *Marcel Dekker Inc. N.Y.*

Hanurav, T.V. (1966). "Some aspects of unified sampling theory", *Sankhyá*, **28**, 175-204.

——————— (1968). "Hyperadmissibility and optimum estimators for sampling finite populations", *Ann. Math. Stat.*, **39**, 621-642.

Hartley, H.O. & Rao, J.N.K. (1962). "Sampling with unequal probabilities and without replacement", *Ann. Math. Stat.* **33**, 350-374.

Hege, V.S. (1965). "Sampling designs which admit uniformly minimum variance unbiased estimators", *Cal. Stat. Assoc. Bull.*, **14**, 160-162.

Horvitz., D.G. & Thompson, D.J. (1952). "A generalization of sampling without replacement from a finite universe", *Jour. Amer. Stat. Assoc.*, **47**, 663-685.

Isaki, C.T.& Fuller, W.A. (1982). "Survey design under the regression superpopulation model",*Jour. Amer. Stat. Assoc.*, **77**, 89-96.

Kunte, Sudhakar (1978). "A note on circular systematic sampling design", *Sankhyā*, C, **40**, 72-73.

Kuk, A.Y.C (1990). "Asking sensitive questions indirectly", *Biometrika*, **77**, 436-438.

Murthy, M.N. (1957). "Ordered and unordered estimators in sampling without replaccement", *Sankhyá*, **18**, 379-390.

——————— (1967). "Sampling theory and methods", *Statistical Publishing Society, Calcutta.*

Mangat, N.S. (1992). "Two stage randomized response sampling procedure using unrelated questions", *Jour. Ind. Soc. Agri. Stat.*, **44**(1), 87-88.

Mangat, N.S. & Singh, R. (1990). "An alternative randomized response procedure", *Biometrika*, **77**, 439-442.

Mangat, N.S. & Singh, R. & Singh, S. (1992). "An improved unrelated question randomized response strategies", *Cal. Stat. Assoc. Bull.*, **42**, 277-281.

Narain, R.D. (1951). "On sampling with varying probabilities without replacement", *Jour. Ind. Soc. Agri. Stat.*, **3**, 169-175.

Ogus, J.K. & Clark, D.F. (1971). "A report on methodology ", *Technical Report No. 24, U.S.Bureau of Census, Washington D.C.*, **77**, 436-438.

Prasad, N.G.N. & Rao, J.N.K. (1990). "The estimation of the mean square error of small area estimates", *Jour. Amer. Stat. Assoc.*, **85**, 163-171.

Raj. Des. (1954). "Ratio estimation in sampling with equal and unequal probabilities", *Jour. Ind. Soc. Agri. Stat.*, **6**, 127-138.

———— (1968). "Sampling Theory", Mc-graw Hill, N.Y.

Rao, J.N.K. (1975). "Unbiased variance estimation for multi-stage designs", *Sankhyá*, C, **37**, 133-139.

———— (1979). "On deriving mean square errors and other non-negative unbiased estimators in finite population sampling", *Jour. Ind. Stat. Assoc.* **17**, 125-136.

Rao, J.N.K., Hartley, H.O. & Cochran, W.G. (1962). "On a simple procedure of unequal probability sampling without replacement", *Jour. Roy. Stat. Soc.* B, **24**, 482-491.

Rao, J.N.K. & Wu, C.F.J. (1987). "Methods for standard errors and confidence intervals from sample survey data: Some recent work.", *Bull. Int. Stat. Ins.*, 1-17.

———— (1988). "Resampling inference with complex survey data", *Jour. Amer. Stat. Assoc.*, **83**, 231-241.

Rao, J.N.K. & Vijayan, K. (1977). "On estimating the variance in sampling with probability proportional to aggregate size", *Jour. Amer. Stat. Assoc.*, **72**, 579-584.

Ray, S. & Das, M.N. (1997). "Circular systematic sampling with drawback", *Jour. Ind. Soc. Agri. Stat.* **50**(1), 70-74.

Royall, R.M. (1970). "On finite population sampling theory under certain linear regression models", *Biometrika*, **57**, 377-387.

Salehi, M.N. & Seber, G.A.F. (1997). "Adaptive cluster sampling with networks selected without replacement", *Biometrika.*, **84**, 209-219.

―――――――――――― (2002). "Unbiased estimators for restricted adaptive cluster sampling", *Aust. N. Z. Jour. Stat.*, **44**(1), 63-74.

Särndal, C.E. (1980). "On π-inverse weighting versus best linear weighting in probability sampling", *Biometrika*, **67**(3), 639-650.

―――――――― (1981). "Frameworks for inference in survey sampling with applications to small area estimation and adjustment for non-response", *Bull. Int.Stat. Ins.*, **29**, 494-513.

―――――――― (1982). "Implications of survey design for generalized regression estimation of linear functions", *Jour. Stat. Plan. Inf.*, **7**, 155-170.

―――――――― (1996). "Efficient estimators with simple variance in unequal probability sampling", *Jour. Amer. Stat. Assoc.*, **91**, 1289-1300.

Särndal, C.E., Swensson, B.E. & Wretman, J.H. (1992). *"Model Assisted Survey Sampling"*, Springer Verlag, N.Y.

Schabenberger, O. & Gregoire, T.G. (1994). "Competitors to genuine πPS sample designs: A comparison.", *Survey Meth.*, **20** (2), 185-192.

Sen, A.R. (1953). "On the estimator of the variance in sampling with varying probabilities", *Jour. Ind. Soc. Agri. Stat.*, **5**(2) 119-127.

Seth, G.R. (1966). "On estimators of variance of estimate of population total in varying probabilities", *Jour. Ind. Soc. Agri. Stat.*, **18**(2), 52-56.

Singh, S. & Singh, R. (1992). "Improved Franklin's model for randomized response sampling", *Jour. Ind. Stat. Assoc.*, **30**, 109-122.

———————— (1993). "Generalized Franklin's model for randomized response sampling", *Comm. Stat.-Theo. Meth.*, **22**, 741-755.

Singh, S. & Joarder Anwar H. (1997). "Unknown repeated trials in randomized response sampling", *Jour. Ind. Soc. Agri. Stat.*, **50**, 70-74.

Stehman, S. & Overton, W.S. (1994)."Comparison of variance estimators of the Horvitz Thompson estimator for randomized variable probability systematic sampling ", *Jour. Amer. Stat. Assoc.*, **89**, 30-43.

Stenger, H. (1977). "Sequential sampling from finite populations", *Sankhyá*, C, **39**, 10-20.

Thompson, S.K. (1990). *"Adaptive cluster sampling",Jour. Amer. Stat. Assoc.*, **85**, 1050-1059.

———————— (1992). *"Sampling"*, John Wiley & Sons, N.Y.

Thompson, S.K. & Seber, G.A.F. (1996). "*Adaptive Sampling*", John Wiley & Sons, N.Y.

Vijayan, K. (1975). "On estimating the variance in unequal probability sampling", *Jour. Amer. Stat. Assoc.*, **70**, 713-716.

Warner, S.L. (1965). "RR : a survey technique for eliminating evasive answer bias", *Jour. Amer. Stat. Assoc.*, **60**, 63-69.

Woodruff, R.S. (1952). "Confidence intervals for medians and other position measures.", *Jour. Amer. Stat. Assoc.*, **47**, 635-646.

Yates, F. & Grundy, P.M. (1953). "Selection without replacement from within strata with probability proportional to size", *Jour Roy. Stat. Soc.* B, **(15)**, 253-261.