International Statistical Review (1987), **55**, 3, pp. 279–294. Printed in Great Britain

A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse

Carl-Erik Särndal[1] and Bengt Swensson[2]

[1]Département de mathématiques et de statistique, Université de Montréal, Case postale 6128, succursâle "A", Montréal, P.Q. H3C 3J7, Canada. [2]Institutionen för dataanalys, Högskolan i Örebro, Box 923, S-701 30 Örebro, Sweden.

Summary

This paper is meant as a contribution to estimation (a) in the case of two-phase sampling, (b) in the case of nonresponse. In both cases, there is subsample selection. The difference is that the subsampling obeys a known distribution in two-phase sampling, an unknown distribution ('the response mechanism') in the nonresponse case. General results are given for two-phase sampling, with emphasis on regression estimation and on the problem of variance estimation. The results, with necessary modifications, are applied to the nonresponse situation. In particular, if the assumed response mechanism is false, we find that regression-type estimators give more robust confidence statements than if weighting alone is employed.

Key words: Generalized regression estimation; Nonresponse; Randomization theory; Response mechanism; Two-phase sampling; Variance estimation.

1 Introduction

It has been observed several times in the literature that a basic similarity exists between 'sampling with subsequent nonresponse' and 'two-phase sampling'. In each case a sample, s, is initially drawn, but the 'ultimate sample', r, that is the sample for which we actually measure the variable(s) of study, is only a subset of s. Given this parallel, it is possible, in the nonresponse situation to profit directly from two-phase sampling theory. A systematic effort in this direction is made in this paper.

Two-phase sampling means that a controlled, randomized scheme is used for subsampling. The first part of this paper, §§ 2–6, develops general results for this case: arbitrary sampling designs are admitted in each of the two phases, and a general framework for regression estimation is employed. Nonresponse, on the other hand, means nonrandomized subsampling. The response is assumed to obey a probability distribution; however, this distribution is unknown to the statistician. The second part of the paper, §§ 7–10, develops this case by borrowing and modifying the results obtained in the first part. Our estimators are constructed with special effort to control nonresponse bias so as to avoid serious distortion of the confidence statements. Variance estimation and confidence intervals for nonresponse situations are important aspects of our paper.

2 The two-phase sampling set-up

Since its introduction by Neyman (1938), two-phase sampling (or double sampling) has been part of the standard repertoire of sampling techniques, as witnessed by the fact that

standard texts devote some space to the area; see, for example, Raj (1968, pp. 139–152).

Whereas simple random sampling has often been assumed earlier in one or both of the phases, we admit, more generally, arbitrary designs, and we obtain a general approach to estimating the variance of the two phase estimate. More recent work going in the direction of the generality that we have in mind includes Chaudhuri & Adhikary (1983).

Consider a finite population $U = \{1, \ldots, k, \ldots, N\}$. Let y be the variable of study, and let y_k be the value of y for the kth unit. We seek to estimate the population total $t = \Sigma_U y_k$ from a sample r, obtained through two phases of selection. If $A \subseteq U$ is a set of units, we write $\Sigma_A y_k$ for $\sum_{k \in A} y_k$. We allow a general sampling design in each phase; that is, the inclusion probabilities in each phase are arbitrary. Our notation for the sampling designs will be as follows.

(a) The first-phase sample s ($s \subset U$) of size n_s, not necessarily fixed, is drawn by a design denoted $p_a(.)$, such that $p_a(s)$ is the probability of choosing s. The inclusion probabilities are defined by

$$\pi_{ak} = \sum_{s \ni k} p_a(s), \quad \pi_{akl} = \sum_{s \ni k,l} p_a(s),$$

with $\pi_{akk} = \pi_{ak}$. Set $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$. We assume that $\pi_{ak} > 0$ for all k, and, when it comes to variance estimation, that $\pi_{akl} > 0$ for all $k \neq l$.

(b) Given s, the second-phase sample r ($r \subset s$) of size m_r, not necessarily fixed, is drawn according to a sampling design $p(. \mid s)$, such that $p(r \mid s)$ is the conditional probability of choosing r. The inclusion probabilities given s are defined by

$$\pi_{k|s} = \sum_{r \ni k} p(r \mid s), \quad \pi_{kl|s} = \sum_{r \ni k,l} p(r \mid s),$$

with $\pi_{kk|s} = \pi_{k|s}$. Set $\Delta_{kl|s} = \pi_{kl|s} - \pi_{k|s}\pi_{l|s}$. We assume that, for any s, $\pi_{k|s} > 0$ for all $k \in s$, and that in variance estimation $\pi_{kl|s} > 0$ for all $k \neq l \in s$.

For example, the first-phase sample s may be selected by a two-stage sampling design in which geographical or administrative clusters of individuals are first drawn, followed by subsampling of individuals within drawn clusters. Certain information is gathered for individuals in the sample thus selected. Some of that information, say, concerning age/sex categories, may serve to divide the first-phase sample into strata, whereupon stratified sampling is used in the second phase. The complication that the clusters used at the first stage of the first phase may cut across the strata used for the second-phase sampling causes no conceptual difficulty in our approach.

3 Estimation by π^*-expanded sums

Define, for all $k, l \in s$, and any s,

$$\pi_k^* = \pi_{ak}\pi_{k|s}, \quad \pi_{kl}^* = \pi_{akl}\pi_{kl|s},$$

with $\pi_{kk}^* = \pi_k^*$. Set $\Delta_{kl}^* = \pi_{kl}^* - \pi_k^*\pi_l^*$. We also define expanded y-values and expanded Δ-values by

$$\check{y}_k = y_k/\pi_{ak}, \quad \check{y}_k^\dagger = \check{y}_k/\pi_{k|s} = y_k/\pi_k^*, \quad \check{\Delta}_{akl} = \Delta_{akl}/\pi_{akl}.$$

Note that "$\check{}$" indicates first-phase expansion, while "$\check{}$" and "$\dagger$" together indicate double expansion.

The basic estimator in two phase sampling, which we call the π^*ES *estimator* for π^*-expanded sum, is described in the following result. If $A \subseteq U$ is a set of units, $\Sigma\Sigma_A c_{kl}$ means $\sum_{k \in A} \sum_{l \in A} c_{kl}$.

RESULT 0. *In two-phase sampling, a design unbiased estimator of the population total* $t = \Sigma_U y_k$ *is given by the* π^*ES *estimator,*

$$\hat{t}_{\pi^*} = \Sigma_r \check{y}_k^\dagger = \Sigma_r y_k / \pi_k^*. \tag{3.1}$$

Its design variance is

$$V(\hat{t}_{\pi^*}) = \Sigma\Sigma_U \Delta_{akl} \check{y}_k \check{y}_l + E_a\{\Sigma\Sigma_s \Delta_{kl|s} \check{y}_k^\dagger \check{y}_l^\dagger\}, \tag{3.2}$$

where $E_a(.)$ *denotes expectation with respect to the sampling design in phase one. A design unbiased variance estimator is given by*

$$\hat{V}(\hat{t}_{\pi^*}) = \Sigma\Sigma_r \check{\Delta}_{akl} \check{y}_k \check{y}_l / \pi_{kl|s} + \Sigma\Sigma_r \Delta_{kl|s} \check{y}_k^\dagger \check{y}_l^\dagger / \pi_{kl|s} = \Sigma\Sigma_r \Delta_{kl}^* \check{y}_k^\dagger \check{y}_l^\dagger / \pi_{kl}^*. \tag{3.3}$$

Note that the second component of (3.2) must be left in the form of an expected value, since the $\Delta_{kl|s}$ may depend on s. Note also that, despite a certain similarity, the Horvitz–Thompson estimator, $\hat{t}_{\mathrm{HT}} = \Sigma_r y_k / \pi_k$, is in general different from the π^*ES estimator, since the unconditional inclusion probability π_k is given by

$$\pi_k = \sum_{r \ni k} \sum_{s \supset r} p_a(s) p(r \mid s) = \sum_{s \ni k} p_a(s) \pi_{k|s}.$$

To determine the π_k, we must thus know $\pi_{k|s}$ for *every* s, knowledge often unavailable in an applied situation, since $\pi_{k|s}$ will often depend on information collected for the units in the particular first-phase sample s actually drawn. The Horvitz–Thompson estimator is thus impractical.

To prove Result 0, express the error of the estimator as

$$\begin{aligned} \hat{t}_{\pi^*} - t &= (\Sigma_s \check{y}_k - \Sigma_U y_k) + (\Sigma_r \check{y}_k^\dagger - \Sigma_s \check{y}_k) \\ &= A_s + B_r, \end{aligned} \tag{3.4}$$

say. Then use the rules $E(.) = E_a E(. \mid s)$, $V(.) = V_a E(. \mid s) + E_a V(. \mid s)$, where the operators E_a and V_a refer to phase one, and $E(. \mid s)$ and $V(. \mid s)$ to phase two, given the outcome s of phase one. Now, $E(A_s \mid s) = A_s$ and $E(B_r \mid s) = 0$, so that $E(\hat{t}_{\pi^*}) - t = E_a(A_s) = 0$. Moreover, $V(A_s \mid s) = 0$ and

$$V(B_r \mid s) = \Sigma\Sigma_s \Delta_{kl|s} \check{y}_k^\dagger \check{y}_l^\dagger, \quad V_a(A_s) = \Sigma\Sigma_U \Delta_{akl} \check{y}_k \check{y}_l,$$

whereby the variance result follows. That (3.3) is an unbiased estimator of the variance (3.2) follows by noting that, for arbitrary constants c_{kl},

$$E_a E(\Sigma\Sigma_r c_{kl} / \pi_{kl|s} \mid s) = E_a(\Sigma\Sigma_s c_{kl}) = \Sigma\Sigma_U \pi_{akl} c_{kl}. \tag{3.5}$$

This equation establishes the unbiasedness of the estimated first component if we take $c_{kl} = \check{\Delta}_{akl} \check{y}_k \check{y}_l$. As for the estimated second component, we use only the first equation of (3.5); the c_{kl} may then depend on s, and the appropriate choice is $c_{kl} = \Delta_{kl|s} \check{y}_k^\dagger \check{y}_l^\dagger$.

If $\hat{t}$ stands for an estimator of t, that is $\hat{t}$ could be $\hat{t}_{\pi^*}$, or any of the estimators presented later, it is understood that an approximately $100(1-\alpha)\%$ confidence interval for t is constructed as $\hat{t} \pm z_{1-\alpha/2}\{\hat{V}(\hat{t})\}^{\frac{1}{2}}$, where the constant $z_{1-\alpha/2}$ is exceeded with probability $\alpha/2$ by the unit normal random variable.

4 Application: Two-phase sampling for stratification

As an application of the preceding, we consider stratified random sampling in phase two. However, more generally than in the usual sampling texts, we here permit an arbitrary design, $p_a(s)$, for the first phase. Information is collected for the n_s units in s and used to partition s into H_s strata s_h, for $h = 1, \ldots, H_s$. Denote by n_h the size of s_h. From

s_h, a subsample r_h of size m_h is drawn ($h = 1, \ldots, H_s$). The complete second-phase sample, r, is the union of the H_s sets r_h. The size of r, m_r, is the sum of the m_h. We examine two stratified designs for the second phase:

Case STSI (STratified Simple random sampling), where r_h is drawn from s_h by Simple random sampling (SI):

Case STBE (STratified BErnoulli sampling), where r_h is drawn from s_h by Bernoulli sampling.

Here, STSI is of great practical interest for two-phase sampling; our interest in STBE is more motivated by the applications to nonresponse theory given in §§ 7–10 below. Sampling variance is interpreted via a repeated sampling process about which we assume, in both cases, that exactly the same second-phase stratified design (same strata, same sampling fractions within the strata) would be used every time a specific first phase sample is realized. However, for two nonidentical first-phase samples, suitably different stratifications may be used. There may be differences in the number of strata, H_s, as well as in the principle used to demarcate the strata.

Case STSI. Given the strata s_h, the statistician specifies certain subsample sizes m_h. For units in s_h,

$$\pi_{k|s} = \pi_{kk|s} = m_h/n_h = f_h, \quad \pi_{kl|s} = f_h(m_h - 1)/(n_h - 1) \quad (k \neq l), \tag{4.1}$$

while $\pi_{kl|s} = f_h f_{h'}$ when k and l belong to two different strata, s_h and $s_{h'}$. The π^*ES estimator (3.1) takes the form

$$\hat{t}_{\pi^*} = \Sigma_r \check{y}_k/\pi_{k|s} = \sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_h} \check{y}_k. \tag{4.2}$$

The variance, obtained from (3.2), is

$$V(\hat{t}_{\pi^*}) = \Sigma\Sigma_U \Delta_{akl} \check{y}_k \check{y}_l + E_a\left\{\sum_{h=1}^{H_s} n_h^2(1 - f_h) S^2_{\check{y}s_h}/m_h\right\}, \tag{4.3}$$

where $S^2_{\check{y}s_h}$ is the variance of $\check{y}_k = y_k/\pi_{ak}$ in s_h. By the variance of certain numbers z_k in a set A of n_A units, we mean $\Sigma_A(z_k - \bar{z}_A)^2/(n_A - 1) = S^2_{zA}$, where $\bar{z}_A = \Sigma_A z_k/n_A$. The variance estimator, obtained from (3.3), is, provided $m_h \geq 2$ for all h,

$$\hat{V}(\hat{t}_{\pi^*}) = \Sigma\Sigma_r \check{\Delta}_{akl} \check{y}_k \check{y}_l/\pi_{kl|s} + \sum_{h=1}^{H_s} n_h^2(1 - f_h) S^2_{\check{y}r_h}/m_h, \tag{4.4}$$

where $S^2_{\check{y}r_h}$ is the variance of $\check{y}_k$ in the set r_h and $\pi_{kl|s}$ is given by (4.1). The estimated second component is of a form that is familiar in the context of stratified sampling.

Example 1. Let the first-phase design be simple random sampling (SI) with fixed sample size $n_s = n$, and let $w_h = n_h/n$; $f = n/N$. Then, with STSI in phase-two, the π^*ES estimator is

$$\hat{t}_{\pi^*} = N \sum_{h=1}^{H_s} w_h \bar{y}_{r_h} = N\hat{\mu}, \tag{4.5}$$

say. Its design variance, from (4.3), is

$$V(\hat{t}_{\pi^*}) = N^2(1 - f) S^2_{yU}/n + E_{\mathrm{SI}}\left\{N^2 \sum_{h=1}^{H_s} w_h^2(1 - f_h) S^2_{ys_h}/m_h\right\}. \tag{4.6}$$

From (4.4), the unbiased estimators of the two components are given by

$$\hat{V}_1 = N^2 \frac{1-f}{n} \sum_{h=1}^{H_s} w_h \left\{ (1-Q_h) S_{yr_h}^2 + \frac{n}{n-1} (\bar{y}_{r_h} - \hat{\mu})^2 \right\},$$

$$\hat{V}_2 = \frac{N^2}{n^2} \sum_{h=1}^{H_s} n_h^2 \frac{1-f_h}{m_h} S_{yr_h}^2,$$

with $Q_h = (n - n_h)/\{(n-1)m_h\}$. Thus $V(\hat{t})$ is estimated by $\hat{V}_1 + \hat{V}_2$, which can be expressed as

$$\hat{V}(\hat{t}_{\pi^*}) = N(N-1) \sum_{h=1}^{H_s} \left\{ \frac{n_h - 1}{n-1} - \frac{m_h - 1}{N-1} \right\} w_h S_{yr_h}^2 / m_h + \frac{N(N-n)}{n-1} \sum_{h=1}^{H_s} w_h (\bar{y}_{r_h} - \hat{\mu})^2. \quad (4.7)$$

Note that these results do not require the same stratification principle for every s. Rao (1973) and Cochran (1977, p. 328) also obtained the result (4.7), but for the conceptually different situation where there exists a predetermined set of population strata, $U_1, \ldots, U_h, \ldots, U_H$, although with unknown sizes, that serve as the common basis for stratifying each possible s, so that $s_h = s \cap U_h$, for $h = 1, \ldots, H$.

Case STBE. The subsample r_h from s_h is realized as follows. The inclusion or noninclusion in r_h of a certain unit k in s_h is decided by a Bernoulli experiment, the probability for inclusion being specified as θ_{hs} for all $k \in s_h$. The experiments are independent. The notation θ_{hs} is chosen to indicate that the probability may differ from one stratum to another, and from one sample s to another. For the given stratification of s, the subsample size m_h is random with the expected value $\theta_{hs} n_h$. Two alternatives emerge for the analysis: (a) the θ_{hs} are treated as known; (b) the θ_{hs} are treated as unknown and estimated from the sample. Even though the θ_{hs} are ordinarily known in two-phase sampling, little is probably lost by always following (b). Under the heading Case STBE, we shall only pursue alternative (b), the case of interest for the treatment of nonresponse, where the θ_{hs} play the role of unknown response probabilities. It is convenient to condition on $\mathbf{m} = (m_1, \ldots, m_h, \ldots, m_{H_s})$, the vector of realized counts. Define

$$\pi_{k|s,\mathbf{m}} = \Pr(k \in r \mid s, \mathbf{m}), \quad \pi_{kl|s,\mathbf{m}} = \Pr(k, l \in r \mid s, \mathbf{m}). \quad (4.8)$$

When s is fixed, so is $\mathbf{n} = (n_1, \ldots, n_{H_s})$. If $\mathbf{m}$ is also fixed, then r_h is an SI sample of m_h units from n_h. Thus, for units in s_h,

$$\pi_{k|s,\mathbf{m}} = \pi_{kk|s,\mathbf{m}} = m_h/n_h = f_h, \quad \pi_{kl|s,\mathbf{m}} = f_h(m_h - 1)/(n_h - 1) \quad (k \neq l), \quad (4.9)$$

while $\pi_{kl|s,\mathbf{m}} = f_h f_{h'}$ if k and l belong to different strata, s_h and $s_{h'}$. Let A_{1s} be the event that $m_h \geq 1$ for $h = 1, \ldots, H_s$. Supposing A_{1s} occurs, we can consider the 'conditional π^*ES estimator'

$$\hat{t}_{c\pi^*} = \Sigma_r \check{y}_k / \pi_{k|s,\mathbf{m}} = \sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_h} \check{y}_k, \quad (4.10)$$

with the variance expression

$$V(\hat{t}_{c\pi^*}) = \Sigma\Sigma_U \Delta_{akl} \check{y}_k \check{y}_l + E_a E_{\mathbf{m}} \left\{ \sum_{h=1}^{H_s} n_h^2 (1-f_h) S_{\check{y}s_h}^2 / m_h \right\}, \quad (4.11)$$

where $E_{\mathbf{m}}(.)$ indicates expectation over all realizations $\mathbf{m}$, given s and A_{1s}. Further let A_{2s} denote the event that $m_h \geq 2$ for $h = 1, \ldots, H_s$. If A_{2s} occurs, a variance estimator is

given by

$$\hat{V}(\hat{t}_{c\pi^*}) = \Sigma\Sigma_r \check{\Delta}_{akl}\check{y}_k\check{y}_l/\pi_{kl|s,\mathbf{m}} + \sum_{h=1}^{H_r} n_h^2(1-f_h)S^2_{\check{y}r_h}/m_h, \tag{4.12}$$

where $\pi_{kl|s,\mathbf{m}}$ is given by (4.9).

Remark 1. A comparison of Cases STSI and STBE reveals a useful analogy. The two estimators (4.2) and (4.10) agree formally, since $\pi_{k|s} = \pi_{k|s,\mathbf{m}}$. The respective variances, (4.3) and (4.11), differ, since the two sampling designs generate two different sampling distributions for $\hat{t}$. But coming to the estimated variances, (4.4) and (4.12), the two cases again agree formally, since $\pi_{kl|s} = \pi_{kl|s,\mathbf{m}}$. In Case STBE, if the probability of A_{1s} is very near unity, a confidence interval constructed around $\hat{t}_{c\pi^*}$, with (4.12) as the variance estimator, will, in most practical situations, yield essentially correct confidence levels.

Example 2. As in Example 1, let the first-phase design be simple random sampling. If the design STBE is used in phase two, the estimator of t is given, as in the case of STSI, by (4.5). Moreover, the variance estimator is given by (4.7).

5 Regression estimation in two-phase sampling

The π^*ES estimator can be described as a pure 'weighting-type' estimator. For example, in double sampling for stratification, § 4, the information recorded after phase one is used to form strata, and consequently the weights $1/\pi_k^* = 1/\{(m_h/n_h)\pi_{ak}\}$ in the π^*ES estimator (4.2) reflect the stratified sampling in the second phase. In the regression-type estimators now to be considered, recorded auxiliary variables enter explicitly into the formula for the estimator. We distinguish three situations depending on the nature of the auxiliary information.

Situation 1. The value $\mathbf{x}_k = (x_{1k}, \ldots, x_{qk})'$ of the auxiliary vector $\mathbf{x} = (x_1, \ldots, x_q)'$ is recorded for the units $k \in s$.

Situation 2. The value $\mathbf{x}_k$ is available for all units k in the entire population U.

Situation 3. A combination of Situations 1 and 2. The value $\mathbf{x}_k$ is recorded for $k \in s$, and some other (perhaps 'weaker') information $\mathbf{z}_k = (z_{1k}, \ldots, z_{pk})'$ is known for all $k \in U$.

Let us first develop Situations 1 and 2, using extensions of the regression approach (Cassel, Särndal & Wretman, 1976; Särndal, 1982). We assume the existence of a regression relationship between y_k and $\mathbf{x}_k$ in the finite population. Then, although y_k may be observed only in a smallish second-phase sample, the scarcity of the y-information can be partly compensated for by using a design-based regression estimator derived as follows. Suppose a scatter plot of y_k on $\mathbf{x}_k$ in the entire finite population looks as if a reasonably good fit could be obtained by the linear model ξ stating that $E_\xi(y_k) = \mathbf{x}_k'\boldsymbol{\beta}$, $V_\xi(y_k) = \sigma_k^2$ and the y_k are independent. Here, the σ_k^2 are known up to multiplicative constants that vanish when $\mathbf{b}$ below is calculated. The model ξ is just a tool to express the presumed relationship between y and $\mathbf{x}$ in the finite population; it is not assumed to be 'true' and does not in any way serve below to analyse the statistical properties of our estimators. If all N points $(y_k, \mathbf{x}_k)$ were available, the weighted least-squares $\boldsymbol{\beta}$-estimator and the associated residuals are

$$\mathbf{B} = (\Sigma_U \mathbf{x}_k\mathbf{x}_k'/\sigma_k^2)^{-1}\Sigma_U \mathbf{x}_k y_k/\sigma_k^2, \quad E_k = y_k - \mathbf{x}_k'\mathbf{B}. \tag{5.1}$$

However, $(y_k, \mathbf{x}_k)$ is observed for $k \in r$ only, and the kth unit carries the weight π_k^{*-1}. Let

us therefore estimate $\mathbf{B}$ by

$$\mathbf{b} = \{\Sigma_r \mathbf{x}_k \mathbf{x}_k'/(\sigma_k^2 \pi_k^*)\}^{-1} \Sigma_r \mathbf{x}_k y_k/(\sigma_k^2 \pi_k^*). \tag{5.2}$$

The estimator $\mathbf{b}$ will serve to calculate the predicted values $\hat{y}_k = \mathbf{x}_k'\mathbf{b}$, for $k \in s$, since $\mathbf{x}_k$ is known for $k \in s$, and the residuals

$$e_k = y_k - \hat{y}_k = y_k - \mathbf{x}_k'\mathbf{b} \quad (k \in r), \tag{5.3}$$

since y_k is known for $k \in r$ only. In Situation 1, we can form the regression estimator described in Result 1 below, where the variance expression is approximate, therefore denoted AV.

RESULT 1. *In two-phase sampling, when $\mathbf{x}_k$ is recorded for $k \in s$, an approximately design unbiased estimator of $t = \Sigma_U y_k$ is given by*

$$\hat{t}_{1\text{REG}} = \Sigma_s \hat{y}_k/\pi_{ak} + \Sigma_r (y_k - \hat{y}_k)/\pi_k^*. \tag{5.4}$$

Let E_k be given by (5.1) *and $\check{E}_k^\dagger = E_k/\pi_k^*$. The approximate variance is*

$$\text{AV}(\hat{t}_{1\text{REG}}) = \Sigma\Sigma_U \Delta_{akl} \check{y}_k \check{y}_l + E_a\{\Sigma\Sigma_s \Delta_{kl|s} \check{E}_k^\dagger \check{E}_l^\dagger\}. \tag{5.5}$$

Let e_k be given by (5.3) *and $\check{e}_k^\dagger = e_k/\pi_k^*$. A variance estimator is then given by*

$$\hat{V}(\hat{t}_{1\text{REG}}) = \Sigma\Sigma_r \check{\Delta}_{akl} \check{y}_k \check{y}_l/\pi_{kl|s} + \Sigma\Sigma_r \Delta_{kl|s} \check{e}_k^\dagger \check{e}_l^\dagger/\pi_{kl|s}. \tag{5.6}$$

Without complete detail, let us justify Result 1. The bias and the variance of (5.4) are complicated because of the nonlinear random variable $\mathbf{b}$. Explicit expressions can only be had through approximation. Approximating $\mathbf{b}$ by its population analogue $\mathbf{B}$, a constant, we have

$$\hat{t}_{1\text{REG}} \doteq \hat{t}_{1\text{REG}}^0 = \Sigma_s y_k^0/\pi_{ak} + \Sigma_r (y_k - y_k^0)/\pi_k^*,$$

where $y_k^0 = \mathbf{x}_k'\mathbf{B}$ has replaced $\hat{y}_k = \mathbf{x}_k'\mathbf{b}$. The advantage gained is that $\hat{t}_{1\text{REG}}^0$ is extremely simple to analyse. We have

$$\begin{aligned} \hat{t}_{1\text{REG}} - t \doteq \hat{t}_{1\text{REG}}^0 - t &= (\Sigma_s \check{y}_k - \Sigma_U y_k) + (\Sigma_r \check{E}_k^\dagger - \Sigma_s \check{E}_k), \\ &= A_s + B_r', \end{aligned} \tag{5.7}$$

say, where $\check{E}_k = E_k/\pi_{ak}$, $\check{E}_k^\dagger = E_k/\pi_k^* = \check{E}_k/\pi_{k|s}$. It follows that

$$E(\hat{t}_{1\text{REG}}) - t \doteq E(\hat{t}_{1\text{REG}}^0) - t = 0,$$

so that $\hat{t}_{1\text{REG}}$ is approximately unbiased. To obtain the variance, note the analogy between (5.7) and (3.4). The term A_s is present in both expressions. The terms B_r and B_r' differ only in that the latter is expressed in the residuals E_k, the former in the raw scores y_k. The argument used in proving (3.2) leads directly to (5.5), which is in this case an approximate expression because $\mathbf{b}$ was approximated by $\mathbf{B}$. In obtaining an estimated variance, $E_k = y_k - \mathbf{x}_k'\mathbf{B}$ cannot be used since it contains the unknown $\mathbf{B}$. Instead, substitute $e_k = y_k - \mathbf{x}_k'\mathbf{b}$, where $\mathbf{b}$ is calculated from the sample, and the formula (5.6) follows.

The following Result 2 deals with Situation 2.

RESULT 2. *In two-phase sampling, when $\mathbf{x}_k$ is recorded for all $k \in U$, an approximately design unbiased estimator of $t = \Sigma_U y_k$ is given by*

$$\hat{t}_{2\text{REG}} = \Sigma_U \hat{y}_k + \Sigma_r (y_k - \hat{y}_k)/\pi_k^*. \tag{5.8}$$

Let E_k be given by (5.1), *$\check{E}_k = E_k/\pi_{ak}$ and $\check{E}_k^\dagger = E_k/\pi_k^*$. An approximate variance expression is*

$$\text{AV}(\hat{t}_{2\text{REG}}) = \Sigma\Sigma_U \Delta_{akl} \check{E}_k \check{E}_l + E_a\{\Sigma\Sigma_s \Delta_{kl|s} \check{E}_k^\dagger \check{E}_l^\dagger\}. \tag{5.9}$$

Let e_k be given by (5.3), $\check{e}_k = e_k/\pi_{ak}$ *and* $\check{e}_k^\dagger = e_k/\pi_k^*$. *A variance estimator is*

$$\hat{V}(t_{2\text{REG}}) = \Sigma\Sigma_r \check{\Delta}_{akl}\check{e}_k\check{e}_l/\pi_{kl|s} + \Sigma\Sigma_r \Delta_{kl|s}\check{e}_k^\dagger\check{e}_l^\dagger/\pi_{kl|s} = \Sigma\Sigma_r \Delta_{kl}^*\check{e}_k^\dagger\check{e}_l^\dagger/\pi_{kl}^*. \tag{5.10}$$

A justification of Result 2 can be produced along lines that resemble the argument used above for Result 1. We omit the details.

Let us turn to Situation 3, where the auxiliary information comes from two sources: the vector $\mathbf{x}_k$ is available for $k \in s$, and another vector $\mathbf{z}_k$ for $k \in U$. In this case, one fitted regression will estimate the relation between $\mathbf{x}_k$ and y_k, another that between $\mathbf{z}_k$ and y_k. The first fit is, as in Situations 1 and 2, summarized by formulae (5.1)–(5.3).

Suppose that the finite population scatter of $(y_k, \mathbf{z}_k)$ is such that a reasonably good fit can be obtained by the model $E_{\xi_1}(y_k) = \mathbf{z}_k'\boldsymbol{\beta}_1$, $V_{\xi_1}(y_k) = \sigma_{1k}^2$. If all N points $(y_k, \mathbf{z}_k)$ were observed, the $\boldsymbol{\beta}_1$-estimator and the residuals would be

$$\mathbf{B}_1 = (\Sigma_U \mathbf{z}_k\mathbf{z}_k'/\sigma_{1k}^2)^{-1}\Sigma_U \mathbf{z}_k y_k/\sigma_{1k}^2, \quad E_{1k} = y_k - \mathbf{z}_k'\mathbf{B}_1. \tag{5.11}$$

But the information about y_k is less extensive, so we must estimate $\mathbf{B}_1$. To this end, consider two possibilities.

The first method follows naturally from the fact that y_k is available for the set r only:

$$\mathbf{b}_1 = \{\Sigma_r \mathbf{z}_k\mathbf{z}_k'/(\sigma_{1k}^2\pi_k^*)\}^{-1}\Sigma_r \mathbf{z}_k y_k/(\sigma_{1k}^2\pi_k^*). \tag{5.12}$$

The second method, slightly more complicated, recognizes that the known $\mathbf{x}_k$-values for $k \in s$ make it possible to calculate 'pseudo-observations', y_k^*, for $k \in s$, although y_k itself is known for $k \in r$ only. Let us define the pseudo-observations as

$$y_k^* = \begin{cases} \hat{y}_k + (y_k - \hat{y}_k)/\pi_{k|s} & (k \in r), \\ \hat{y}_k & (k \in s - r), \end{cases}$$

where $\hat{y}_k = \mathbf{x}_k'\mathbf{b}$, with $\mathbf{b}$ given by (5.2). Now, the second estimator of $\mathbf{B}_1$ is taken as

$$\mathbf{b}_1 = \{\Sigma_s \mathbf{z}_k\mathbf{z}_k'/(\sigma_{1k}^2\pi_{ak})\}^{-1}\Sigma_s \mathbf{z}_k y_k^*/(\sigma_{1k}^2\pi_{ak}). \tag{5.13}$$

Both (5.12) and (5.13) are under general conditions consistent estimators of $\mathbf{B}_1$.

Whether (5.12) or (5.13) is used, we calculate predicted values as: $\hat{y}_{1k} = \mathbf{z}_k'\mathbf{b}_1$ for $k \in U$, since z_k is known for all $k \in U$; and residuals according to

$$e_{1k} = y_k - \mathbf{z}_k'\mathbf{b}_1 \quad (k \in r), \tag{5.14}$$

since y_k is available for $k \in r$ only.

The regression estimator proposed for Situation 3 does in fact combine the principles used in Situations 1 and 2.

RESULT 3. *In two-phase sampling, when $\mathbf{x}_k$ is recorded for $k \in s$ and $\mathbf{z}_k$ for $k \in U$, an approximately design unbiased estimator of $t = \Sigma_U y_k$ is*

$$\hat{t}_{3\text{REG}} = \Sigma_U \hat{y}_{1k} + \Sigma_s \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_{ak}} + \Sigma_r \frac{y_k - \hat{y}_k}{\pi_k^*}. \tag{5.15}$$

Let E_{1k} and E_k be given by (5.11) *and* (5.1), *respectively*; $\check{E}_{1k} = E_{1k}/\pi_{ak}$ *and* $\check{E}_k^\dagger = E_k/\pi_k^*$. *An approximate variance expression is then*

$$\text{AV}(\hat{t}_{3\text{REG}}) = \Sigma\Sigma_U \Delta_{akl}\check{E}_{1k}\check{E}_{1l} + E_a\{\Sigma\Sigma_s \Delta_{kl|s}\check{E}_k^\dagger\check{E}_l^\dagger\}. \tag{5.16}$$

Let e_{1k} and e_k be given by (5.14) *and* (5.3), *respectively*; $\check{e}_{1k} = e_{1k}/\pi_{ak}$ *and* $\check{e}_k^\dagger = e_k/\pi_k^*$. *A variance estimator is then*

$$\hat{V}(\hat{t}_{3\text{REG}}) = \Sigma\Sigma_r \check{\Delta}_{akl}\check{e}_{1k}\check{e}_{1l}/\pi_{kl|s} + \Sigma\Sigma_r \Delta_{kl|s}\check{e}_k^\dagger\check{e}_l^\dagger/\pi_{kl|s}. \tag{5.17}$$

Result 3 can be justified along lines similar to those reproduced in detail for Situation 1.

Example 3. Suppose that the first phase involves a two-stage sampling design: classes

of students (PSU's) are selected at the first stage; individual students (SSU's) are then subsampled within selected classes. The students thus selected form the first-phase sample, s, for which the inexpensive information x_k, say, grade point average, is recorded. The sampling weights relevant to phase one are $1/\pi_{ak} = 1/(\pi_{Ii}\pi_{ki})$, where π_{Ii} is the probability of including the ith PSU in the first-stage sample, and π_{ki} the probability of choosing the kth SSU of the ith PSU. A second-phase sample, r, is subsampled from s by simple random selection of, say, m_s of the n_s SSU's in s. For $k \in r$, the value y_k, a more expensive measure of performance, is recorded.

Assume that y is fairly well explained by the ratio model

$$E_\xi(y_k) = \beta x_k, \quad V_\xi(y_k) = \sigma^2 x_k \quad (k \in U). \tag{5.18}$$

The slope estimator and the residuals arising from the fit of this model are

$$b = (\Sigma_r \check{y}_k^\dagger)/(\Sigma_r \check{x}_k^\dagger), \quad e_k = y_k - bx_k \quad (k \in r), \tag{5.19}$$

where the weight $1/\pi_k^* = 1/(\pi_{Ii}\pi_{ki}m_s/n_s)$ is used for the calculation of $\check{y}_k^\dagger = y_k/\pi_k^*$ and $\check{x}_k^\dagger = x_k/\pi_k^*$.

In a case where the size N is the only information available at the level of the entire population, we may apply situation 3 with the 'trivial' model

$$E_{\xi_1}(y_k) = \beta_1, \quad V_{\xi_1}(y_k) = \sigma_1^2 \quad (k \in U), \tag{5.20}$$

corresponding to $z_k = 1$ for all k. Results 1, 2 and 3 yield the regression estimators

$$\hat{t}_{1\text{REG}} = (\Sigma_s \check{x}_k)b, \quad \hat{t}_{2\text{REG}} = (\Sigma_U x_k)b, \quad \hat{t}_{3\text{REG}} = N\bar{x}_s b,$$

with $\bar{x}_s = (\Sigma_s \check{x}_k)/\hat{N}$, $\hat{N} = \Sigma_s 1/\pi_{ak}$, and b given by (5.19). We have used (5.13) in deriving $\hat{t}_{3\text{REG}}$. The estimated variances are obtained from Results 1 to 3, where $e_k = y_k - bx_k$ and $e_{1k} = y_k - b\bar{x}_s$. Ordinarily, $\hat{t}_{3\text{REG}}$ is better than $\hat{t}_{1\text{REG}}$, and $\hat{t}_{2\text{REG}}$ best of the three, thanks to the more extensive x-information.

6 Regression estimation in two-phase sampling for stratification

Let us examine Situations 1, 2 and 3 when phase two involves stratified sampling. The set-up is that of §§ 4 and 5 combined. For units k in the first-phase sample, the statistician collects information by means of which s is partitioned into H_s sets s_h, which serve as strata for phase two. In addition, assume that auxiliary values $\mathbf{x}_k$, and possibly $\mathbf{z}_k$, are available, so as to fit the respective descriptions of Situations 1, 2 and 3. Other notation will be as in §§ 4 and 5. We conclude the following.

Case STSI. Results 1 to 3 apply straightforwardly, with $\pi_{k|s}$ and $\pi_{kl|s}$ determined by (4.1). That is, the kth observation is given the weight $1/\pi_k^*$, with $\pi_k^* = \pi_{ak} f_h$ for $k \in s_h$, where π_{ak} is the first-phase inclusion probability and $f_h = m_h/n_h$ is the sampling fraction used in s_h. For example, the first regression estimator is

$$\hat{t}_{1\text{REG}} = \sum_{h=1}^{H_s} \left\{ \Sigma_{s_h} \frac{\hat{y}_k}{\pi_{ak}} + f_h^{-1} \Sigma_{r_h} \frac{y_k - \hat{y}_k}{\pi_{ak}} \right\}, \tag{6.1}$$

reflecting the stratified nature of phase two. One notes that all three regression estimators share the same estimated second variance component, namely,

$$\hat{V}_2 = \sum_{h=1}^{H_s} n_h^2 (1 - f_h) S_{\check{e}r_h}^2 / m_h, \tag{6.2}$$

a 'stratified expression' in which $S^2_{\check{e}r_h}$ is the variance in the set r_h of the expanded residuals $\check{e}_k = (y_k - \mathbf{x}_k'\mathbf{b})/\pi_{ak}$.

Case STBE. The transition from Case STSI to Case STBE is done by conditioning on $\mathbf{m}$ as in § 4: $\pi_{k|s,\mathbf{m}}$ and $\pi_{kl|s,\mathbf{m}}$, defined by (4.9), will replace their unconditional counterparts $\pi_{k|s}$ and $\pi_{kl|s}$ in Results 1–3. Here, Remark 1 in § 4 is again relevant: in each of the three situations, the estimator and the corresponding estimated variance will be in formal agreement with Case STSI.

In Situation 1, for example, we obtain the following 'conditional regression estimator', approximately unbiased for $t = \Sigma_U y_k$:

$$\hat{t}_{c1\mathrm{REG}} = \Sigma_s \frac{\hat{y}_k}{\pi_{ak}} + \Sigma_r \frac{y_k - \hat{y}_k}{\pi_{ak}\pi_{k|s,\mathbf{m}}} = \sum_{h=1}^{H_s} \left\{ \Sigma_{s_h} \frac{\hat{y}_k}{\pi_{ak}} + f_h^{-1}\Sigma_{r_h} \frac{y_k - \hat{y}_k}{\pi_{ak}} \right\}, \tag{6.3}$$

which is in form identical to (6.1) above. The variance estimator is given by

$$\hat{V}(\hat{t}_{c1\mathrm{REG}}) = \Sigma\Sigma_r \check{\Delta}_{akl}\check{y}_k\check{y}_l/\pi_{kl|s,\mathbf{m}} + \sum_{h=1}^{H_s} n_h^2(1-f_h)S^2_{\check{e}r_h}/n_h.$$

Comparing this with its analogue (4.12) in the case of the conditional π^*ES estimator, we see that the important change is that the estimated second component has become 'residualized'. Ordinarily, $\hat{t}_{c1\mathrm{REG}}$ will yield a shorter confidence interval.

Example 4. We consider the situation outlined in Example 3. In phase one, a two-stage sample of students, s, is selected. Suppose that this first-phase sample is stratified (on the basis of sex and/or age, say) and that stratified sampling is used in phase two. Also, for $k \in s$, the variable x_k, grade point average, is recorded. The relation between y_k, recorded for $k \in r$ only, and x_k is again assumed to follow the ratio model (5.18). If the STSI sampling fraction in stratum h is $f_h = m_h/n_h$, the slope estimate becomes

$$b = \left(\sum_{h=1}^{H_s} f_h^{-1}\Sigma_{r_h}\check{y}_k \right) \Big/ \left(\sum_{h=1}^{H_s} f_h^{-1}\Sigma_{r_h}\check{x}_h \right). \tag{6.4}$$

Also, for Situation 3, let us fit the simple model (5.20). With b determined by (6.4), the first and third regression estimators of $t = \Sigma_U y_k$ are given, in Case STSI, by

$$\hat{t}_{1\mathrm{REG}} = (\Sigma_s \check{x}_k)b, \quad \hat{t}_{3\mathrm{REG}} = N\bar{x}_s b, \tag{6.5}$$

with $\bar{x}_s = (\Sigma_s \check{x}_k)/(\Sigma_s 1/\pi_{ak})$. The residuals necessary for the variance estimation are $e_k = y_k - bx_k$, $e_{1k} = y_k - b\bar{x}_s$. This leads to the estimated variances

$$\hat{V}(\hat{t}_{1\mathrm{REG}}) = \Sigma\Sigma_r \check{\Delta}_{akl}\check{y}_k\check{y}_l/\pi_{kl|s} + \hat{V}_2, \quad \hat{V}(\hat{t}_{3\mathrm{REG}}) = \Sigma\Sigma_r \check{\Delta}_{akl}\check{e}_{1k}\check{e}_{1l}/\pi_{kl|s} + \hat{V}_2, \tag{6.6}$$

where $\hat{V}_2$ is given by (6.2), and $\pi_{kl|s}$ by (4.1).

The results (6.5)–(6.6) apply without any formal change in Case STBE, although the notation should then be $\hat{t}_{1c\mathrm{REG}}$, $\hat{t}_{3c\mathrm{REG}}$ to indicate the conditional nature of the regression estimator.

In a practical situation, the approach to two-phase sampling presented above will clearly require certain judgements on the part of the statistician about the best way to utilize the available auxiliary information, notably the information gathered for the units k in the phase-one sample s. Following phase one, the statistician must:

(i) make a choice of a sampling design for phase two;

(ii) make a choice of an estimator, using the regression modeling approach. He may, for example, choose to use a very simple second-phase design and instead utilize most or all of the gathered information directly in the regression estimator formula. Alternatively,

he may use some (or all) of the gathered information to stratify or in other ways create a more efficient second phase design. It may still be advantageous (but somewhat less imperative) to use an estimator of the regression type.

7 The nonresponse problem

In §§ 7–10, we consider the situation of randomized sample selection, followed by a certain nonresponse. We assume that the set s, the 'intended sample', is drawn by some given (arbitrary) sampling design. For units $k \in s$, certain information may be recorded. Subselection occurs by fact that the measurement y_k is obtained for units $k \in r$ only, where $r \subseteq s$. We call r the response set; $s - r$ is the nonresponse set. We invoke the assumption, commonly made in recent nonresponse literature, that r is realized, given s, through a probabilistic mechanism of unknown form, called the response mechanism.

As far as possible we shall use the same notation as in §§ 2–6 for concepts that directly correspond to each other. The response mechanism (which corresponds to the second-phase sampling design) will thus be denoted $p(r \mid s)$. Here the statistician is forced to make an explicit assumption about the unknown form of $p(r \mid s)$; earlier $p(r \mid s)$ was completely known.

Our discussion includes the 'adjustment group technique', one of the widely used attempts to eliminate or reduce bias due to unit nonresponse. In this technique, one applies a weight, n_h/m_h, equaling the inverse of the response rate in the hth group, to every respondent value y_k from the group. In addition, the ordinary sampling weight is applied. Thus, if simple random sampling of n out of N is used to draw the sample, the estimator of the population total $t = \Sigma y_k$, where the sum is over $k = 1, \ldots, N$, becomes

$$\hat{t} = \sum_{h=1}^{H} (N/n)(n_h/m_h)\Sigma_{r_h} y_k = N \sum_{h=1}^{H} w_h \bar{y}_{r_h}, \tag{7.1}$$

where H is the fixed number of adjustment groups, $w_h = n_h/n$ is the sample portion in the hth group and $\bar{y}_{r_h}$ is the respondent mean of y in the hth group. The estimator (7.1) has been analysed, in different ways, by Thomsen (1973), Oh & Scheuren (1983) and others. Different points of departure may be used in the analysis of (7.1). An analysis that was standard up until recently used the 'deterministic model' of a dichotomized population, as described by Cochran (1977, p. 359):

> In the study of nonresponse it is convenient to think of the population as divided into two 'strata', the first consisting of all units for which measurements would be obtained if the units happened to fall in the sample, the second of the units for which no measurements would be obtained.

Under this model, units in the response stratum respond with probability one, the other units with probability zero. This places the survey statistician in the uncomfortable position that valid conclusions from the sample data (which come from respondents only) can only be extended to the response stratum of the population. Cochran (1977, p. 360) is quick to admit the limitations of the deterministic model:

> This division into two distinct strata is, of course, an oversimplification. Chance plays a part in determining whether a unit is found and measured in a given number of attempts. In a more complete specification of the problem we would attach to each unit a probability representing the chance that it would be measured by a given field method if it fell in the sample.

In the direction hinted at by Cochran, more recent analyses of estimators for the nonresponse situation favour a framework where the response behavior is considered stochastic rather than deterministic. This spirit penetrates several of the papers in the recent and authoritative 'incomplete data in sample surveys', for example, Oh & Scheuren (1983), Platek & Gray (1983) and Cassel et al. (1983).

The estimator (7.1) can be justified through the assumption that the population U is composed of a fixed set of disjoint subpopulations such that all units in a subpopulation have the same response probability, and that units respond independently of each other. This model is called a 'uniform response mechanism within subpopulations' by Oh & Scheuren (1983), who fittingly describe the set-up with probability sampling augmented with a response model as 'quasi-randomization'. Assuming simple random sampling and a uniform response mechanism within subpopulations, they analyse the bias, variance and mean squared error of the estimator (7.1), conditionally on the n_h and the m_h, as well as unconditionally. In our opinion, a model for the response mechanism should be formulated *given the sample s*, with consideration given to the survey operations to which the units in the *particular sample s* are exposed (Dalenius, 1983). Consider, for example, the case where a team of interviewers carry out the field work. The required number of interviewers may depend on the geographical spread of the particular sample s. Differences in interviewing skill, in age, sex and race of the interviewers will create differences in response rates. This should be reflected in the response model, for example, by partitioning the particular s into groups that correspond to the interviewers, or, if the data base is sufficiently large, to interviewers crossed with respondent age-sex groups. (The model with fixed subpopulations may be adequate for a one-interviewer situation, or for a mail survey.) We shall therefore formulate a more general response model.

8 Theoretical results for the nonresponse situation

We assume that the intended sample s, of size n_s, is drawn from the population $U = \{1, \ldots, k, \ldots, N\}$ by the arbitrary design $p_a(s)$. The quantities π_{ak}, π_{akl} and Δ_{akl} associated with this design are defined as in § 2. In particular, $p_a(.)$ may be a 'complex' design in two or more stages. Once drawn, s is partitioned into H_s groups, s_h, for $h = 1, \ldots, H_s$. Denote by n_h the size of s_h, by r_h the responding subset of s_h, and by m_h the size of r_h. The total set of respondents, r, is the union of the r_h; the size of r, m_r, is the sum of the m_h. We assume that the individual response probability is the same for all units in s_h, for $h = 1, \ldots, H_s$, which we call *response homogeneity groups* (RHG's). Units are assumed to respond independently of each other. Thus we have the RHG *model*: for $h = 1, \ldots, H_s$,

$$\Pr(k \in r \mid s) = \pi_{k|s} = \theta_{hs} \quad (k \in s_h), \quad \Pr(k, l \in r \mid s) = \pi_{kl|s} = \Pr(k \in r \mid s) \Pr(l \in r \mid s) \quad (k \neq l). \quad (8.1)$$

The number of groups, H_s, and their definition may change with s; the principle for forming the groups is not necessarily the same for all samples s. The RHG model is an exact copy of the randomization imposed under stratified Bernoulli sampling, Case STBE, in §§ 4 and 6. However, in contrast to Case STBE, the RHG model is 'only' an assumption, not an actually imposed randomization scheme. Assuming that the RHG model (8.1) holds, we can thus directly borrow the results reported above in §§ 4 and 6, under the heading Case STBE. Note that $\mathbf{m} = (m_1, \ldots, m_{H_s})$ is now the vector of respondent counts, and $f_h = m_h/n_h$ is the response rate in the hth group, s_h. As in § 4, we can identify a 'basic situation', which leads to a 'conditional π^*ES estimator', and Situations 1, 2 and 3 (as defined in § 5), with different extents of auxiliary information and leading to three different 'conditional regression estimators'. For easy reference, we list the estimators for the four different situations as follows.

The conditional π^*ES estimator becomes

$$\hat{t}_{c\pi^*} = \sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_h} \check{y}_k. \quad (8.2)$$

The conditional regression estimators are given by

$$\hat{t}_{c1\text{REG}} = \sum_{h=1}^{H_s} \left(\Sigma_{s_h} \frac{\hat{y}_k}{\pi_{ak}} + f_h^{-1} \Sigma_{r_h} \frac{y_k - \hat{y}_k}{\pi_{ak}} \right), \tag{8.3}$$

$$\hat{t}_{c2\text{REG}} = \Sigma_U \hat{y}_k + \sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_h} \frac{y_k - \hat{y}_k}{\pi_{ak}}, \tag{8.4}$$

$$\hat{t}_{c3\text{REG}} = \Sigma_U \hat{y}_{1k} + \sum_{h=1}^{H_s} \left(\Sigma_{s_h} \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_{ak}} + f_h^{-1} \Sigma_{r_h} \frac{y_k - \hat{y}_k}{\pi_{ak}} \right). \tag{8.5}$$

The estimator (8.2), its variance and its estimated variance (see below) are discussed, in the context of nonresponse, by Singh & Singh (1979).

If the assumed RHG model holds, these four estimators are at least approximately unbiased for t. If $\hat{t}$ denotes one of these estimators, the estimated variance is of the form

$$\hat{V}(\hat{t}) = \hat{V}_1(\hat{t}) + \hat{V}_2(\hat{t}), \tag{8.6}$$

where $\hat{V}_1(\hat{t})$ estimates the variance contribution due to the randomized selection of the intended sample s, and $\hat{V}_2(\hat{t})$ estimates the variance due to nonresponse under the RHG model. If the response is complete, that is $r = s$, then $\hat{V}_2(\hat{t}) = 0$.

The estimated first component, with $\pi_{kl|s,\mathbf{m}}$ given by (4.9), is

$$\hat{V}_1(\hat{t}) = \Sigma\Sigma_r \check{\Delta}_{akl} \check{a}_k \check{a}_l / \pi_{kl|s,\mathbf{m}}, \tag{8.7}$$

where the definition of the quantities $\check{a}_k$ depends on the estimator. For $\hat{t} = \hat{t}_{c\pi^*}$ and for $\hat{t} = \hat{t}_{c1\text{REG}}$, we have $\check{a}_k = \check{y}_k = y_k/\pi_{ak}$; for $\hat{t} = \hat{t}_{c2\text{REG}}$, $\check{a}_k = e_k/\pi_{ak} = (y_k - \hat{y}_k)/\pi_{ak}$, and finally for $\hat{t} = \hat{t}_{c3\text{REG}}$, $\check{a}_k = e_{1k}/\pi_{ak} = (y_k - \hat{y}_{1k})/\pi_{ak}$. The fitted values $\hat{y}_k$ and $\hat{y}_{1k}$ are as defined in § 5. The estimated second component is given by the 'stratified form'

$$\hat{V}_2(\hat{t}) = \sum_{h=1}^{H_s} n_h^2 (1 - f_h) S_{\check{a}r_h}^2 / m_h, \tag{8.8}$$

where $S_{\check{a}r_h}^2$ is the variance in the set r_h of the quantities $\check{a}_k$ defined as follows: for $\hat{t} = \hat{t}_{c\pi^*}$, $\check{a}_k = \check{y}_k = y_k/\pi_{ak}$; for the other three estimators, $\check{a}_k = \check{e}_k = (y_k - \hat{y}_k)/\pi_{ak}$.

An approximately $100(1-\alpha)\%$ confidence interval is formed as $\hat{t} \pm z_{1-\alpha/2}\sqrt{\{\hat{V}(\hat{t})\}}$, where $z_{1-\alpha/2}$ is exceeded with probability $\alpha/2$ by the unit normal variate. This interval takes nonresponse into account and assumes that a correct RHG model has been formulated. (Otherwise the estimator is more or less biased, and the interval tends to be off-center.)

Our frequentist interpretation of variance and confidence intervals appeals to an imagined two-step process of repeated samples s and, for each s, repreated realizations r under the RHG model (8.1). We assume that each time a given sample s is selected, the repeated realizations of the model (8.1) are always with the same number of groups, H_s, and by the same grouping principle, but that these factors may change with s.

Example 5. Two-stage sampling with nonresponse. Suppose the PSU's are large and cutting across the RHG's. Assume that the SI design is used at each of the two stages: a sample s_I of n_I clusters is drawn from N_I at the first stage ($f_I = n_I/N_I$); if the ith PSU is selected, a sample s_i of n_i units is drawn from N_i at the second stage ($f_i = n_i/N_i$). The resulting sample of SSU's, which is the union of the n_I sets s_i, is divided into RHG's s_h, for $h = 1, \ldots, H_s$, and $f_h = m_h/n_h$ is the response rate in the hth group. The conditional π^*ES estimator is, from (8.2),

$$\hat{t}_{c\pi^*} = f_I^{-1} \sum_{i \in s_I} f_i^{-1} \left(\sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_{ih}} y_k \right), \tag{8.9}$$

where r_{ih} is the set of respondents in the cross-classification of the ith PSU with the hth RHG. Here three different inverted fractions, f_I^{-1}, f_i^{-1} and f_h^{-1}, intervene as weights. The weight due to the sample selection is $\pi_{ak}^{-1} = f_I^{-1} f_i^{-1}$ for all SSU's k in the ith PSU, whereas f_h^{-1} is the weight associated with correction for nonresponse in the hth group.

Now suppose in addition that an auxiliary variable x_k is recorded for $k \in s$, and that a ratio model is a decent description of the x-to-y relationship, $E(y_k) = \beta x_k$, $V(y_k) = \sigma^2 x_k$. For this model, the regression estimator (8.3) becomes

$$\hat{t}_{c1\text{REG}} = \left(f_I^{-1} \sum_{i \in s_I} f_i^{-1} \Sigma_{s_i} x_k \right) b, \tag{8.10}$$

where

$$b = \left\{ \sum_{i \in s_I} f_i^{-1} \left(\sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_{ih}} y_k \right) \right\} \Big/ \left\{ \sum_{i \in s_I} f_i^{-1} \left(\sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_{ih}} x_k \right) \right\}.$$

Note that, if $x_k = 1$ for all k, the estimator (8.10) will still be different from (8.9). The estimated variances follow easily from (8.6)–(8.8), in observing that $\pi_{ak} = f_I f_i$ and that the required quantities are $\check{a}_k = y_k / \pi_{ak}$ in the case of (8.9) and $\check{a}_k = (y_k - b x_k)/\pi_{ak}$ in the case of (8.10).

It cannot be enough emphasized that in practice we must always be conscious of the possibility of, not to say the high likelihood of, misspecification of the RHG model. For example, even if groups do exist within which the individual response probabilities are essentially equal, these 'true' groups may not coincide with the groups assumed in formulating the RHG model.

In other words, the estimation procedure described here (and most other procedures for estimation when there is nonresponse) requires an *assumed response model,* to be abbreviated ARM. (Here we consider only models that involve RHG's, but in a more general setting, the ARM may have a structure not involving the group assumption.) The ARM decision is crucial, for it will determine the estimator formula, and thereby it determines the numerical value of the estimate as well as the confidence interval estimate of t ultimately released by the statistician. With some other ARM, (perhaps markedly) different point and interval estimates would be published by the statistical agency.

In settling on a certain ARM, the statistician believes that, with due consideration given, point and interval estimates produced under his assumption will be reasonably well 'nonresponse adjusted'. But he would be naive to consider his ARM a 'true response model'. As Kalton (1983) puts it,

> sampling practitioners do not believe that the nonresponse models on which their adjustments are based hold exactly: they simply hope that they are improvements on the model of data missing at random.

9 A simulation study

If the assumed RHG model is false, the estimators (8.2)–(8.5) will be biased to an extent that depends on the degree of model breakdown. We carried out a Monte Carlo experiment to get some idea of the consequences of a falsely specified ARM. We hypothesized that the regression estimator $\hat{t}_{c1\text{REG}}$, which uses an auxiliary vector $\mathbf{x}$ in addition to weighting, would show a more robust behaviour (less bias, closer to nominal confidence levels) under ARM breakdown than $\hat{t}_{c\pi^*}$, which uses weighting only. If this were true, the important consequence is a strong incentive, whenever possible in situations with nonresponse, to observe one or more strongly explantory x-variables for units k in the intended sample s, and to incorporate these variables in a regression estimator, (8.3), or (8.5). Not only would this reduce variance if the ARM is true or almost

true, but, more importantly, it would improve the chances for correct confidence statements when the ARM is false. Our hypothesis was confirmed by the Monte Carlo experiment, of which a more detailed account is given by Särndal & Swensson (1985).

Repeated simple random samples s, of size $n = 400$ each, were drawn from a real population of $N = 1227$ Swedish households. Once selected, each sample was exposed to simulated unit nonresponse. The true response mechanism, chosen by us and used to generate unit nonresponse, conformed to an RHG model with four RHG's (four household types). The ARM (that is, the RHG model used as a basis for calculating the estimator, the variance estimator and the confidence interval) was stated as a RHG model with: (a) four groups, those of the true ARM: (b) two groups; and (c) one group. Here, (b) and (c) are false ARM's. The regression estimator (8.10) was used in the simulation.

In summary, the Monte Carlo study led to the following conclusions.

When the ARM *was true*:

(1) both $\hat{t}_{c\pi^*}$ and $\hat{t}_{c1\text{REG}}$ were essentially unbiased;
(2) in both cases, the variance estimators were essentially unbiased;
(3) the variance of $\hat{t}_{c1\text{REG}}$ was considerably smaller than that of $\hat{t}_{c\pi^*}$, the reduction lying in the second variance component and being due to the fact that y was fairly well explained by x;
(4) the achieved coverage rates of the confidence intervals were near the nominal rate for both estimators, the average length of interval being considerably shorter for $\hat{t}_{c1\text{REG}}$.

These results confirm what theory leads us to believe.

When the ARM *was false*:

(1) both $\hat{t}_{c\pi^*}$ and $\hat{t}_{c1\text{REG}}$ were now biased, but $\hat{t}_{c1\text{REG}}$ had much smaller bias;
(2) the respective variance estimators were fairly insensitive to ARM breakdown;
(3) the variance was again considerably smaller for the regression estimator $\hat{t}_{c1\text{REG}}$ than for the weighting-only estimator $\hat{t}_{c\pi^*}$;
(4) the achieved coverage rates were much closer to nominal levels for $\hat{t}_{c1\text{REG}}$ than for $\hat{t}_{c\pi^*}$.

10 Discussion

Let us examine a statement by Oh & Scheuren (1983), in which 'subgroups' refers to our RHG's:

> A seemingly robust approach is to choose the subgroups such that for the variable(s) to be analyzed, the within-group variation for nonrespondents is small (and the between-group mean differences are large); then, even if the response mechanism is postulated incorrectly, the bias impact will be small.

In our opinion, one must separate the role of the RHG's from that of other information (the x-variables) recorded for $k \in s$. Two different concepts are involved. The sole criterion for the RHG's should be that they eliminate bias as far as possible. Every effort should be made, and all prior knowledge used, to settle on groups likely to display response homogeneity. But in addition it is imperative to measure, for $k \in s$, a concomitant vector $\mathbf{x}_k$, that will reduce variance and give added protection against bias. Groups that eliminate or reduce bias are not necessarily variance reducing, and, contrary to what the quotation seems to suggest, the criterion of maximizing between-to-within variation in y does not necessarily create groups that work well for removing bias.

In summary, we find that:

(1) in order to *eliminate* bias due to nonresponse, it is vital to identify the true response

model; as this is usually impossible, bias can be *greatly reduced* if powerful explanatory x-variables can be found and incorporated in a regression-type estimator;

(2) a second reason to incorporate such x-variables into the estimator is that the inevitable increase in variance caused by the nonresponse, 'the second variance component', is kept at low levels.

Acknowledgement

The authors are grateful to the referees, whose constructive comments led to improvement in the paper. The work of C.-E. Särndal was supported by the Natural Sciences and Engineering Research Council of Canada.

References

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.

Cassel, C.M., Särndal, C.E. & Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys,* **3,** Ed. W.G. Madow and I. Olkin, pp. 143–160. New York: Academic Press.

Chaudhuri, A. & Adhikary, A.K. (1983). On optimality of double sampling strategies with varying probabilities. *J. Statist. Plan. Inf.* **8,** 257–265.

Cochran, W.G. (1977). *Sampling Techniques,* 3rd ed. New York: Wiley.

Dalenius, T. (1983). Some reflections on the problem of missing data. In *Incomplete Data in Sample Surveys,* **3,** Ed. W.G. Madow and I. Olkin, pp. 411–413. New York: Academic Press.

Kalton, G. (1983). Models in the practice of survey sampling. *Int. Statist. Rev.* **51,** 175–188.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *J. Am. Statist. Assoc.* **33,** 101–116.

Oh, H.L. & Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys,* **2,** Ed. W.G. Madow, I. Olkin and D.B. Rubin, pp. 143–184. New York: Academic Press.

Platek, R. & Gray, G.B. (1983). Imputation Methodology. In *Incomplete Data in Sample Surveys,* **2,** Ed. W.G. Madow, I. Olkin and D.B. Rubin, pp. 255–293. New York: Academic Press.

Raj, D. (1968). *Sampling theory.* New York: McGraw-Hill.

Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika* **60,** 125–133.

Särndal, C.E. (1982). Implications of survey design for generalized regression estimation of linear functions. *J. Statist. Plan. Inf.* **7,** 155–170.

Särndal, C.E. & Swensson, B. (1985). Incorporating nonresponse modelling in a general randomization theory approach. *Bull. Int. Statist. Inst.* (45:th session), LI-3, 15.2.1-16.

Singh, S. & Singh, R. (1979). On random non-response in unequal probability sampling. *Sankhyā* C **41,** 127–137.

Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statist. Tidskrift* **11,** 278–285.

Résumé

Cet article comporte deux parties: (a) théorie d'estimation pour l'échantillonnage en deux phases, et (b) théorie d'estimation dans une enquête avec de la non-réponse. Dans les deux cas, un premier échantillonnage est suivi d'un "sous-échantillonnage". Ce sous-échantillonnage est régi dans le premier cas par une loi connue (un plan d'échantillonnage fixé), dans le deuxième cas par une loi hypothétique et inconnue (un modèle du mécanisme de la réponse). Malgré ces différences, on pourra, comme le démontre cet article, tirer profit du premier cas pour traiter le deuxième.

Dans la première partie nous présentons, pour l'échantillonnage en deux phases, des résultats généraux concernant certains estimateurs par la régression. Nous étudions surtout la question d'estimation de la variance de ces estimateurs, avec les intervalles de confiance qui s'ensuivent.

Ensuite nous transplantons ces résultats, après des modifications nécessaires mais mineures, dans le cadre d'une enquête avec de la non-réponse. Nos résultats montrent que les estimateurs par la régression proposés dans l'article donnent lieu à des intervalles de confiance assez robustes, même si l'hypothèse concernant le mécanisme de la réponse est dans une certaine mesure fausse.

[*Received October* 1985, *revised April* 1987]