SCB STATISTISKA CENTRALBYRÅN

1985-12-05

PROMEMORIOR FRÅN P/STM

NR 20

A GENERAL VIEW OF ESTIMATION FOR TWO PHASES OF SELECTION

AV CARL-ERIK SÄRNDAL OCH BENGT SWENSSON

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1985:20. A general view of estimation for two phases of selection / Carl-Erik Särndal; Bengt Swensson.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1985-20

A GENERAL VIEW OF ESTIMATION FOR TWO PHASES OF SELECTION

by

Carl-Erik Särndal
Université de Montréal

Bengt Swensson
Statistics Sweden
and
University of Örebro

SUMMARY

In developing estimation methods for the nonresponse situation, survey samplers have largely failed to profit from a striking analogy of which most of them are nevertheless aware: the classical formulation of the two-phase sampling situation offers a near perfect basis for a methodology for the nonresponse situation. More precisely, in both cases, an initial sample, s, is drawn, certain information (but not the variable of interest) can be observed for the units in s, a subsample, r, is then realized (voluntarily in one situation, involuntarily in the other), and the variable of interest is measured for units in the ultimate sample r only. However, at closer look it is not surprising that the analogy has not been taken advantage of. The theory of two-phase sampling has been insufficiently developed, so far, to handle more complex sampling designs and estimators. In other words, while a perfect natural bridge exists between two situations, not much of value has been close at hand, so far, to transport over the bridge. In this paper we lay certain groundwork on the side that should first be developed: the first part of our work presents a more general two-phase sampling theory, allowing complex (arbitrary) sampling designs in each phase and/or more advanced (regression type) estimators, as well as variance estimators and confidence intervals computed from the ultimate sample r. The second part of the paper shows how these results fit into the nonresponse situation, more precisely, in the case where nonresponse is adjusted for by the often used model of homogeneous response probability within subgroups. The parallel with two-phase sampling produces the desired formulas for calculating estimated variances and confidence intervals for estimates affected by nonresponse. Concluding Part II of the paper is a Monte Carlo study emphasizing that concomitant variables used to model the response mechanism must be distinguished from equally important (but conceptually different) concomitant variables whose immediate role is to be variance reducing. We

want to create awareness of the fact that two different types of variables are involved; their respective roles must not be confused. We study the validity (the empirical coverage rate) of our confidence statements, which is of particular interest when the response modelling breaks down (as is inevitably the case in practice, to a greater or smaller extent). We observe that the robustness of the confidence statements is very much improved by the presence in the estimator of the kind of concomitant variable that we call variance reducing. This type of variable therefore becomes important for the added reason of robustness.

CONTENTS PART I

1.	INTRODUCTION	1
2.	THE TWO-PHASE SAMPLING SETUP	3
3.	ESTIMATION BY π^*-EXPANDED SUMS	5
4.	APPLICATION: TWO-PHASE SAMPLING FOR STRATIFICATION	7
5.	REGRESSION ESTIMATION IN TWO-PHASE SAMPLING	13
6.	REGRESSION ESTIMATION IN TWO-PHASE SAMPLING FOR STRATIFICATION	21
7.	CONCLUSION	23
	REFERENCES	24

CONTENTS PART II

1.	INTRODUCTION TO PART II	25
2.	THEORETICAL RESULTS	28
3.	A SIMULATION STUDY	36
4.	DISCUSSION	42
5.	SOFTWARE AT STATISTICS SWEDEN FOR POINT ESTIMATES AND STANDARD ERRORS	43
	REFERENCES	44

A GENERAL VIEW OF ESTIMATION FOR TWO PHASES OF SELECTION

PART I: RANDOMIZED SUBSAMPLE SELECTION (TWO-PHASE SAMPLING)

by

Carl-Erik Särndal
Université de Montréal

Bengt Swensson
Statistics Sweden
and
University of Örebro

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 should be possible, in the nonresponse situation, to profit directly from two-phase sampling theory. To date, this advantage has not been exploited to any great extent. One reason can perhaps be found in the present state of relative underdevelopment of two-phase sampling theory, for which the basic results were presented long ago. But there has been rather few extensions and refinements of two-phase sampling theory, which, consequently, has been insufficiently equipped to handle the practical problems of the nonresponse situation. Our paper may be seen as a step towards filling the gap. In Part I of the paper, we develop general principles for estimation in situations that involve two-phase sampling. This term will be used with its standard meaning in survey sampling, that is to say that a controlled, randomized scheme is used for subsampling of an initial probability sample. In Part II, we transplant the results, with necessary minor

modifications, into the case of involuntary subsampling caused by nonresponse. We consider estimators constructed with special effort to control the nonresponse bias, so that confidence statements will not be severely distorted. This is obtained by weighting adjustment combined with the explicit use in the estimator of auxiliary variables. Considerable emphasis is put in this paper on estimation of the precision (for use in confidence intervals, for example) of estimators appropriate for the nonresponse situation. The variance estimators that we suggest for the nonresponse case follow directly from our two-phase sampling theory.

The standard two-phase sampling situation calls for the selection of a rather large first phase sample, s, and the collection of some inexpensive information for the units k in s. More formally, say that this consists in recording the value $\underset{\sim}{x}_k$ of the vector $\underset{\sim}{x}$ for $k \in s$. The second phase sampling procedure consists in drawing a subsample r from s. For the units $k \in r$, one records the value y_k of the study variable, y, more expensive to measure than $\underset{\sim}{x}$. Now $\underset{\sim}{x}_k$ can be used (a) to create a highly efficient design for drawing the second phase sample, and/or (b) to create a highly efficient (regression type) estimator of the characteristic of interest, say, the population total of y. Conforming to this description, we present in Section 2 a general view of two-phase sampling allowing an arbitrary, possibly "complex" sampling design in each of the two phases. In particular, the first phase design may be a two- or multi-stage design.

Let us compare with the nonresponse situation: An "intended sample", s, is drawn according to a given (but arbitrary) sampling design, which may be "complex", e.g., involving two or more stages of sampling. The sample s is affected by unit nonresponse, resulting in a measurement y_k only for $k \in r$, the subset of responding units $(r \subset s)$. It is common now to assume that r results from s through a probabilistic "response mechanism", corresponding to the second phase sampling design in the two-phase situation. The difference is that the response mechanism

is ordinarily unknown, forcing the statistician to make assumptions about it, whereas in the two-phase situation, the second phase sampling obeys a known probability distribution, chosen and executed by the statistician.

2- *THE TWO-PHASE SAMPLING SETUP*

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. For example, Raj (1968, p. 139-152) considers topics such as two-phase sampling for difference estimation, for pps estimation, for (biased or unbiased) ratio estimation, for simple regression estimation, for stratification, etc.

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 and 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 k:th 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 $p_a(\cdot)$, such that $p_a(s)$ is the probability of choosing s . The inclusion probabilities are defined by

$$\pi_{ak} = \sum_{s \ni k} p_a(s) \; ; \; \pi_{ak\ell} = \sum_{s \ni k \& \ell} p_a(s) \; ,$$

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

(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 $p(\cdot|s)$, such that $p(r|s)$ is the conditional probability of choosing r . The conditional inclusion probabilities are defined by

$$\pi_{k|s} = \sum_{r \ni k} p(r|s) \; ; \; \pi_{k\ell|s} = \sum_{r \ni k \& \ell} p(r|s) \; ,$$

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

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.

In turning now to estimation, we note that the Horvitz-Thompson estimator, despite its great flexibility for unequal probability sampling designs, does not fit (except in some special cases) our general two phase sampling setup. The reasons are as follows:

Let $p(r)$ be the probability that the set r is realized by the procedure defined by (a) and (b) above. If it were possible to determine the unconditional inclusion probabilities

$$\pi_k = \sum_{r \ni k} p(r) = \sum_{r \ni k} \sum_{s \supset r} p_a(s)p(r|s) , \qquad (2.1)$$

we might consider the Horvitz-Thompson estimator of $t = \Sigma_U y_k$, that is, $\hat{t}_{HT} = \Sigma_r y_k/\pi_k$. Now, from (2.1),

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

To determine the π_k, we must thus know $\pi_{k|s}$ for <u>every</u> 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. Thus $\pi_{k|s}$ will be known <u>only</u> for this particular s. To stress this point still more: It is not until after s has been drawn and information has been gathered about the units in s (and no other units) that an exact specification of the second phase design $p(\cdot|s)$ can be given. Consequently the π_k (which may depend on all the other possible s) can not be determined in many real-life situations. The Horvitz-Thompson estimator is thus impractical. We shall construct a more useful design unbiased estimator.

3. *ESTIMATION BY π^*-EXPANDED SUMS*

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

$$\pi_k^* = \pi_{ak}\pi_{k|s} \; ; \; \pi_{k\ell}^* = \pi_{ak\ell}\pi_{k\ell|s}$$

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

$$\breve{y}_k = y_k/\pi_{ak} \; ; \; \breve{\breve{y}}_k = \breve{y}_k/\pi_{k|s} = y_k/\pi_k^* \; ; \; \breve{\Delta}_{ak\ell} = \Delta_{ak\ell}/\pi_{ak\ell} \; ; \; \breve{\Delta}_{k\ell}^* = \Delta_{k\ell}^*/\pi_{k\ell}^* .$$

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, $\Sigma\Sigma_A\, c_{k\ell}$ means $\sum_{k\in A}\sum_{\ell\in A} c_{k\ell}$.) Despite a certain similarity, the Horvitz-Thompson estimator is in general different from the π*es estimator.

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,

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

Its design variance is

$$V(\hat{t}_{\pi*}) = \Sigma\Sigma_U \Delta_{ak\ell} \breve{y}_k \breve{y}_\ell + E_a\{\Sigma\Sigma_s \Delta_{k\ell|s} \breve{\breve{y}}_k \breve{\breve{y}}_\ell\} \,, \tag{3.2}$$

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

$$\hat{V}(\hat{t}_{\pi*}) = \Sigma\Sigma_r \breve{\Delta}_{ak\ell} \breve{y}_k \breve{y}_\ell/\pi_{k\ell|s} + \Sigma\Sigma_r \Delta_{k\ell|s} \breve{\breve{y}}_k \breve{\breve{y}}_\ell/\pi_{k\ell|s} = \Sigma\Sigma_r \breve{\breve{\Delta}}_{k\ell} \breve{\breve{y}}_k \breve{\breve{y}}_\ell \,. \;\square \tag{3.3}$$

Note that the second component of (3.2) must be left in the form of an expected value, since the $\Delta_{k\ell|s}$ may depend on s.

The following observation is pertinent to Result 0 as well as to other similarly presented results below: The variance consists of a "first component" due to sampling in phase one, and a "second component" due to sampling in phase two. Correspondingly, the variance estimator is composed of an "estimated first component" and an "estimated second component". Note that if no subsampling is carried out (so that only the first phase prevails), then the (estimated) second component vanishes.

If $\hat{t}$ stands for an estimator of t ($\hat{t}$ 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.

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

$$\hat{t}_{\pi*} - t = \underbrace{\Sigma_s \check{y}_k - \Sigma_U y_k}_{A_s} + \underbrace{\Sigma_r \check{\check{y}}_k - \Sigma_s \check{y}_k}_{B_r} . \tag{3.4}$$

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

$$V(B_r|s) = \Sigma\Sigma_s \Delta_{k\ell|s} \check{\check{y}}_k \check{\check{y}}_\ell \; ; \; V_a(A_s) = \Sigma\Sigma_U \Delta_{ak\ell} \check{\check{y}}_k \check{\check{y}}_\ell ,$$

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

$$E_a E(\Sigma\Sigma_r \, c_{k\ell}/\pi_{k\ell|s}|s) = E_a(\Sigma\Sigma_s c_{k\ell}) = \Sigma\Sigma_U \pi_{ak\ell} c_{k\ell} \tag{3.5}$$

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

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, $h = 1,\ldots,H_s$. 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 (short for stratified simple random sampling), where r_h is drawn from s_h by simple random sampling (SI); Case STBE (for 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 non-response theory given in Part II of the paper. 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 non-identical 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,

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

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

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

The variance, obtained from (3.2), is

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

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)

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

where $S^2_{\check{y}r_h}$ is the variance of $\check{y}_k$ in the set r_h. 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

$$\hat{t}_{\pi*} = N \sum_{h=1}^{H_s} w_h \bar{y}_{r_h} = N\hat{\bar{y}}_U \ ; \qquad (4.5)$$

its design variance, from (4.3), is

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

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

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

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

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

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

Note that these results do not require the same stratification principle for every s. □

EXAMPLE 2. Conceptually different from Example 1 is the case where the same stratification principle applies to every conceivable first phase sample s. We can then say that there exists H "conceptual strata" $U_1,\ldots,U_h,\ldots,U_H$ in the

population, for example, a fixed set of age-sex groups. Although not identified prior to drawing s, these groups (of unknown sizes) define the predetermined principle by which every possible s would be stratified, if selected.

That is, $s_h = s \cap U_h$, $h = 1,\ldots,H$. Let n_h be the size of s_h. Further suppose that a subsample r_h of size $m_h = \nu_h n_h$ is drawn from s_h by SI, $h = 1,\ldots,H$, where $\nu_1,\ldots,\nu_H$ are a priori fixed positive fractions. The results (4.5) and (4.7) will apply (although with H_s replaced by H, since the number of strata is now constant for all s.) The variance (4.6) can be stated on a more explicit form, see Rao (1973) and Cochran (1977, p. 328),

$$V(\hat{t}) = N^2(1-f)S_{yU}^2/n + \frac{N^2}{n}\sum_{h=1}^{H} W_h S_{yU_h}^2 \left(\frac{1}{\nu_h} - 1\right), \qquad (4.8)$$

where $W_h = N_h/N$ is the relative size of U_h. Note that in this example, there is a non-zero probability that an empty set s_h ($n_h=0$) occurs for at least one $h = 1,\ldots,H$. (In Example 2, this problem does not arise, since the strata are formed only after s has been realized, and every stratum is necessarily subsampled.) The estimator $\hat{t} = N\bar{\hat{y}}_U$ is undefined if s_h is empty for some h. If the probability of this event were ignorable, (4.8) would be an exact variance; otherwise, it holds only approximately. □

Case STBE. The subsample r_h (from s_h) is realized as follows. The inclusion or non-inclusion 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 $\underset{\sim}{m} = (m_1,\ldots,m_h,\ldots,m_{H_s})$, the vector of realized counts. Define

$$\pi_{k|s,\underset{\sim}{m}} = \Pr(k\epsilon r|s,\underset{\sim}{m}) \; ; \; \pi_{k\ell|s,\underset{\sim}{m}} = \Pr(k\&\ell\epsilon r|s,\underset{\sim}{m}) \; .$$

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

$$\pi_{k|s,\underset{\sim}{m}} = \pi_{kk|s,\underset{\sim}{m}} = m_h/n_h = f_h \; ; \; \pi_{k\ell|s,\underset{\sim}{m}} = f_h(m_h-1)/(n_h-1) \quad (k\neq\ell) \tag{4.9}$$

while $\pi_{k\ell|s,\underset{\sim}{m}} = f_h f_{h'}$ 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"

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

with the variance expression

$$V(\hat{t}_{c\pi*}) = \Sigma\Sigma_U \Delta_{ak\ell}\check{y}_k\check{y}_\ell + E_a E_{\underset{\sim}{m}}\{\sum_{h=1}^{H_s} n_h^2(1-f_h)S^2_{\check{y}s_h}/m_h\} \tag{4.11}$$

where $E_{\underset{\sim}{m}}(\cdot)$ indicates expectation over all realizations $\underset{\sim}{m}$, 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

$$\hat{V}(\hat{t}_{c\pi*}) = \Sigma\Sigma_r \check{\Delta}_{ak\ell}\check{y}_k\check{y}_\ell/\pi_{k\ell|s,\underset{\sim}{m}} + \sum_{h=1}^{H_s} n_h^2(1-f_h)S^2_{\check{y}r_h}/m_h \tag{4.12}$$

where $\pi_{k\ell|s,\underset{\sim}{m}}$ 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 $\pi_{k|s} = \pi_{k|s,\underset{\sim}{m}}$. 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 $\pi_{k\ell|s} = \pi_{k\ell|s,m}$. (Here, (4.12) is an approximation; see below.) This fortuitous identity of the two variance estimators will be used again later. □

EXAMPLE 3. 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). □

In Case STBE, the estimator (4.10) is approximately unbiased for t, and (4.11) is an approximate variance expression. Let us substantiate these claims. (This digression from the main theme of the paper can be skipped at first reading.) The estimator $\hat{t}_{c\pi*}$ is undefined when $\overline{A}_{1s}$ = "not A_{1s}" occurs. This need not be of great concern in practice, for one would set up the groups to keep $Pr(\overline{A}_{1s})$ close to zero. But from the formal perspective, it is unsatisfactory that the estimator could possibly be undefined. Let us extend its definition to cover all possible outcomes. Set $t_h = \Sigma_{s_h} \check{y}_k$, a quantity which we estimate (given s) by $\hat{t}_h = f_h^{-1} \Sigma_{r_h} \check{y}_k$ if $m_h > 0$ and $\hat{t}_h = t_{ho}$ if $m_h = 0$, where t_{ho} is any given constant (which may be taken as zero). We now estimate $t = \Sigma_U y_k$ by

$$\hat{t} = \sum_{h=1}^{H_s} \hat{t}_h ,$$

which is always defined and equal to $\hat{t}_{c\pi*}$ if A_{1s} occurs. Set

$$\varepsilon_{ho} = Pr(m_h=0|s) \; ; \; d_{ho} = t_{ho} - t_h . \tag{4.13}$$

Now, $E(\hat{t}_h|s,m_h) = t_{ho}$ if $m_h = 0$, $E(\hat{t}_h|s,m_h) = t_h$ if $m_h > 0$, and it is easily seen that

$$E(\hat{t}) - t = E_a\left(\sum_{h=1}^{H_s} \varepsilon_{ho} d_{ho}\right) , \tag{4.14}$$

which is the bias of $\hat{t}$. Under the STBE design, $\varepsilon_{ho} = (1-\theta_{hs})^{n_h}$. If the group count n_h is large, and θ_{hs} not too near zero, ε_{ho} will be near zero. If this is true for all groups (which implies that $\bar{A}_{1s}$ has near zero probability), then the bias will be negligible.

A similarly detailed analysis can be carried out to see that the variance (4.11) is in effect the variance of the "extended" estimator $\hat{t}$, in the limit as the ε_{ho} approach zero. Finally, the proposed variance estimator (4.12) is natural, given the composition of the variance expression (4.11). Despite the approximative nature of the procedure summarized by (4.10) - (4.12), a confidence interval with essentially correct level will be obtained, for most practical situations, by means of $\hat{t}_{c\pi^*}$, with the variance estimator (4.12).

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 (Section 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 at 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 $\underset{\sim}{x}_k = (x_{1k},\ldots,x_{qk})'$ of the auxiliary vector $\underset{\sim}{x} = (x_1,\ldots,x_q)$ is recorded for the units $k \in s$.

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

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

Let us first develop Situations 1 and 2, using extensions of the regression approach; Cassel, Särndal and Wretman (1976); Särndal (1982). We assume the existence of a strong regression relationship between y_k and $\underset{\sim}{x}_k$. Then, although y_k may be observed in a smallish second phase sample only, the relative scarcity of the y-information can to a large degree be compensated for by using a regression estimator. We model the relationship of y on $\underset{\sim}{x}$ by assuming that the y_k are independent (throughout) with $E_\xi(y_k) = \underset{\sim}{x}_k'\underset{\sim}{\beta}$, $V_\xi(y_k) = \sigma_k^2$, where ξ indicates moments with respect to the model. Here, $\underset{\sim}{\beta}$ is unknown. The σ_k^2 are known up to multiplicative constants that vanish in the calculation of the estimator $\underset{\sim}{b}$ given below by (5.2). If the entire population were observed, one would estimate $\underset{\sim}{\beta}$ and form residuals according to

$$\underset{\sim}{B} = (\Sigma_U \underset{\sim}{x}_k\underset{\sim}{x}_k'/\sigma_k^2)^{-1}\Sigma_U \underset{\sim}{x}_k y_k/\sigma_k^2 \; ; \; E_k = y_k - \underset{\sim}{x}_k'\underset{\sim}{B} \; . \tag{5.1}$$

However, $(y_k,\underset{\sim}{x}_k)$ is observed for $k \in r$ only, and the k:th unit carries the weight π_k^{*-1}. Let us therefore estimate $\underset{\sim}{B}$ by

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

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

$$e_k = y_k - \hat{y}_k = y_k - \underset{\sim}{x}_k'\underset{\sim}{b} \; , \; k \in r \; , \tag{5.3}$$

(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 $\underset{\sim}{x}_k$ is recorded for $k \in s$, an approximately design unbiased estimator of $t = \Sigma_U y_k$ is given by

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

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

$$AV(\hat{t}_{1REG}) = \Sigma\Sigma_U \Delta_{ak\ell} \check{y}_k \check{y}_\ell + E_a\{\Sigma\Sigma_s \Delta_{k\ell|s} \check{\check{E}}_k \check{\check{E}}_\ell\} \ . \tag{5.5}$$

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

$$\hat{V}(\hat{t}_{1REG}) = \Sigma\Sigma_r \check{\Delta}_{ak\ell} \check{y}_k \check{y}_\ell/\pi_{k\ell|s} + \Sigma\Sigma_r \Delta_{k\ell|s} \check{\check{e}}_k \check{\check{e}}_\ell/\pi_{k\ell|s} \ . \ \square \tag{5.6}$$

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

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

where $y_k^o = \underset{\sim}{x}_k' \underset{\sim}{B}$ has replaced $\hat{y}_k = \underset{\sim}{x}_k' \underset{\sim}{b}$. The advantage gained is that $\hat{t}^o_{1REG}$ is extremely simple to analyze. We have

$$\hat{t}_{1REG} - t \doteq \hat{t}^o_{1REG} - t = \underbrace{\Sigma_s \check{y}_k - \Sigma_U y_k}_{A_s} + \underbrace{\Sigma_r \check{\check{E}}_k - \Sigma_s \check{E}_k}_{B_r'} \tag{5.7}$$

where $\check{E}_k = E_k/\pi_{ak}$; $\check{\check{E}}_k = E_k/\pi_k^* = \check{E}_k/\pi_{k|s}$. It follows that $E(\hat{t}_{1REG}) - t \doteq E(\hat{t}^o_{1REG}) - t = 0$, so that $\hat{t}_{1REG}$ 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 $\underset{\sim}{b}$ was approximated by $\underset{\sim}{B}$). In obtaining an estimated variance,

$E_k = y_k - \underset{\sim}{x}_k'\underset{\sim}{B}$ can not be used since it contains the unknown $\underset{\sim}{B}$. Instead, substitute $e_k = y_k - \underset{\sim}{x}_k'\underset{\sim}{b}$, where $\underset{\sim}{b}$ 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 $\underset{\sim}{x}_k$ is recorded for all $k \in U$, an approximately design unbiased estimator of $t = \Sigma_U y_k$ is given by

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

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

$$AV(\hat{t}_{2REG}) = \Sigma\Sigma_U \Delta_{ak\ell} \check{E}_k \check{E}_\ell + E_a\{\Sigma\Sigma_s \Delta_{k\ell|s} \check{\check{E}}_k \check{\check{E}}_\ell\} . \tag{5.9}$$

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

$$\hat{V}(t_{2REG}) = \Sigma\Sigma_r \check{\Delta}_{ak\ell} \check{e}_k \check{e}_\ell/\pi_{k\ell|s} + \Sigma\Sigma_r \Delta_{k\ell|s} \check{\check{e}}_k \check{\check{e}}_\ell/\pi_{k\ell|s} = \Sigma\Sigma_r \Delta_{k\ell}^* \check{\check{e}}_k \check{\check{e}}_\ell . \tag{5.10}$$

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

Compare the three variance expressions (3.2), (5.5) and (5.9). They correspond to three different levels of $\underset{\sim}{x}$-information: None at all; $\underset{\sim}{x}_k$ known only for $k \in s$; and $\underset{\sim}{x}_k$ known for all $k \in U$, respectively. The variance components reflect this progression: In (3.2), no regression residuals enter into the variance components. In (5.5), residuals enter in the second (but not the first) variance component, since the $\underset{\sim}{x}$-information extends only to the first phase sample. Finally, in (5.9), the residuals enter into both variance components, since $\underset{\sim}{x}_k$ is then known throughout the population. Clearly, "residualization" of a variance component will normally reduce its numerical value.

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

As a model for the relationship between $\underset{\sim}{z}_k$ and y_k, assume that the y_k are independent with $E_{\xi_1}(y_k) = \underset{\sim}{z}_k'\underset{\sim}{\beta}_1$; $V_{\xi_1}(y_k) = \sigma_{1k}^2$ where $\underset{\sim}{\beta}_1$ is to be estimated and the σ_{1k}^2 are new model variances (in the simplest case, $\sigma_{1k}^2 = \sigma^2$, for all k). If all y_k, $k \in U$, were observed, the $\underset{\sim}{\beta}_1$-estimator and the residuals would be

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

But the information about y_k is less extensive, so we must estimate 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:

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

The second method, slightly more complicated, recognizes that the known $\underset{\sim}{x}_k$-values for $k \in s$ makes 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

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

Now, the second estimator of $\underset{\sim}{\beta}_1$ is taken as

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

(Both (5.12) and (5.13) are in fact consistent estimators of $\underset{\sim}{B}_1$.)

Whether (5.12) or (5.13) is used, we calculate predicted values as

$$\hat{y}_{1k} = \underset{\sim}{z}_k'\underset{\sim}{b}_1 \quad \text{for} \quad k \in U$$

(since $\underset{\sim}{z}_k$ is known for all $k \in U$) , and residuals according to

$$e_{1k} = y_k - \underset{\sim}{z}_k'\underset{\sim}{b}_1 \quad \text{for} \quad k \in r \tag{5.14}$$

(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 $\underset{\sim}{x}_k$ is recorded for $k \in s$ and $\underset{\sim}{z}_k$ for $k \in U$, an approximately design unbiased estimator of $t = \Sigma_U y_k$ is

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

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

$$AV(\hat{t}_{3REG}) = \Sigma\Sigma_U \Delta_{ak\ell} \check{E}_{1k} \check{E}_{1\ell} + E_a\{\Sigma\Sigma_s \Delta_{k\ell|s} \check{\check{E}}_k \check{\check{E}}_\ell\} . \tag{5.16}$$

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

$$\hat{V}(\hat{t}_{3REG}) = \Sigma\Sigma_r \check{\Delta}_{ak\ell} \check{e}_{1k} \check{e}_{1\ell}/\pi_{k\ell|s} + \Sigma\Sigma_r \Delta_{k\ell|s} \check{\check{e}}_k \check{\check{e}}_\ell/\pi_{k\ell|s} . \quad \square \tag{5.17}$$

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

Comparing the respective variances of the three regression estimators, (5.5), (5.9) and (5.16), we note that the second variance component, expressed in the residuals E_k , is common to the three regression estimators. Differences

occur in the first variance component, which is "residualized" in the case of $\hat{t}_{2REG}$ and $\hat{t}_{3REG}$, but not in the case of $\hat{t}_{1REG}$. The first component will therefore ordinarily be smaller for $\hat{t}_{2REG}$ and $\hat{t}_{3REG}$ than for $\hat{t}_{1REG}$.

EXAMPLE 4. 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 i:th psu in the first stage sample, and π_{ki} the probability of choosing the k:th ssu of the i:th 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

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

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

$$b = (\Sigma_r \breve{y}_k)/(\Sigma_r \breve{x}_k) \; ; \; e_k = y_k - bx_k \; , \; k \in r \; , \tag{5.19}$$

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

In addition, Situation 3 requires a second model, which we take to be the "trivial" one with

$$E_\xi(y_k) = \beta \; ; \; V_\xi(y_k) = \sigma^2 \; , \; k \in U \; . \tag{5.20}$$

(This corresponds to $z_k = 1$ for all k. For estimation of t, this model does

require some additional information, namely, the knowledge of $\Sigma_U z_k = N$, the population size). Results 1, 2 and 3 yield the regression estimators

$$\hat{t}_{1REG} = (\Sigma_s \check{x}_k)b \; ; \; \hat{t}_{2REG} = (\Sigma_U x_k)b \; ; \; \hat{t}_{3REG} = N\tilde{x}_s b$$

with

$$\tilde{x}_s = (\Sigma_s \check{x}_k)/\hat{N} \; ; \; \hat{N} = \Sigma_s \, 1/\pi_{ak} \; ,$$

and b given by (5.19). We have used (5.13) in deriving $\hat{t}_{3REG}$.

Both $\hat{t}_{1REG}$ and $\hat{t}_{3REG}$ require knowledge of the values x_k for $k \in s$. In addition, $\hat{t}_{3REG}$ requires that N be known. In cases where two stage sampling must be resorted to, N is normally unknown; consequently, for estimating the total t , it is $\hat{t}_{1REG}$ rather than $\hat{t}_{3REG}$ that will be used. However, if N were known, $\hat{t}_{3REG}$ would be preferred because of a variance advantage. Now $\hat{t}_{2REG}$ requires the x-total for the entire population U ; correspondingly, this estimator will ordinarily have the smallest variance of the three. The estimated variances are obtained from Results 1 to 3, where $e_k = y_k - bx_k$ and $e_{1k} = y_k - b\tilde{x}_s$.

The choice between $\hat{t}_{1REG}$ and $\hat{t}_{3REG}$ appears in a different perspective if an estimate is sought not of the total t but rather of the mean $\bar{y} = t/N$. The estimators of $\bar{y}$ formed (by division by N) are

$$\hat{\bar{y}}_{1REG} = \hat{t}_{1REG}/N = N^{-1}(\Sigma_s x_k)b \; ; \; \hat{\bar{y}}_{3REG} = \hat{t}_{3REG}/N = \tilde{x}_s b \; .$$

The latter formula has two advantages: it does not depend on N , and the variance is usually smaller. Thus, in two stage sampling with N unknown, it is $\hat{\bar{y}}_{3REG}$ rather than $\hat{\bar{y}}_{1REG}$ that will be used. □

6- *REGRESSION ESTIMATION IN TWO-PHASE SAMPLING FOR STRATIFICATION*

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

Case STSI. Results 1 to 3 apply straightforwardly, with $\pi_{k|s}$ and $\pi_{k\ell|s}$ determined by (4.1). That is, the k:th 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

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

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

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

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

Case STBE. The transition from Case STSI to Case STBE is done by conditioning on $\underset{\sim}{m}$ as in Section 4: $\pi_{k|s,\underset{\sim}{m}}$ and $\pi_{k\ell|s,\underset{\sim}{m}}$, defined by (4.9), will replace their unconditional counterparts $\pi_{k|s}$ and $\pi_{k\ell|s}$ in Results 1 to 3. Here, Remark 1 in Section 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 $t = \Sigma_U y_k$:

$$\hat{t}_{c1REG} = \Sigma_s \frac{\hat{y}_k}{\pi_{ak}} + \Sigma_r \frac{y_k - \hat{y}_k}{\pi_{ak}\pi_{k|s,\underset{\sim}{m}}} = \sum_{h=1}^{H_s} \left\{ \Sigma_{s_h} \frac{\hat{y}_k}{\pi_{ak}} + f_h^{-1} \Sigma_{r_h} \frac{y_k - \hat{y}_k}{\pi_{ak}} \right\} . \qquad (6.3)$$

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

$$\hat{V}(\hat{t}_{c1REG}) = \Sigma\Sigma_r \check{\Delta}_{ak\ell} \check{y}_k \check{y}_\ell / \pi_{k\ell|s,\underset{\sim}{m}} + \sum_{h=1}^{H_s} n_h^2 (1-f_h) s^2_{\check{e}r_h} / n_h .$$

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, $\hat{t}_{c1REG}$ will yield a shorter confidence interval.

EXAMPLE 5. We reconsider the situation outlined in Example 4. 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 sampling fraction in stratum h is $f_h = m_h/n_h$, the slope estimate becomes

$$b = \frac{\sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_h} \check{y}_k}{\sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_h} \check{x}_k} . \qquad (6.4)$$

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

$$\hat{t}_{1REG} = (\Sigma_s \check{x}_k) b \; ; \; \hat{t}_{3REG} = N \tilde{x}_s b \qquad (6.5)$$

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

$$\hat{V}(\hat{t}_{1REG}) = \Sigma\Sigma_r \check{\Delta}_{ak\ell} \check{y}_k \check{y}_\ell / \pi_{k\ell|s} + \hat{V}_2 \tag{6.6}$$

$$\hat{V}(\hat{t}_{3REG}) = \Sigma\Sigma_r \check{\Delta}_{ak\ell} \check{e}_{1k} \check{e}_{1\ell} / \pi_{k\ell|s} + \hat{V}_2 \tag{6.7}$$

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

The results (6.5) to (6.7) apply without any formal change in Case STBE (although the notation should then be $\hat{t}_{1cREG}$, $\hat{t}_{3cREG}$ to indicate the conditional nature of the regression estimator). □

7- *CONCLUSION*

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

1. Make a choice of a sampling design for phase two.

2. 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.

In many applications where two-phase sampling is likely to be used, there may be little auxiliary information available prior to phase one, that is, information about <u>all</u> units in the population U . In such cases, the first stage design will ordinarily be the simplest possible under the circumstances.

REFERENCES

CASSEL, C.M., SÄRNDAL, C.E. and WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika, 63, 615-620.

CHAUDHURI, A. and ADHIKARY, A.K. (1983). On optimality of double sampling strategies with varying probabilities. Journal of Statistical Planning and Inference, 8, 257-265.

COCHRAN, W.G. (1977). Sampling Techniques, 3rd edition. New York: Wiley.

NEYMAN, J. (1938). Contribution to the theory of sampling human populations. Journal of the American Statistical Association, 33, 101-116.

RAJ, D. (1968). Sampling Theory. New York: McGraw-Hill.

RAO, J.N.K. (1973). On double sampling for stratification and analytic surveys. Biometrka, 60, 125-133.

SÄRNDAL, C.E. (1982). Implications of survey design for generalized regression estimation of linear functions. Journal of Statistical Planning and Inference, 7, 155-170.

PART II: NON-RANDOMIZED SUBSAMPLE SELECTION (NONRESPONSE)

1- INTRODUCTION TO PART II

The bridge between the present Part II and the preceding Part I of this paper is the common feature of selection of a first set, s, from the population of $U = \{1,\ldots,k,\ldots,N\}$, followed by subselection of the set r from s, and observation of the variable of interest, y, in the set r only.

More specifically, in Part II, the first set s, now to be called the intended sample, in 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 in Parts I and II for concepts that directly correspond to each other. The response mechanism (which corresponds to the second phase sampling design in Part I) will thus be denoted $p(r|s)$. Here the statistician is forced to make an explicit assumption about the unknown form of $p(r|s)$.

A difference between Parts I and II should be signaled: In the two phase sampling situation, the first phase sample, s, is typically large, because inexpensive, and the more costly second phase sample, r, is often a small fraction, say, 20% or less of the first phase sample. The variance due to the second phase can then be considerable. In the nonresponse situation, the size of the intended sample s may still be large, but, by contrast, if the rate of nonresponse is not too pronounced, the ultimate sample r is, say, 80% or more of the intended sample s. The variance attributed to the nonresponse "phase" may thus be relatively modest in comparison to the variance due to selection of s; the main problem is instead the bias due to the systematic (rather than random) manner in which the nonresponse decimates the intended sample s.

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 h:th 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 $t = \Sigma_1^N y_k$ becomes

$$\hat{t} = \sum_{h=1}^{H} (N/n)(n_h/m_h)\Sigma_{r_h} y_k = N \sum_{h=1}^{H} w_h \bar{y}_{r_h} , \qquad (1.1)$$

where H is the fixed number of adjustment groups, $w_h = n_h/n$ is the sample portion in the h:th group and $\bar{y}_{r_h}$ is the respondent mean of y in the h:th group. The estimator (1.1) has been analyzed, in different ways, by Thomsen (1973), Oh and Scheuren (1983) and others. Different points of departure may be used in the analysis of (1.1). 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 district 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. For example, this spirit penetrates a number of the contributions to the recent and authoritative "Incomplete Data in Sample Surveys", volumes 1 to 3. Here we can distinguish two lines of reasoning: One of them extends the classical randomization theory. The set of (known) inclusion probabilities is supplemented with the set of (unknown but modelled) response probabilities, to form the necessary material for a modified randomization theory that can address the nonresponse situation. In "Incomplete Data in Sample Surveys", this line of thought is emphasized in Platek and Gray (1983), Oh and Scheuren (1983), Cassel, Särndal and Wretman (1983), whereas the second line of reasoning, fully model based inference conditionally on the set r, is present in other contributions, such as Little (1983) and Rubin (1983). Here, we shall use the former approach.

The estimator (1.1) can be justified through the assumption that the population U is composed of fixed set of disjoint subpopulations such that all units within a subpopulation have the same reponse probability, and that units respond

independently of each other. This model is called a "uniform response mechanism within subpopulations" by Oh and Scheuren (1983), who fittingly describe the setup with probability sampling augmented with a response model as "quasi-randomization". Assuming simple random sampling and a uniform response mechanism within subpopulations, they analyze the bias, variance and mean squared error of the estimator (1.1), 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 (cf. 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. Difference 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 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, where it may make sense to partition s according to an unchanging rule.) We shall therefore formulate a more general response model.

2. *THEORETICAL RESULTS*

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}, $\pi_{ak\ell}$ and $\Delta_{ak\ell}$ associated with this design are defined as in Section 2 of part I. In particular, $p_a(\cdot)$ may be a "complex" design in two or more stages. Once drawn, s is partitioned into H_s groups, s_h, $h = 1,\ldots,H_s$. 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 the m_h. We assume that the individual response probability is

the same for all units in s_h , $h = 1,\ldots,H_s$, which we call response homogeneity groups (Rhg's). Units are assumed to respond independently of each other. Thus we have the Rgh model: for $h = 1,\ldots,H_s$,

$$P(k\epsilon r|s) = \pi_{k|s} = \theta_{hs} \ , \ k \in s_h \ ; \ P(k\&\ell\epsilon r|s) = \pi_{k\ell|s} = P(k\epsilon r|s)P(\ell\epsilon r|s) \ , \ k \neq \ell \ . \quad (2.1)$$

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 Sections 4 and 6 of Part I. (However, in contrast to Case STBE, the Rhg model is "only" an assumption, not an actually imposed randomization scheme.) Assuming that the Rgh model (2.1) holds, we can thus directly borrow the results reported in Part I, Sections 4 and 6, under the heading Case STBE. Note that $\underset{\sim}{m} = (m_1,\ldots,m_{H_s})$ is now the vector of respondent counts, and $f_h = m_h/n_h$ is the response rate in the h:th group, s_h . As in Part I, one can identify a "basic situation", which leads to a "conditional π*es estimator", and Situations 1, 2 and 3 (as defined in Section 5 of Part I), with different levels of auxiliary information and leading to three different "conditional regression estimators". For easy reference, we list the estimators for the four different situations:

The conditional π*es estimator becomes

$$\hat{t}_{c\pi*} = \sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_h} \check{y}_k \ . \quad (2.2)$$

The conditional regression estimators are given by

$$\hat{t}_{c1REG} = \sum_{h=1}^{H_s} \left(\Sigma_{s_h} \frac{\hat{y}_k}{\pi_{ak}} + f_h^{-1} \Sigma_{r_h} \frac{y_k - \hat{y}_k}{\pi_{ak}} \right) \ , \quad (2.3)$$

$$\hat{t}_{c2REG} = \Sigma_U \hat{y}_k + \sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_h} \frac{y_k - \hat{y}_k}{\pi_{ak}} \ , \quad (2.4)$$

$$\hat{t}_{c3REG} = \Sigma_U \hat{y}_{1k} + \sum_{h=1}^{H_s} \left(\Sigma_{s_h} \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_{ak}} + f_h^{-1} \Sigma_{r_h} \frac{y_k - \hat{y}_k}{\pi_{ak}} \right) \ . \quad (2.5)$$

The estimator (2.2), its variance and its estimated variance (see below) are discussed, in the context of nonresponse, by Singh and 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

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

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 is of the form

$$\hat{V}_1(\hat{t}) = \Sigma\Sigma_r \check{\Delta}_{ak\ell} \check{a}_k \check{a}_\ell / \pi_{k\ell|s,\underset{\sim}{m}}$$

where the definition of the quantities $\check{a}_k$ depends on the estimator. For $\hat{t} = \hat{t}_{c\pi*}$ and for $\hat{t} = \hat{t}_{c1REG}$, we have $\check{a}_k = \check{y}_k = y_k/\pi_{ak}$; for $\hat{t} = \hat{t}_{c2REG}$, $\check{a}_k = e_k/\pi_{ak} = (y_k - \hat{y}_k)/\pi_{ak}$, and finally for $\hat{t} = \hat{t}_{c3REG}$, $\check{a}_k = e_{1k}/\pi_{ak} = (y_k - \hat{y}_{1k})/\pi_{ak}$. The fitted values $\hat{y}_k$ and $\hat{y}_{1k}$ are as defined in Section 5 of Part I. The estimated second component is given by the "stratified form"

$$\hat{V}_2(\hat{t}) = \sum_{h=1}^{H_s} n_h^2 (1-f_h) S^2_{\check{a}r_h} / m_h , \qquad (2.6)$$

where $S^2_{\check{a}r_h}$ 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, repeated realizations r under the Rhg model (2.1). We assume that each time a given sample s is selected, the repeated realizations of the model (2.1) 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 1. Alternative ways to use the sampling weights. Assume that a complex (non-self-weighting) design in two or more stages is used to draw the intended sample s. Let the sampling weights associated with this design be $1/\pi_{ak}$, and $\check{y}_k = y_k/\pi_{ak}$. Information is obtained that permits s to be divided into Rhg's s_h, $h = 1,\ldots,H_s$; no further information is gathered. In the h:th group, the responses y_k are obtained for the set r_h $(r_h \subseteq s_h)$. The statistician used to working with sample weighted observations can now easily think of at least three different estimators that attempt to correct for nonresponse through the Rhg-groups:

$$\hat{t}_1 = \sum_{h=1}^{H_s} f_h^{-1} \sum_{r_h} \check{y}_k$$

$$\hat{t}_2 = \sum_{h=1}^{H_s} \left(\sum_{s_h} \pi_{ak}^{-1}\right) \frac{\sum_{r_h} \check{y}_k}{\sum_{r_h} \pi_{ak}^{-1}}$$

$$\hat{t}_3 = \left(\sum_{s} \pi_{ak}^{-1}\right) \frac{\sum_{h=1}^{H_s} f_h^{-1} \sum_{r_h} \check{y}_k}{\sum_{h=1}^{H_s} f_h^{-1} \sum_{r_h} \pi_{ak}^{-1}}$$

with $f_h = m_h/n_h$. Which is the "correct" estimator? Note that if the design is self-weighting (that is, π_{ak} = constant for all k), there is no issue, since the three estimators will then agree. Let us analyse the origin of each formula. Here, $\hat{t}_1$ is the conditional π*es estimator (2.2); it is based purely on a weighting argument. By contrast, $\hat{t}_2$ and $\hat{t}_3$ also require some modeling; they arise from

the $\hat{t}_{clREG}$ formula (2.3), for two different model formulations. The model that generates $\hat{t}_2$ is

$$E_\xi(y_k) = \beta_h \ ; \ V_\xi(y_k) = \sigma_h^2 \quad \text{for} \quad k \in s_h \ . \tag{2.7}$$

The fit of this model yields

$$b_h = \tilde{y}_{r_h} = \Sigma_{r_h} \check{y}_k / \Sigma_{r_h} \pi_{ak}^{-1} \tag{2.8}$$

and $\hat{y}_k = \tilde{y}_{r_h}$ for all $k \in s_h$. Inserting into (2.3) we obtain $\hat{t}_2$, where $(\Sigma_{s_h} \pi_{ak}^{-1})/(\Sigma_{r_h} \pi_{ak}^{-1})$ is a "sample weighted response rate"; see Platek and Gray (1983).

Underlying $\hat{t}_3$ is the "trivial" model $E(y_k) = \beta$; $V(y_k) = \sigma^2$ for all k . Here

$$b = \tilde{y}_r = \Big(\sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_h} \check{y}_k\Big) \Big/ \Big(\sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_h} \pi_{ak}^{-1}\Big)$$

and $\hat{y}_k = \tilde{y}_r$ for all k . Inserting into (2.3), the estimator $\hat{t}_3$ follows.

Although the three estimators yield slightly different estimates of t for a given data set, they will be accompanied by the same estimated variance, so the width of the three confidence intervals will be the same. (The large sample efficiency of the three methods is the same.) It is easy to see that the three estimators share the same estimated first component. The common estimated second component is

$$\hat{V}_2(\hat{t}) = \sum_{h=1}^{H_s} n_h^2 (1-f_h) s^2_{\check{y}r_h} / m_h \ . \tag{2.9}$$

This follows since the respective quantities $\check{a}_k$ to use in (2.6) are $\check{a}_k = \check{y}_k$ (for $\hat{t}_1$), $\check{a}_k = (y_k - \tilde{y}_{r_h})/\pi_{ak}$ for $k \in r_h$ (for $\hat{t}_2$) and $\check{a}_k = (y_k - \tilde{y}_r)/\pi_{ak}$ for $k \in r$ (for $\hat{t}_3$); all three cases lead to (2.9).

Now, if the population size N were known, further alternatives to $\hat{t}_1$, $\hat{t}_2$ and $\hat{t}_3$ include

$$\hat{t}_4 = (N/\Sigma_s \pi_{ak}^{-1})\hat{t}_2 \; ; \; \hat{t}_5 = (N/\Sigma_s \pi_{ak}^{-1})\hat{t}_3 \; .$$

These can be shown to derive from $\hat{t}_{c3REG}$ given by (2.5).

If the intended sample s is drawn by simple random sampling, then all five estimators, $\hat{t}_1$ to $\hat{t}_5$, become

$$N \sum_{h=1}^{H_s} (n_h/n)\bar{y}_{r_h} \; ,$$

the wellknown weighting class estimator. □

EXAMPLE 2. The case of known population group sizes. A standard estimator in the nonresponse literature is

$$\hat{t} = \sum_{h=1}^{H} N_h \bar{y}_{r_h} \; , \tag{2.10}$$

where $N_1,\ldots,N_H$ are the known sizes of certain fixed population groups $U_1,\ldots,U_H$, and $\bar{y}_{r_h}$ the simple mean of y_k in the set r_h. Thomsen (1973) and Oh and Scheuren (1983) among others, have analyzed this estimator, which we can identify, in our approach, as the $\hat{t}_{c2REG}$ estimator for the ANOVA model (2.7). The fitted values are $\hat{y}_k = \tilde{y}_{r_h}$ for $k \in s_h$, where $\tilde{y}_{r_h}$ is given by (2.8). Insertion into (2.4) gives

$$\hat{t}_{c1REG} = \sum_{h=1}^{H} N_h \tilde{y}_{r_h} \; .$$

In particular, (2.10) is the special case corresponding to a self-weighting design ($\pi_k = f$, say, for all k). □

EXAMPLE 3. 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 i:th 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;

$h = 1,\ldots,H_s$, and $f_h = m_h/n_h$ is the response rate in the h:th group. The conditional π*es estimator is, from (2.2),

$$\hat{t}_{c\pi*} = f_I^{-1} \sum_{i\in s_I} f_i^{-1}\Big(\sum_{h=1}^{H_s} f_h^{-1} \Sigma_{r_{ih}} y_k\Big) , \qquad (2.11)$$

where r_{ih} is the set of respondents in the crossclassification of the i:th psu with the h:th 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 i:th psu, whereas f_h^{-1} is the weight associated with correction for nonresponse in the h:th 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 (2.3) becomes

$$\hat{t}_{clREG} = (f_I^{-1} \sum_{i\in s_I} f_i^{-1} \Sigma_{s_i} x_k)b \qquad (2.12)$$

where

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

(Note that if $x_k = 1$ for all k , the estimator (2.12) will still be different from (2.11).) The estimated variance follows easily from the general formulas, 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 (2.11) and $\check{a}_k = (y_k - bx_k)/\pi_{ak}$ in the case of (2.12). □

It can not 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".

If the assumed Rhg model is false, the estimators will be biased to an extent that depends on the degree of model breakdown. The Monte Carlo study in Section 3 illustrates that regression estimators (when based on strong concomitant variables) are more resistant to bias than the estimators of straight weighting (π*es) type.

3- *A SIMULATION STUDY*

We carried out a small scale Monte Carlo experiment involving repeated draws of simple random samples of size $n = 400$ from a real population U of $N = 1227$ Swedish households, for which we have access to y_k, the disposable income of the household, and w_k, the taxable income of the household, $k = 1,\ldots,N$. We studied alternative estimators of $t = \Sigma_U y_k$, based on samples affected by nonresponse. In our study, w_k serves as a concomitant variable. We can pretend that w_k is available from the tax returns and not affected by nonresponse and that y_k is obtained from a survey, for responding households only. (The fact that simple random sampling was used to select s is not seen as a limitation. The objective here is to study the effects of nonresponse, and our conclusions would have been similar under some other sampling design.) The program for the simulation study was written in APL (VSAPL) and run on an IBM 370/158. The authors are thankful to Mr. Claes Andersson for his assistance with the simulation.

Once selected, each sample s was exposed to simulated unit nonresponse. By the true response mechanism (TRM) we mean the random mechanism chosen by us to generate unit nonresponse. (Note that in this setting, a true mechanism does exist, since the experiment is fully controlled.) We studied two TRM's, each conforming to an Rhg model with four groups (which we may assume to correspond to four different household types). The assumed response model (ARM) is the Rhg model actually used in the calculation of estimator, variance estimator and confidence interval for t. Our objectives were (a) to see if the preceding theory (based partly on large sample approximations) holds fairly well for moderate sample sizes when the ARM is true (that is, equal to the TRM); (b) to study the bias and the validity of the confidence statements when the ARM is false (that is, deviating from the TRM). The population regression of y_k on $\sqrt{w_k} = x_k$ is heteroscedastic and roughly linear through the origin; a decent (in no sense perfect) description is the ratio model

$$E(y_k) = \beta x_k \ ; \ V(y_k) = \sigma^2 x_k \ . \qquad (3.1)$$

The population correlation coefficient between x_k and y_k is $r_{xy} = 0.831$. Our scenario assumes that x_k is observed for units k in the intended sample s , while y_k is observed for k in the response set r only.

The TRM was determined by dividing U into four subpopulations, U_h , $h = 1,\ldots,4$. These were created from the 1227 y_k-values through a process that combined an element of random assignment with some deliberate steering to separate the four subpopulation y-means. We allowed considerable overlap between the groups, as far as the y_k-values were concerned; the separation between the group means is consequently far from maximal. If N_h denotes the number of units, $\bar{y}_h$ the mean of y and $\bar{x}_h$ the mean of x in U_h , then the triple $(N_h, \bar{y}_h \times 10^2, \bar{x}_h \times 10)$ had the following values:

For U_1 : (373, 4.18, 1.68) ; for U_2 : (303, 5.65, 2.25) ;
for U_3 : (280, 6.73, 2.54) ; for U_4 : (271, 8.31, 2.79) .

For $h = 1,\ldots,4$, the TRM was given its final specification by attaching to each unit in U_h the same fixed value, θ_h , used in the simulation as the true individual response probability of the unit. The θ_h-values 0.45, 0.60, 0.75, 0.90 were used as follows to create two different TRM's:

TRM 1. For U_1 : $\theta_1 = 0.45$; for U_2 : $\theta_2 = 0.60$; for U_3 : $\theta_3 = 0.75$; for U_4 : $\theta_4 = 0.90$. Consequently, there is a (moderate) tendency for the response probability to increase with the y_k-value (and with the x_k-value). We have $r_{\theta y} = 0.44$; $r_{\theta x} = 0.39$, where $r_{\theta y}(r_{\theta x})$ is the correlation coefficient (calculated over the $N = 1227$ units) between the individual response probability and the y_k-value (x_k-value).

TRM 2. The same set of θ_h-values were attached to the U_h in the reverse

order: For U_1 : $\theta_1 = 0.90$; for U_2 : $\theta_2 = 0.75$; for U_3 : $\theta_3 = 0.60$; for U_4 : $\theta_4 = 0.45$. Here there is a moderate tendency for the response probability to decrease with the y_k-value (and with the x_k-value); we have $r_{\theta y} = -0.44$; $r_{\theta x} = -0.39$.

As a result of the considerable overlap allowed between the subpopulations, the individual response probability has a rather modest correlation with the y_k-value, and with the x_k-value. (It was because we wanted to keep these correlations low that the groups were created with overlap.) Despite these modest correlations, large biases are created in the estimates of t , for both TRM's, unless effective corrective action is taken.

For each of the two TRM's, we studied three different ARM's (for all s , we used a fixed number of groups, $H = 4$) : Situation TRUE: The ARM is true, that is, stated in terms of $H = 4$ groups identical to the four Rhg's of the TRM; $s_h = s \cap U_h$; $h = 1,\ldots,4$. Situation FALSE 2: The ARM is falsely stated in terms of $H = 2$ groups, each formed by merging two neighbouring Rhg's of the TRM; s_1 with s_2 ; s_3 with s_4 . Situation FALSE 1. The ARM is falsely stated in terms of $H = 1$ group, formed by merging all four Rhg's of the TRM; that is, the response probability is incorrectly assumed to be uniform throughout the population.

Three estimators were studied:

$$\hat{t}_A = \frac{N}{n} \sum_{h=1}^{H} n_h \bar{y}_{r_h} \; ; \; \hat{t}_B = \frac{N}{n} (\Sigma_s x_k) b \; ; \; \hat{t}_C = \frac{N}{n} \sum_{g=1}^{G} n_{g\cdot} b_g \; ,$$

where

$$b = \Big(\sum_{h=1}^{H} f_h^{-1} \Sigma_{r_h} y_k \Big) \Big/ \Big(\sum_{h=1}^{H} f_h^{-1} \Sigma_{r_h} x_k \Big) \; ; \; b_g = \Big(\sum_{h=1}^{H} f_h^{-1} \Sigma_{r_{gh}} y_k \Big) \Big/ \Big(\sum_{h=1}^{H} f_h^{-1} \Sigma_{r_{gh}} x_k \Big) \; .$$

Here, $\hat{t}_A$ is the usual weighting class estimator, while $\hat{t}_B$ and $\hat{t}_C$ result from the regression estimator $\hat{t}_{1cREG}$ given by (2.3). $\hat{t}_B$ is generated by the ratio model (3.1). In this model, the x-variable appears in its original,

continuous form, but an alternative (with no appreciable information loss) is to group the x_k-values. This gives rise to "auxiliary groups", which are conceptually different from the Rhg's. More explicitly, for each realized sample s, suppose that G equal-sized auxiliary groups are formed by ordering the n values x_k from the smallest to the largest, letting the first group, $s_{1\cdot}$, consist of $100/G\%$ of the sample with the smallest x_k-values, the second group, $s_{2\cdot}$, of the next $100/G\%$ of the x-ordered sample, etc.

A reasonably good alternative model for explaining the y-variable is then

$$E(y_k) = \beta_g \; ; \; V(y_k) = \sigma_g^2 \quad , \text{ all } \; k \in s_{g\cdot}$$

where $s_{g\cdot}$, $g = 1,\ldots,G$. The crossclassification of the G auxiliary groups with the H Rhg's of the ARM gives rise to GH cells. Let r_{gh} be the response set in the cell gh, m_{gh} the size of r_{gh}, and

$$m_h = \sum_{g=1}^{G} m_{gh} \; .$$

Also let n_{gh} be the size of the subset of the sample s that falls in cell gh, and

$$n_h = \sum_{g=1}^{G} n_{gh} \; ; \; n_{g\cdot} = \sum_{h=1}^{H} n_{gh} \; .$$

In our experiment, $G = H = 4$. Thus, $\hat{t}_A$ uses only the Rhg's of the ASM, whereas $\hat{t}_B$ and $\hat{t}_C$ incorporate both the x-information and the Rhg's. As the empirical results will show, the presence of the x-variable in $\hat{t}_B$ and $\hat{t}_C$ serves as a variance reducing device (whether the ARM is true or false), and, perhaps more importantly, as a bias reducing device when the ARM is false.

If $\hat{t}$ is one of the studied estimators, the variance estimate is composed as

$$\hat{V}(\hat{t}) = \hat{V}_1(\hat{t}) + \hat{V}_2(\hat{t}) \; .$$

Observing that, in our experiment, $\pi_{ak} = n/N$ · for all k and $\pi_{ak\ell} = n(n-1)/N(N-1)$ for all $k \neq \ell$, the general results in Section 2 lead us to conclude that $\hat{t}_A$, $\hat{t}_B$ and $\hat{t}_C$ share the same first variance component, estimated by

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

where $f = n/N$; $Q_h = (n-n_h)/(n-1)m_h$; $w_h = n_h/n$; $\bar{y}_{r_h} = \Sigma_{r_h} y_k/m_h$; $\bar{\bar{y}}_r = \sum_{h=1}^{H} w_h \bar{y}_{r_h}$.
The second variance component (the "nonresponse variance") is estimated by

$$\hat{V}_2(\hat{t}) = \frac{N^2}{n^2} \sum_{h=1}^{H} n_h^2 \frac{1-f_h}{m_h} s^2_{ar_h},$$

where $s^2_{ar_h}$ is the variance in the set r_h of the numbers a_k such that, for $\hat{t} = \hat{t}_A$, $a_k = y_k$; for $\hat{t} = \hat{t}_B$, $a_k = y_k - bx_k$; for $\hat{t} = \hat{t}_C$, $a_k = y_k - b_g x_k$ when k is in group g ; $g = 1,\ldots,G$.

For each of our two TRM's, the simulation proceeded as follows: A first SI sample s of size $n = 400$ is drawn from U . For each unit k in s, the value x_k and the Rhg membership (according to the TRM) are recorded. If k is found to belong to group h, a Bernoulli trial is carried out with the known probability θ_h of "success" (= response) and $1-\theta_h$ of "failure" (= nonresponse). The n independent trials generate a response set r . Then y_k is recorded for each $k \in r$, and $\hat{t}_A$, $\hat{t}_B$ and $\hat{t}_C$, as well as their respective variance estimators and confidence intervals, are computed for each of the situations TRUE, FALSE 2 and FALSE 1. The procedure is repeated for a total of $K = 1000$ generated response sets r . If $\hat{t}_\nu$ denotes the value for the ν:th response set of one of the three estimators, the following summary performance measures were computed:

$$\text{BIAS} = \sum_{\nu=1}^{K} (\hat{t}_\nu - t)/K \; ; \; \text{MSE} = \sum_{\nu=1}^{K} (\hat{t}_\nu - t)^2/K \; ; \; \text{VAR} = \text{MSE} - (\text{BIAS})^2 \; ;$$

$$\text{MEAN } \hat{V}_1 \; ; \; \text{MEAN } \hat{V}_2 \; ; \; \text{MEAN } \hat{V} \; ; \; \text{CVR90} \; ; \; \text{CVR95}$$

where MEAN $\hat{V}_1$, MEAN $\hat{V}_2$ and MEAN $\hat{V}$ are the means of $\hat{V}_1(\hat{t}_\nu), \hat{V}_2(t_\nu)$ and

$\hat{V}(\hat{t}_\nu) = \hat{V}_1(\hat{t}_\nu) + \hat{V}_2(\hat{t}_\nu)$, respectively, over the $K = 1000$ repetitions. Finally, CVR90 is 100 times the proportion of the 1000 confidence intervals, with $z_{0.950} = 1.645$, that contain the true total t. For CVR95, 1.960 replaces 1.645.

Table 1. Performance measures for $\hat{t}_A$, $\hat{t}_B$ and $\hat{t}_C$ under three ARM's: TRUE, FALSE 2 and FALSE 1. Upper portion of the table: TRM 1 ; lower portion of the table: TRM 2. True value to estimate: $t = 74.05$.

ASM		BIAS	MSE	VAR	MEAN $\hat{V}$	MEAN $\hat{V}_2$	CVR90	CVR95
	$\hat{t}_A$	0.00	4.75	4.75	5.07	1.99	90.1	95.2
TRUE	$\hat{t}_B$	-0.01	3.78	3.78	3.86	0.78	91.6	95.5
	$\hat{t}_C$	0.00	3.75	3.75	3.82	0.74	91.1	95.6
	$\hat{t}_A$	1.11	6.06	4.83	5.17	2.06	88.0	93.3
FALSE2	$\hat{t}_B$	0.19	3.82	3.79	3.89	0.78	91.3	95.6
	$\hat{t}_C$	0.38	3.92	3.77	3.88	0.77	91.0	95.6
	$\hat{t}_A$	4.85	28.75	5.28	5.81	2.55	33.1	46.3
FALSE1	$\hat{t}_B$	1.26	5.30	3.71	4.04	0.78	84.3	92.6
	$\hat{t}_C$	1.19	5.14	3.72	4.14	0.88	86.7	93.9
	$\hat{t}_A$	0.04	5.96	5.96	5.67	2.58	90.1	94.4
TRUE	$\hat{t}_B$	0.02	4.09	4.09	3.93	0.84	89.8	94.8
	$\hat{t}_C$	0.02	4.30	4.30	4.07	0.98	90.3	95.4
	$\hat{t}_A$	-1.00	6.53	5.53	5.25	2.25	86.1	91.0
FALSE2	$\hat{t}_B$	-0.31	4.01	3.91	3.74	0.74	88.8	94.0
	$\hat{t}_C$	-0.50	4.35	4.11	3.90	0.90	89.0	93.8
	$\hat{t}_A$	-4.55	25.54	4.86	4.69	1.85	34.3	45.1
FALSE1	$\hat{t}_B$	-1.28	5.37	3.74	3.43	0.59	80.3	86.6
	$\hat{t}_C$	-1.18	5.34	3.96	3.45	0.61	80.1	86.5

Table 1 shows the following:

Situation TRUE: (1) Each estimator is approximately unbiased (BIAS $\doteq$ 0) ; (2) Each estimator has an approximately unbiased variance estimator (VAR is close

to MEAN $\hat{V}$) ; (3) $\hat{t}_B$ and $\hat{t}_C$ (which use the x-variable) have considerably smaller MEAN $\hat{V}$ than $\hat{t}_A$ (which ignores x) ; (4) the reason for (3) is that MEAN $\hat{V}_2$ is much smaller for $\hat{t}_B$ and $\hat{t}_C$ than for $\hat{t}_A$, whereas MEAN $\hat{V}_1$ is the same for all three cases; (5) the coverage rate for each estimator is near the nominal rate (CVR 90 ≐ 90%, CVR 95 ≐ 95%). The conclusions confirm what theory leads us to except.

Situations FALSE 1 and FALSE 2: (1) All three estimators are biased, although $\hat{t}_B$ and $\hat{t}_C$ are much less so than $\hat{t}_A$; (2) all three variance estimators are fairly insensitive to breakdown of the ARM , that is, VAR and MEAN $\hat{V}$ are still fairly close; (3) VAR is again considerably smaller for $\hat{t}_B$ and $\hat{t}_C$ than for $\hat{t}_A$, as a result of a greatly reduced second variance component; (4) the coverage rates CVR 90 and CVR 95 are much closer to nominal levels for $\hat{t}_B$ and $\hat{t}_C$ than for $\hat{t}_A$. (Extremely poor CVR's are recorded for $\hat{t}_A$ in the case FALSE 1.) The primary explanation is the lower bias of $\hat{t}_B$ and $\hat{t}_C$. Here, point (1) confirms earlier work (for example, Sarndal and Hui, 1981) indicating that regression estimators are more bias resistant. Additional work is needed to see if point (2) holds more generally.

4- *DISCUSSION*

Our ambition with the foregoing simulation and theory was partly to create increased awareness about the forces at play in the nonresponse situation. To illustrate, let us examine a statement by Oh and 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. ... A further difficulty with this prescription is that it is only for the respondents that within-group variability can be observed." A statement such as this reflects,

we believe, a not untypical hesitation and uncertainty that survey practitioners feel about the proper role of adjustment groups. Some of the confusion may have its roots in the old and crude dichotomy response stratum/nonresponse stratum. In our opinion, it is necessary to distinguish the role of the Rhg's from that of other information (the x-variables) recorded for $k \in s$. Two very different concepts are involved. The sole criterion for the Rhg's should be that they eliminate bias, to the fullest extent 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, $\underset{\sim}{x}_k$, that will yield variance reduction and added protection against bias. Groups that eliminate or reduce bias are not necessarily variance reducing, and, contrary to the cited statement, 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 often 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.

5- *SOFTWARE AT STATISTICS SWEDEN FOR POINT ESTIMATES AND STANDARD ERRORS*

Statistics Sweden can often rely, for its surveys, on good sampling frames with up-to-date addresses for most units in the population under study. Many surveys involve mail inquiry, often with follow-ups by telephone, attempting to reach all (or a subsample of) nonrespondents. (Of course, not all attempts will result in a completed telephone interview, and therefore some nonresponse remains). In these situations, stratified sampling is efficient, especially since one can often

let some of the domains of study correspond to strata. Much of the auxiliary information in the frame is used for determining the sampling design, and a simple estimator (of the π^*es type) will often be efficient. TAB 68, developed at Statistics Sweden, is the principal software used by the agency in the production of statistical tables. An advantage of TAB 68 is great liberty to specify the features of desired tables. One major draw-back is that to date, TAB 68 has been limited to the calculation of point estimates. Statistics Sweden is in the process of implementing new software, SMED 83, which, while maintaining the flexibility of TAB 68 for producing tables, will make possible not only the calculation of point estimates of totals, means or ratios, but also their estimated standard errors (or coefficients of variation), all for presentation in the same table. Underlying SMED 83 is the theory that we have presented in this paper.

REFERENCES

BINDER, D.A. (1983). Some models for non-response and other censoring in sample surveys. Report, Statistics Canada, Dec. 1983.

CASSEL, C.M., SÄRNDAL, C.E. and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In Madow, W.G., Olkin, I. (eds) Incomplete Data in Sample Surveys, Vol. 3. New York: Academic Press, 143-160.

COCHRAN, W.G. (1977). Sampling Techniques, 3rd edition. New York: Wiley.

DALENIUS, T. (1983). Some reflections on the problem of missing data. In Madow, W.G., Olkin, I. (eds) Incomplete Data in Sample Surveys, vol. 3. New York: Academic Press, 411-413.

KALTON, G. (1983). Models in the practice of survey sampling. International Statistical Review, 51, 175-188.

LITTLE, R.J.A. (1983). Superpopulation models for nonresponse. In Madow, W.G., Olkin, I. and Rubin, D.B. (eds) Incomplete Data in Sample Surveys, vol. 2. New York: Academic Press, 337-413.

OH, H.L. and SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. In Madow, W.G., Olkin, I. and Rubin, D.B. (eds) Incomplete Data in Sample Surveys, vol. 2. New York: Academic Press, 143-184.

PLATEK, R. and GRAY, G.B. (1983). Imputation Methodology. In Madow, W.G., Olkin, I. and Rubin, D.B. (eds) Incomplete Data in Sample Surveys, vol. 2. New York: Academic Press, 255-293.

RUBIN, D.B. (1983). Conceptual issues in the presence of nonresponse. In Madow, W.G., Olkin, I. and Rubin, D.B. (eds) Incomplete Data in Sample Surveys, vol. 2. New York: Academic Press, 125-142.

SÄRNDAL, C.E. and HUI, T.K. (1981). Estimation for nonresponse situations: To what extent must we rely on models? In Krewski, D., Platek, R. and Rao, J.N.K. (eds) Current Topics in Survey Sampling. New York: Academic Press, 227-246.

SINGH, S. and SINGH, R. (1979). On random non-response in unequal probability sampling. Sankhya 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. Statistisk Tidskrift, 11, 278-285.

Tidigare nummer av Promemorior från P/STM:

NR

1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)

2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)

3 En presentation av Box-Jenkins metod för analys och prognos av tidsserier. Åke Holmén (1979-12-20)

4 Handledning i AID-analys. Anders Norberg (1980-10-22)

5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)

6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)

7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)

8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)

9 Regression analysis and ratio analysis for domains. A randomization theory approach. Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)

10 Current survey research at Statistics Sweden. Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)

11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata. Lars-Otto Sjöberg (1984-01-11)

12 Regressionsanalys för f d statistikstuderande. Harry Lütjohann (1984-02-01)

13 Estimating Gini and Entropy inequality parameters. Fredrik Nygård och Arne Sandström (1985-01-09)

14 Income inequality measures based on sample surveys. Fredrik Nygård och Arne Sandström (1985-05-20)

15 Granskning och evalvering av surveymodeller, tiden före 1960. Gösta Forsman (1985-05-30)

16 Variance estimators of the Gini coefficient - simple random sampling. Arne Sandström, Jan Wretman och Bertil Waldén (Memo, Februari 1985)

17 Variance estimators of the Gini coefficient - probability sampling. Arne Sandström, Jan Wretman och Bertil Waldén (1985-07-05)

18 Reconciling tables and margins using least-squares. Harry Lütjohann (1985-08-01)

19 Ersättningens och uppgiftslämnarbördans betydelse för kvaliteten i undersökningarna om hushållens utgifter. Håkan L. Lindström (1985-11-29)

Kvarvarande exemplar av ovanstående promemorior kan rekvireras från Elseliv Lindfors, P/STM, SCB, 115 81 Stockholm, eller per telefon 08 7834178

☆Garnisonstryckeriet Stockholm 1985☆