SOME OPTIMALITY PROBLEMS WHEN ESTIMATING HOUSEHOLD DATA ON THE BASIS OF A "PRIMARY", STRATIFIED SAMPLE OF INDIVIDUALS

Bengt Rosén, Statistics Sweden
U/Ledn, Statistics Sweden, S-115 81, STOCKHOLM

Abstract. Statistics Sweden conducts a yearly sample survey (called HINK) with the objective to describe income conditions for different domains of households, household being determined by factual cohabitation and not by marital status.

The national population register is used as a frame for drawing a primary, stratified sample of adults. By interviews, the households of the sampled individuals are identified. Then various income (and expenditure) data are collected for the "entire" households, and used to achieve desired estimates for domains of households.

Within the framework of this sampling design, the statistician has various options; how to form strata in the population of adults, how to allocate the sample among strata, which estimator weights to use, to mention the most important ones. The multitude of objectives for the survey will also be an essential feature of the problem.

1 Background

Since 1975 Statistics Sweden has conducted a yearly survey, called HINK, with the main purpose of providing data on income for different classes of (cohabitation) households, the socioeconomic classification being of chief interest. The survey can also be, and is, used to yield data for classes of individuals. However, in this paper we shall confine ourself to the most important aspect, i.e. the household aspect.

In a recent revision of the HINK survey, we examined the efficiency of its design-estimation strategy. This paper reports on some general findings from that study. The design-estimation procedure in HINK is somewhat complicated, a main reason for this is that no sampling frame (in the form of a register) exists for cohabitation households (i.e. households determined by factual cohabitation and not by marital status). Therefore the sampling procedure which is used has sampling of adults as its "kernel". Further discussion of sampling frame, the sampling design etc., is given in Section 3.

Hence, efficiency problems concerning HINK fall under the following general heading; "Optimization of household surveys, where households are sampled via a stratified sample of adults", and this general topic will be our main theme. The presentation will be linked to the HINK survey, though, for the following reasons. The general problems and results will hopefully become more comprehensible if they are given a concrete background and moreover, a fairly concrete application will enable illustration of the orders of magnitude of the effects under consideration. We shall confine ourselves, though, to an "idealized" version of the HINK survey and work under the following simplifying assumptions; (i) No population changes occur during the survey period. (ii) The population is sampled without under- as well as overcoverage. (iii) All sampled units respond. It can be shown, though, that the analysis of a factual survey (as e.g. HINK) can be conceptualized as the simplified case.

2 Some terminology and notation

In the HINK survey a household is defined as follows. Its "core" is its "adult part" (adult = individual ≥ 18 years), which is either a cohabitation couple of opposite sexes (be they married or not) or a single adult. The complete household also includes the children (< 18 years) under "everyday care" of the adult(s). Let

U^H denote the population of households, and let d denote a generic element in U, (2.1)

$\underline{x} = \{x_d; d \in U\}$ denote a household variable, (2.2)

G denote a domain of study (i.e. a subset of U^H). (2.3)

The $\underline{x}$-total over G, the size of G and the $\underline{x}$-mean over G are denoted as follows, where $1_G(\cdot)$ stands for the indicator of the set G and $\underline{1}$ for the household variable $\underline{1}=\{\equiv 1; d \in U^H\}$,

$$\tau(\underline{x};G) = \sum_{d \in U^H} x_d \cdot 1_G(d), \quad (2.4)$$

$$g(G) = \sum_{d \in U^H} 1_G(d) = \tau(\underline{1};G), \quad (2.5)$$

$$\mu(\underline{x};G) = \tau(\underline{x};G)/g(G) = \tau(\underline{x};G)/\tau(\underline{1};G). \quad (2.6)$$

3 Chief aims and main features of the sampling procedure

3.1 Chief aims

A rough formulation of the problem we shall consider is as follows.

> Let $G_1, G_2, \ldots, G_R$ be a specified set of disjoint household domains. Achieve, under prevailing constraints, the best possible estimates of $\{\mu(\underline{x};G_r), g(G_r), \tau(\underline{x};G_r);\ r=1,2,\ldots,R\}$, for a specified collection of $\underline{x}$-variables. (3.1)

Here the order μ, g, τ should be regarded as an ordering according to importance, domain means being of greatest interest while domain totals are of comparatively less interest.

In the HINK survey the $\underline{x}$-variable of greatest interest is disposable income, which roughly is defined as income from work and capital plus social benefits minus tax. The study domains of chief interest are the socioeconomic classes of households listed in Table 1 below. (There are of course rules for classifying a household when partners belong to different socioeconomic classes.)

Domains, and their notation	Appr. group size, g(G)
Unskilled worker househ., G_1	805 000
Skilled worker househ., G_2	525 000
Junior sal. empl. househ., G_3	395 000
Interm. sal. empl. househ., G_4	425 000
Senior sal. empl. househ., G_5	275 000
Enterpreneur househ., G_6	145 000
Farmer househ., G_7	75 000
Pensioner househ., G_8	1 065 000

Table 1. The major socioeconomic classes in the HINK survey.

3.2 Main features of the sampling procedure

Since the chief aim of the HINK survey is to describe household conditions, the sampling procedure would ideally involve a properly designed sample from a frame containing all households. However, although we have many registers in Sweden, there is no register of (cohabitation) households. Lacking such an ideal frame, the HINK survey uses the register of the total population (RTB), which includes adults as well as children. RTB does contain information on marriages, but the frequency of nonmarital cohabitation is quite high in Sweden and, therefore, individuals are chosen as the "primary" sampling units.

A sample of households is generated in the following way. In the first round a "primary", stratified sample of adults is drawn. Then, by interviewing the primary individuals, the composition of their households is determined and thereby the sample of households is obtained. Once the individuals in the sampled households are identified, data for each household member are collected, mainly from various public agencies (tax authorities, different social welfare agencies etc.). Let

V^I denote the population of adults i.e. the adults in the RTB-register. (RTB contains information on age.) (3.2)

Next we discuss methods for drawing an efficient sample from V^I. Suppose one drew a simple random sample. Then, to the first order of approximation, the estimates of the domain means, $\mu(\underline{x};G_1),\mu(\underline{x};G_2),\ldots,\mu(\underline{x};G_R)$ will have variances which are roughly inversely proportional to the sizes of the domains, i.e. to $g(G_1),g(G_2),\ldots,g(G_R)$. If there is great variation among domain sizes, this type of picture would be nonconcordant with essentially any design principle for comparison of domain means. Even if design principles often disagree, there seem to be rough concensus on the rule of thumb that, when the aim is to compare means, one should strive for fairly equal precisions in the estimates of the means of interest, and this rule of thumb will be a guide for future considerations.

As is seen in Table 1, in HINK the domain sizes differ considerably. The largest domain (pensioners) contains roughly 15 times as many households as the smallest one (farmers). One way to adjust for this unbalance, at least as a "first step", is to introduce strata A_1, A_2, A_3 and A_4 in the sampling population V^I, which have the following properties.

A_1 is "directed" towards the smallest domain G_7 of farmer households, in the sense that there is (at least one hopes) a great chance that an individual from Stratum A_1 leads to a farmer household. Similarly, assume that A_2 is directed towards the (next smallest) domain G_6 and A_3 towards the largest domain G_8. Finally let A_4 be the remaining part of the population V^I.

This type of stratification should then be followed by a sample allocation structure of the following type. Sample "high" (i.e. with a sample fraction above average) in the strata A_1 and A_2 which are directed towards small domains and sample "low" in the stratum A_3 which is directed towards the large domain.

If the directing of the strata is good (to be discussed in more detail later on) and if the sample allocation is as just described, the following will occur. Extra observations (compared with simple random sampling) are "pumped" into the domains G_7 and G_6, thereby improving estimation precision in these domains as compared with "inversely proportional to domain size", while A_3 stears away observations from the large domain G_8, thereby avoiding resource waste by an "overly" good estimation precision for this domain.

Hence, we have presented a main motivation (but others exist) for stratification of the sampling population V^I of individuals. We pursue the matter in a more general setting.

Let $A_1, A_2, \ldots, A_k$ denote a stratification (i.e. a partitioning) of the sampling population V^I, and let the corresponding stratum sizes be denoted by $N_1, N_2, \ldots, N_k$. We assume that the primary sample of adults consists of independent, simple random samples from the different strata, with sample sizes $n_1, n_2, \ldots, n_k$. The corresponding sampling fractions are denoted by

$$f_h = n_h/N_h, \qquad h = 1,2,\ldots,k. \tag{3.3}$$

4 Estimators and their variances

To estimate the quantities τ,g and μ in (2.4)-(2.6) we follow the "ordinary route" by letting estimates $\hat{\tau}(\underline{x};G)$ of domain totals be the fundamental building blocks. Domain sizes and domain means are then estimated as the special case $\hat{g}(G) = \hat{\tau}(\underline{1};G)$ and by the ratio estimator $\hat{\mu}(\underline{x};G) = \hat{\tau}(\underline{x};G)/\hat{\tau}(\underline{1};G)$.

As estimators of domain totals we consider the following type of statistics (explanation of new notation is given afterwards)

$$\hat{\tau}(\underline{x};G;\underline{\alpha}) = \sum_{h=1}^{k} \frac{N_h}{n_h} \sum_{i\in A_h} x_{d(i)}\cdot\alpha_i\cdot 1_G(d(i))\cdot I_i, \tag{4.1}$$

where

d(i) = the household to which individual i belongs, (4.2)

m(i) = the partner of individual i, when i is cohabiting, (4.3)

$\underline{\alpha} = \{\alpha_i; i\in V^I\}$ is a set of numbers, called estimation weights. (4.4)

I_i is the sample inclusion indicator for individual i. (4.5)

The following result is fairly straightforward.

LEMMA 4.1:. The statistic $\hat{\tau}(\underline{x};G;\underline{\alpha})$ in (4.1) yields unbiased estimation of $\tau(\underline{x};G)$ if, and only if, the estimation weights satisfy the following condition (4.6), which we call household balancedness,

$$\alpha_i + \alpha_{m(i)} = 1, \text{ if individual i cohabits, and } \alpha_i = 1, \text{ if individual i is single,} \quad (4.6)$$

Remark 4.1: As a special case of the lemma we have that $\hat{g}(G;\underline{\alpha})=\hat{\tau}(\underline{1};G;\underline{\alpha})$ yields unbiased estimation of g(G) as soon as $\underline{\alpha}$ is household balanced.

Furthermore, if we neglect the bias of the ratio estimator (as usually can be done), $\hat{\mu}(\underline{x};G;\underline{\alpha};\underline{\beta})=\hat{\tau}(\underline{x};G;\underline{\alpha})/\hat{\tau}(\underline{1};G;\underline{\beta})$ yields unbiased estimation of $\mu(\underline{x};G)$ as soon as $\underline{\alpha}$ and $\underline{\beta}$ both are household balanced. In the sequel, estimation weights are presumed to be household balanced. •

Remark 4.2: The present estimation situation can be regarded as a special case whithin the general framework known as "network sampling", in particular "stratified network sampling", and the following papers treat problems which are related to ours; Birnbaum & Sirken(1965), Sirken(1972) and Levy(1977). Their considerations do not cover our situation, though, for the following main reason. We allow a wider class of estimator weights in (4.1) than is done in the mentioned papers, where the interest is confined to so called multiplicity estimators. A crucial step in our analysis will be to derive optimal weights within our wider class, and the weights which turn out to be optimal, see (6.6), yield in fact an estimator outside the class of multiplicity estimators. Moreover, one of the aims in this paper is to show that optimal weights can lead to considerable efficiency gains compared with the multiplicity estimator.

In our context the multiplicity estimator corresponds to the following α-weights, which are readily seen to be household balanced,

$$\alpha_i = \alpha_{m(i)} = 1/2, \text{ for i cohabiting.} \quad (4.7)$$

We shall refer to this weighting system as half-weighting. •

Remark 4.3: A household should contribute twice in (4.1) if both adults in a cohabitation household happen to be sampled. However, in the HINK survey such double counting is omitted for practical reasons (and the omittance is adjusted for). In the sequel we neglect this complication, which in fact is practically negligible when sampling fractions are as small as in HINK (of the order 0.1 per cent).

Another matter which relates to the question of "simple or double counting of households" is the following. Let $\hat{\tau}^*$ denote the Horvitz-Thompson estimator of $\tau(\underline{x};G)$ based on the household sample which is generated by the sample of adults i.e., I_d and π_d denoting the inclusion indicator respectively the inclusion probability for household d,

$$\hat{\tau}^*(\underline{x};G) = \sum_{d\in U^H} x_d \cdot 1_G(d) \cdot \frac{I_d}{\pi_d} . \quad (4.8)$$

The following claim is fairly straightforward to check, and we omit details. Under the assumption that sampling fractions are such that the frequency of "two adults from the same household" is low, $\hat{\tau}^*(\underline{x};G)$ is, with very good approximation, an estimator within the class (4.1), namely the one given by the weighting system which is introduced in Section 6, notably in (6.6). •

Next we turn to the variances of the estimators. The general structure of the estimator $\hat{\tau}$ in (4.1) is quite simple. It is a domain total estimator based on a stratified sample. By employing this fact, variance formulas for $\hat{\tau}$, $\hat{g}$ and $\hat{\mu}$ can be reached in a fairly straightforward way, the details of which we omit. We shall adapt our formulas to a further assumption on the estimation weights which we introduce next, and which we assume to be in force in the rest of the paper. Set

$$h(i) = \text{the stratum to which individual i belongs, } i\in V^I. \quad (4.9)$$

The estimation weights $\underline{\alpha}$ are said to be stratum combination constant if the following relation holds true,

$$\alpha_i=\alpha_j \text{ as soon as } (h(i),h(m(i)))=(h(j),h(m(j))),\ i,j\in V^I \text{ and have partners.} \quad (4.10)$$

When (4.10) is in force we change the α-parameters to a-parameters as follows,

$$a_{h\ell} = \text{is the common value for the α-weights of individuals in stratum } A_h \text{ which have partner in stratum } A_\ell. \quad (4.11)$$

The previous household balancedness condition, (4.6), then takes the form,

$$a_{h\ell} + a_{\ell h} = 1, \quad h,\ell = 1,2,\ldots,k. \quad (4.12)$$

For a (fixed) domain G in U^H, set

$$B_h = \text{the set of single-adult households in G for which the adult belongs to stratum } A_h \text{ in } V^I, \quad h=1,2,\ldots,k, \quad (4.13)$$

$$B_{h\ell} = \text{the set of two-adult households in G for which one of the adults belongs to stratum } A_h \text{ and the other one to stratum } A_\ell, \quad h,\ell=1,2,\ldots,k. \quad (4.14)$$

Note the relation

$$B_{h\ell} = B_{\ell h}, \quad h,\ell=1,2,\ldots,k. \quad (4.15)$$

Set, with # denoting the number of elements in a set,

$$g_h = \#B_h, \quad g_{h\ell}=\#B_{h\ell}, \quad h,\ell=1,2,\ldots,k. \quad (4.16)$$

Furthermore, for a household variable $\underline{x}$ let

$$\mu_\varrho = \text{the } \underline{x}\text{-mean over } B_\varrho \text{ , } \varrho = h \text{ or } (h,\ell),\ h,\ell=1,2,\ldots,k, \tag{4.17}$$

$$\sigma^2_\varrho = \text{the } \underline{x}\text{-variance over } B_\varrho \text{ , } \varrho = h \text{ or } (h,\ell),\ h,\ell=1,2,\ldots,k, \tag{4.18}$$

$$\chi^2_\varrho = \sigma^2_\varrho + (\mu_\varrho - \mu(\underline{x};G))^2,\ \varrho = h \text{ or } (h,\ell),\ h,\ell=1,2,\ldots,k. \tag{4.19}$$

Remark 4.4: Note that the quantities in (4.13), (4.14) and (4.16) depend on the domain G, while the quantities in (4.17)-(4.19) depend on the domain G as well as on the variable $\underline{x}$, although we have surpressed this dependence in the notation. ●

Remark 4.5: The following relations are straightforward consequences of (4.15);

$$g_{h\ell} = g_{\ell h}\ ,\quad \mu_{h\ell} = \mu_{\ell h}\ ,\quad \sigma^2_{h\ell} = \sigma^2_{\ell h}\ ,$$

$$\chi^2_{h\ell} = \chi^2_{h\ell}\ ,\ h,\ell=1,2,\ldots,k \tag{4.20}$$

●

We are now prepared to write down the desired variance formulas. Let us state that, for the sake of simplicity, we have made some approximations of the following types; finite population corrections are neglected, N-1 and N are regarded as equal, etc.. In view of (4.11) we change the α-parameter in the previous notation to an a-parameter. Below and henceforth V denotes variance.

$$V(\hat{\tau}(\underline{x};G;\underline{a})) = \sum_{h=1}^{k} \frac{N_h}{n_h} \cdot \{g_h(\sigma^2_h+\mu^2_h) + \frac{1}{2}\cdot g_{hh} \cdot (\sigma^2_{hh}+\mu^2_{hh}) + \sum_{\ell\neq h} a^2_{h\ell}\cdot g_{h\ell}\cdot(\sigma^2_{h\ell}+\mu^2_{h\ell})\}$$

$$- \sum_{h=1}^{k} \frac{1}{n_h}\cdot(g_h\cdot\mu_h + g_{hh}\cdot\mu_{hh} + \sum_{\ell\neq h} a_{h\ell}\cdot g_{h\ell}\cdot\mu_{h\ell})^2. \tag{4.21}$$

As special case of (4.21), obtained by setting $\underline{x}=1$ (which yields $\mu_\varrho=1$ and $\sigma^2_\varrho=0$) we get,

$$V(\hat{g}(G;\underline{a})) = \sum_{h=1}^{k} \frac{N_h}{n_h}\cdot(g_h + \frac{g_{hh}}{2} + \sum_{\ell\neq h} a^2_{h\ell}\cdot g_{h\ell}) - \sum_{h=1}^{k} \frac{1}{n_h}\cdot(g_h + g_{hh} + \sum_{\ell\neq h} a_{h\ell}\cdot g_{h\ell})^2. \tag{4.22}$$

Next, by applying the usual approximation formula for the variance of a ratio estimator and adopting the following approximation assumption,

> the "squared mean" part of V(μ) is negligible compared with the "mean of squares" part, (4.23)

we arrive at the following formula,

$$V(\hat{\mu}(\underline{x};G;\underline{a};\underline{a})) = \frac{1}{g(G)^2}\sum_{h=1}^{k} \frac{N_h}{n_h}\cdot(g_h\cdot\chi^2_h + \frac{1}{2}\cdot g_{hh}\cdot\chi^2_{hh} + \sum_{\ell\neq h} a^2_{h\ell}\cdot g_{h\ell}\cdot\chi^2_{h\ell}). \tag{4.24}$$

Remark 4.6: There is no general guarantee that (4.23) should hold, but it can be expected to hold in "many" (maybe even in "most") situations. We have checked (4.23) empirically for HINK, and found it to hold with very good approximation there. ●

5 On the directing of strata

As discussed in Section 3, the main idea behind the stratification of the population of individuals is that the strata should serve as "directors" towards certain domains of study. In the following discussion, we regard G as a "target" domain and assume that stratum A_q is directed towards G. For simplicity we assume that A_q is the only stratum which is directed towards G. A stratification will, however, usually not be perfectly directed. "Misses" will occur, and we shall distinguish between two types of misses; a miss of type I if a household in G has no adult in A_q, and a miss of type II if an individual in A_q leads to a household outside G. (Misses of types I and II can be viewed as respectively under- and overcoverage when sampling G via A_q.)

Quantification of the number of misses of the two types can be given as follows,

$$\sum_{h\neq q} g_h + \sum_{1\leq h\leq \ell\leq k;\ h,\ell\neq q} g_{h\ell} \quad \text{tells the number of households in G which are misses of type I,} \tag{5.1}$$

while

$$N_q - (g_q + 2g_{qq} + \sum_{\ell\neq q} g_{q\ell}) \quad \text{tells the number of individuals in stratum } A_q \text{ which yield misses of type II.} \tag{5.2}$$

We shall later on give a more quantitative account of how the efficiency of a stratification depends on its "missing" (or positively formulated "hitting") properties. At this stage we confine ourselves to the following qualitative claim, which we believe to sound intuitively very plausible.

> If, ceteris paribus, misses of type I and/or type II are reduced then estimation precision for target domain characteristics are improved. (5.3)

6 Optimization of the survey

When planning a survey with the general structure outlined in Section 3, the statistician has (at least) the following options;

- choice of stratification $\mathcal{S}$ (definitions of strata as well as the number of strata),

- choice of sample alloction $\mathcal{A}$ (among the strata decided upon),
- choice of estimation procedure $\mathcal{E}$. (In our setting this means choosing estimation weights.)

Note that the above choices are "chained"; in the practical situation they must be carried out in the order $\mathcal{S}, \mathcal{A}, \mathcal{E}$.

Our previous "roughly" formulated problem (cf. (3,1)) can now be given the label "How to optimize the chain $(\mathcal{S}, \mathcal{A}, \mathcal{E})$?" When seeking to give this problem a precise formulation we encounter the well known obstacle of "multipurposeness". We refer to Section 7.3 in Kish(1987) for a general discussion of multipurpose design problems, where also further references can be found. We adopt the approach of minimizing the "total imprecision" under given survey resourses. Hence, to make the optimization problem mathematically well posed we notably have to specify an overall criterion for estimation precision, but also to give precise specifications of constraints. Since the last point is the simplest, we start with that.

We lay the following constraints on $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{E}$.

- For $\mathcal{S}$ we make no other assumption than "realizability", i.e. the information which is needed for a stratification should actually be available in the sampling frame. (6.1)
- For the sample allocation $\mathcal{A}$, we assume, for simplicity, fixed sample size, i.e.,

$$n_1+n_2+\ldots+n_k=n \text{ is given.} \qquad (6.2)$$

  In subsequent considerations, the assumption (6.2) could easily be changed to a fixed cost constraint with a cost function which is linear in stratum sample sizes.
- For $\mathcal{E}$ we stick to the assumptions which have been introduced previously; the estimation weights $\underline{a}$ should be stratum combination constant (see (4.11)) and household balanced (see (4.12)) (6.3)

Next we turn to the specification of an overall criterion for estimation precision. Regarding the precision for a single estimator, we employ the usual criterion; the smaller the variance, the better the precision. In our situation we meet the "multipurpose complication" in that we are interested in several domains of study $G_1, G_2, \ldots, G_R$ and (at least possibly) in many different study variables $\underline{x}$. Moreover, we are concerned with different types of population characteristics; μ, g and τ. We shall consider measures of overall estimation imprecision of the following type (recall notation introduced in Remark 4.1),

$$\Psi(V(\hat{\mu}(\underline{x};G_r;\underline{a},\underline{b}))),V(\hat{g}(G_r;\underline{a}^*)),$$
$$V(\hat{\tau}(\underline{x};G_r;\underline{a}^{**})); \; r=1,2,\ldots,R). \qquad (6.4)$$

The choice of a specific overall function Ψ is intricate and probably also controversal. However, for the time being we regard Ψ as decided upon. Thereby our problem is well posed at least from a mathematical point of view, and it runs as follows.

> Find the tripple $(\mathcal{S}, \mathcal{A}, \mathcal{E})$ which minimizes the quantity in (6.4) under the contraints (6.1)-(6.3). (6.5)

In general such an optimization problem is quite messy. In particular we have; For the optimal strategy $(\mathcal{S}_0, \mathcal{A}_0, \mathcal{E}_0)$ all the quantities $\mathcal{S}_0$, $\mathcal{A}_0$ and $\mathcal{E}_0$ will in general depend on

- the $\underline{x}$-variable,
- the domains of study $G_1, G_2, \ldots, G_R$,
- the overall precision criterion Ψ.

In our situation, though, by a strike of good luck the optimization problem simplifies considerably, and the salient result to that effect is presented below. Although this result does not give our optimization problem a one stroke solution, it brings it down to "manageable".

(APPROXIMATE) OPTIMALITY THEOREM: Assume that the sampling fractions, f_h, are small. Then, under general conditions on $\underline{x}$ and G the following estimation weights

$$\tilde{a}_{h\ell}(=\tilde{b}_{h\ell})=f_h/(f_h+f_\ell),$$
$$h,\ell=1,2,\ldots,k, \qquad (6.6)$$

simultaneously minimize all three variances

$$V(\hat{\mu}(\underline{x};G;\underline{a},\underline{b})),\; V(\hat{g}(G;\underline{a})),\; V(\hat{\tau}(\underline{x};G;\underline{a})). \qquad (6.7)$$

Remark 6.1: The estimation weights $\tilde{\underline{a}}$ according to (6.6) will be referred to as the (sampling fraction) proportional weights. •

Remark 6.2: As indicated in the naming of the above theorem, it is not true in a strict mathematical sense. However, the relative differences between the V:s for $\underline{a} = \tilde{\underline{a}}$ and $\underline{a}$ = the truly optimal weights, are so small that the result can be regarded as true from a practical point of view, at least over a wide range of $\underline{x}$:s and G:s. We have checked this claim in the HINK situation, and found the approximation to be good there.

One exception should be pointed to, though. The weights in (6.6) can be distinctly non-optimal for estimation of domain totals, $\tau(\underline{x};G)$ and g(G), for domains G of the following type. G contains a great number of two-adult households with one adult in a low-sample stratum and the other adult in an average/high-sample stratum. •

Remark 6.3: Proofs of the above approximation results can be given along the following lines. Minimize the expressions for $V(\hat{\tau}(\underline{x};G;\underline{a}))$, $V(\hat{g}(G))$ and $V(\hat{\mu}(\underline{x};G;\underline{a},\underline{b}))$ (cf. (4.21)-(4.23)), which are quadratic functions of $\underline{a}$ and $\underline{b}$, under the constraint (4.12), which is linear in $\underline{a}$ and $\underline{b}$. Lagrange's multiplier method leads to a system of linear equations. Then it can be shown that $\underline{a} = \tilde{\underline{a}}$ not only is an approximate solution to the linear system, but also a good approximation to the original optimization problem. We do not give details. •

The most pertinent conclusion we shall draw from the approximation theorem, thereby using it as an "exact" theorem, is stated in (6.10) below. We start with the following observation.

> The optimal weights $\tilde{\underline{a}}$ do not depend on the "nuisance" parameters $\underline{x}$ and G. (6.8)

Next, even if statisticians may disagree on what should be the "proper" choice of the overall criterion Ψ, we presume they do agree that any reasonable Ψ has the following property,

> Ψ is (strictly) increasing in each of its arguments. (6.9)

Under the assumption (6.9), (6.8) leads to the following conclusion.

> When seeking the optimal tripple $(\mathcal{I}, \mathcal{A}, \mathcal{E})$, the estimation part $\mathcal{E}$ can be "factored out" since it has a "universal" solution (which is independent of $\underline{x}$, $G_1, G_2, \ldots, G_R$ and Ψ), namely the solution given by (6.6). (6.10)

In the rest of this paper we assume that (4.23) is satisfied, and hence that (4.24) applies. By inserting $\underline{a} = \tilde{\underline{a}}$ into (4.24) and paying regard to (6.10) the following result is obtained after some straightforward algebra.

For a given stratification and a given sample allocation, the variance of the (universally) optimal domain mean estimator $\hat{\mu}(\underline{x};G;P) := \hat{\mu}(\underline{x};G;\tilde{\underline{a}};\tilde{\underline{a}})$ (P for proportional) is

$$V(\mu(\underline{x};G;P)) = \frac{1}{g(G)2} \{ \sum_{h=1}^{k} \frac{g_h \cdot x_h^2 + \frac{1}{2} g_{hh} \cdot x_{hh}^2}{n_h/N_h} + \sum_{1 \le h < \ell \le k} \frac{g_{h\ell} \cdot x_{h\ell}^2}{\frac{n_h}{N_h} + \frac{n_\ell}{N_\ell}} \}. \quad (6.11)$$

For HINK, there has been uncertainty and debate how the P-versions of the estimators compare with the half weighted versions, in the sequal denoted H-versions (see (4.8)). The above approximation theorem tells that the P-versions never perform worse than the H-versions, but so far we have not given any quantitative measure of how much optimality pays. It is therefore of interest to have an expression also for the variance of $\hat{\mu}(\underline{x};G;H)$. Such an expression is obtained by setting $a_{h\ell}=1/2$ in (4.24). We give the resulting formula in a somewhat implicite fashion, which has the merit that it clearly shows that $\hat{\mu}(\underline{x};G;P)$ is superior to $\hat{\mu}(\underline{x};G;H)$ (as it should be according to the approximation theorem). It also gives a quantitative expression of the amount of variance reduction the P-version gives compared with the H-version. The following formula is readily obtained from (4.24) and some algebra, which we omit.

$$V(\hat{\mu}(\underline{x};G;H)) = V(\hat{\mu}(\underline{x};G;P)) + \frac{1}{4g(G)2} \sum_{1 \le h < \ell \le k} \frac{\left(\frac{N_h}{n_h} - \frac{N_\ell}{n_\ell}\right)^2}{\frac{N_h}{n_h} + \frac{N_\ell}{n_\ell}} g_{h\ell} \cdot x_{h\ell}^2. \quad (6.12)$$

Analogous formulas can be obtained for the P- and H-versions of $V(\hat{\tau})$ and $V(\hat{g})$ by insertion into (4.21) and (4.22).

The variance formulas which are written out, respectively indicated, above provide tools for theoretical analyses of optimal allocation and optimal stratification under the present design.

7 Some numerical illustrations

The full paper concluded with a fairly elaborate presentation of various numerical findings related to the HINK survey. The main aims with the numerical illustrations are as follows; (i) To illustrate the use of the formulas in Section 4-6. (ii) To shed some light on the following general questions;

- How is a good sample allocation found?
- How is a good stratification found?
- How do proportional and half weighted estimators compare with each other?

As space is limited, we must exclude the numerical illustrations in the present version of the paper. However, the author will be glad to send a copy of the full paper upon request.

Acknowledgements. I want to thank Judith Lessler, Lennart Nordberg and Martin Ribe for valuable comments on drafts of the paper, Judith in particular for information on relevant literature.

References:

Birnbaum, Z. W. & Sirken, M. G. (1965). Design of sample surveys to estimate the prevalance of rare diseases: Three unbiased estimators. U.S. Department of Health, Education and Welfare 12, no 11.

Kish, L. (1987). Statistical design for researchers. John Wiley & Sons, New York.

Levy, P. S. (1977). Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. JASA 72, 758-763.

Sirken, M. G. (1972). Stratified sample surveys with multiplicity. JASA 67, 224-227.