# On Evaluation of Surveys with Samples from the Revised Zimbabwe Master Sample Frame

Bengt Rosén

SCB

R&D Report
Statistics Sweden
Research-Methods-Development
1989:15

| | |
|---|---|
| Från trycket | Augusti 1989 |
| Producent | Statistiska centralbyrån, Utvecklingsavdelningen |
| Ansvarig utgivare | Staffan Wahlström |
| Förfrågningar | Bengt Rosén, tel. 08-783 44 90 |


ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1989

INLEDNING

TILL

**R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.**
**Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.**

**Föregångare:**

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

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

**Efterföljare:**

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

R & D Report 1989:15. On evaluation of surveys with samples from the revised Zimbabwe Master Sample Frame / Bengt Rosén.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1989-X101OP8915

# ON EVALUATION OF SURVEYS WITH SAMPLES FROM THE REVISED ZIMBABWE MASTER SAMPLE FRAME

by

Bengt Rosén

Följande material har tidigare presenterats som del av en Mission Report i SCB International Consulting Offices ZIMSTAT-serie, nämligen som Part 2 i ZIMSTAT 1989:5 "Sampling and Estimation in the Zimbabwe Household Survey Programme II.

ABSTRACT

In Zimbabwe, as in most countries which participate in UN:s National Household Survey Capability Programme, household surveys are conducted with samples from a Master Sample Frame. The Zimbabwe frame was updated/revised in 1987 to the Revised Zimbabwe Master Sample (RZMS). As described, the revision was carried out in a somewhat unorthodox way, leading to some estimation problems of a non-standard nature.

The main aim with the report is that it should provide an estimation manual for surveys with samples from the RZMS, giving formulas for point estimation as well as for estimation of sampling errors. In particular the report comprises parts where the formulas are presented in a way which hopefully is well adapted to EDP implementation of the procedures.

The point estimation formulas can roughly be characterized as particular cases of the "standard formulas" for area sampling. As regards variance estimation, we consider a variant of the method of ultimate clusters/random groups. Some general background theory relating to that method is presented.

# ON EVALUATION OF SURVEYS WITH SAMPLES FROM THE REVISED ZIMBABWE MASTER SAMPLE FRAME.

Contents.

## 1. AN OUTLINE OF THE ZIMBABWE MASTER SAMPLE FRAME AND ITS REVISION.

The first census in Zimbabwe after Independence was carried out by the Central Statistical Office (CSO) in 1982. To enable a subsequent flow of statistics on various aspects of the development in the country, Zimbabwe decided in 1982 to embark on UNs National Household Survey Capability Programme, in its Zimbabwe version named ZNHSCP. The ZNHSC programme included a plan for a series of integrated households surveys on different topics during the period until the next census in 1992 and this plan has since 1982 been guiding for the household surveys carried out by CSO. A main vehicle for the realization of the ZNHSC-programme was the Zimbabwe Master Sample frame (ZMS), which was established in 1983 on the basis of the 1982 census results. As an instrument for achieving good design for the ZNHSCP surveys and in particular a good design for the master sample, an extensive pilot survey was carried out in the first half of 1983.

Household samples from the ZMS can broadly be described as three-stage samples. In the census Zimbabwe was divided into geographical enumeration areas (EAs), for which i.a. enumeration and listing of households were carried out. The primary sampling units in the three-stage sampling, called divisions/-subdivisions or simply PSUs, were created by joining adjacent EAs so as to obtain areas containing approximately 4000 households. In the first stage, a stratified sample of divisions/-subdivisions was selected. The secondary sampling units, called segments or simply SSUs, were obtained by partitioning the selected divisions into subareas containing close to 100 households each. From the divisions/subdivisions selected in the first stage, two or three segments were drawn at random. The so selected segments were visited, they were mapped and lists of their households were made up. The segment household lists then provided frames for selection of the ultimate (third stage) sampling units, the households. The general as well as the detailed structure of the ZMS frame, including various optimality considerations, were worked out by CSO under the assistance of UN consultant M. Tin. An elaborate presentation of the ZMS is given in a CSO report of December 1986.

Next we give a brief review of the probabilistic structure of the original ZMS. The collection of PSUs was stratified with sampling strata formed by the cells in the cross-classification of Zimbabwe's 8 provinces with its 6 "sectors" (reflecting types of administrative areas/land use). However, the most sparsely populated of the 8x6 = 48 potential sampling strata were disregarded and only 30 of them were used as effective sampling strata. From the (effective) sampling strata, independent samples of PSUs were drawn, with inclusion probabilities proportional (within sampling stratum) to size, size being the census number of households in the PSU. In the second stage, two or three segments were drawn from each PSU by systematic (equal probability) sampling. The collection of segments selected by this procedure constituted the master sample.

In the third and final stage, household samples for ZNHSCP surveys were drawn by systematic sampling from the segment household lists. The sizes of these final sub-samples have varied from survey to survey, and some of them have in fact covered all households in the segments.

Already from the outset of the ZNHSC-programme it was decided that the ZMS frame would be updated during the period between the 1982 census and the following census in 1992, the main reasons for the updating being the customary ones;

- The segment household lists would become increasingly inaccurate over time.

- The households in the ZMS would suffer an excessive response burden if the master sample remained unchanged during the whole 10-year period between the censuses.

According to the ZNHSCP planning, the updating of the master sample frame was to be performed in conjunction with (or maybe rather as part of) the Intercensal Demographic Survey (ICDS) which was carried out in the midst of the intercensal period. The aims of the ICDS were formulated as follows;

- To update the sampling units and design currently used in Household Surveys.

- To give statistics on population and on demographic and socio-economic variables.

- To serve as a pilot study for the 1992 Census.

The ICDS consisted of three rounds, called Rounds 0, 1 and 2. Round 0 was specially devoted to the updating/revision of the sampling frame. It covered enumeration and listing of the households in the EAs which were "related" to the ZMS, and resulted in one of the versions of the Revised Zimbabwe Master Sample frame (RZMS). The household sample for the demographic/-socio-economic part of the ICDS (Rounds 1 and 2) was then drawn from the RZMS.

To assist in the planning of the updating of the ZMS and in the planning of the ICDS in general, another UN consultant, C. Scott, visited CSO in the spring of 1987. Scott wrote an instructive report (Scott, May 1987) on the status of the ZMS and on the problems which were faced in the task of updating it. The report gives i.a. a review of the history of the ZMS. A somewhat disturbing part of that history is the fact that some vital documentation on the "early ZMS" in fact no longer was available, with the effect that the ZMS updating programme faced intricate obstacles. We quote Scott;

> "It would of course be possible to abandon the ZMS entirely and select a new master sample from the census EAs or simply decide to use unrelated ad hoc surveys until the next census in 1992. The main argument for a master sample in Zimbabwe is the immobility of the enumerators.

> ..... The consultant therefore recommends as a broad objective an attempt to stay with the existing arrangements as far as possible, until the next census. This raises the question how far it is possible to repair the deficiencies of the ZMS noted above."

However, Scott not only pointed out weaknesses in the ZMS at the current time, but also gave thorough and detailed suggestions for how the deficiencies could be repaired, at least under pragmatically reasonable assumptions concerning undocumented alterations in the original ZMS and some other matters. In our opinion, Scott's advise for how to establish a revised master sample seems competent and wise, and this was obviously also CSO's viewpoint. As we understand it;Scott's report, which originally had the status of suggestions, in fact achieved the status of manual for sampling and estimation procedures to be employed for surveys with samples from the Revised Zimbabwe Master Sample frame (RZMS), in particular for the ICDS. Therefore, in the following discussion of the ZMS and RZMS, we will accept Scott's suggestions and leave out the reservations which Scott adds to them. The reader who is interested in the more detailed arguments is referred to Scott's report.

The revised master sample, RZMS, differs in various respects from the original one. In (point) estimation contexts one can chose to forget about the "heritage" from the ZMS and look at the RZMS as a brand new master sample frame which leads to household samples in a two-stage procedure with EAs or segments as first stage sampling units and households as second stage units. However, when it comes to estimating sampling errors one no longer has an option in how to view a sample from the RZMS, the "full history" has to be taken into account and a RZMS household sample must be viewed as a as a three-stage sample with successive sampling units; divisions/subdivisions, EAs or segments and finally households.

As Scott's report has become a most valuable guideline for the construction of the RZMS and for evaluation of surveys based on samples from it, one may wonder if there is anything more to say on the matter? For a number of reasons we think there is, the following being the main ones;

- Scott's report was written prior to the ICDS, and hence prior to the actual establishment of the RZMS. Although his suggestions were followed in their essentials, there is a need for documentation of the RZMS in its precise realization.

- Scott draws up the general lines for estimation in surveys with samples from the RZMS. His suggestions are maybe sufficient for a statistician, but they are not detailed enough to serve as a basis for the programming work for the processing of data from RZMS surveys in general.

- Although Scott acknowledges the value of estimating sampling errors, he does not enter into a technical discussion of the matter.

1.4

The main aims of the present report are to fill the gaps indicated above. Furthermore, by omitting most of the discussion in Scott's report, although enlightening and valuable, a more concise and comprehensive presentation of the sampling theory for RZMS surveys is hopefully achieved. However, in some respects this report covers less than that of Scott. He also presents various suggestions on how to evaluate surveys with "original" ZMS samples, for which the data are already collected but not yet processed. This report does not touch upon that problem.

We conclude this section with an outline of the rest of the report. In Section 2 we introduce various concepts, terminology and notation relating to the populations under consideration. In Section 3 we specify the probabilistic structure of samples from the RZMS, and in Section 4 we consider the organization of observed data for RZMS surveys. Section 5 deals with (point) estimation for RZMS surveys, while the estimation of sampling errors is treated in Section 6. Section 7 gives a more detailed account of the ICDS survey which is a specific, and in fact the first, RZMS survey. In Appendices 1, 2 and 3 we have collected material of a more theoretical nature.

## 2. SOME TERMINOLOGY AND NOTATION RELATING TO THE POPULATIONS OF HOUSEHOLDS AND INDIVIDUALS.

As we are aiming at precise estimation formulas we must at some place introduce clear and unambiguous terminology and notation for various population concepts, and we chose to do that already at this stage. The following definitions, which are a bit lengthy, can be characterized as the "usual ones". Therefore, the reader who primarily is interested in the revised master sample frame RZMS can skip this section in the first round and proceed to Section 3.

For RZMS surveys there are two natural populations, namely

U = the population of households (in the country),
V = the population of individuals (in the country).

The household population is of relevance in all ZNHSCP surveys. Whether the population of individuals is of interest or not depends on the topic of the specific survey. The ICDS is an example of a survey for which the population of households and the population of individuals both are of interest. In the following we shall introduce various terminology and notation relating to the populations U and V, and we start with the household population.

### 2.1. On the population of households.

We chose a labelling system for the objects in U, i.e. households, which will fit with our "basic view" on RZMS samples (to be elaborated in Section 3), namely as two-stage samples with stratification at the first stage.

We assume that the country is (or at least is imagined to be) partitioned into disjoint areas which we call primary areas (PAs). As will be discussed in more detail in Section 3, we shall be concerned with two different concrete options for the primary areas; either the enumeration areas (EAs) in the 1982 census or the segments in an imagined segmentation of the whole country along the lines used in the RZMS segmentation. In order to keep both options open under one and same word, we use the general term "primary area".

Households are associated with the primary areas in which they reside (according to some prescribed residence rule).

The collection of all primary areas is divided into sampling strata, and the letter h is used as stratum label. Let

H = the number of sampling strata. (2.1)

Furthermore let

$N_h$ = the number of PAs in stratum h, h=1,2,...,H. (2.2)

The letter i is used to label PAs within a sampling stratum, and we refer to PA no i in stratum no h as PA no (hi). Let

$$M_{hi} = \text{the number of households in PA no (hi).} \qquad (2.3)$$

The letter j is used to label households inside PAs, and household no j in PA no (hi) is referred to as household no (hij). Hence, the household population U can be written

$$U = \{(hij);\ j=1,2,..,M_{hi},\ i=1,2,..,N_h,\ h=1,2,..,H\}. \qquad (2.4)$$

By a (household) variable $\underline{x}$ we mean that a number/characteristic is associated with each household in the population U, and the value for household (hij) is denoted by $x_{hij}$. The variable $\underline{x}$ is the collection of all these variable values, i.e. $\underline{x} = \{x_{hij};(hij)\epsilon U\}$, or in full

$$\underline{x} = \{x_{hij};\ j=1,2,..,M_{hi},\ i=1,2,..,N_h,\ h=1,2,..,H\}. \qquad (2.5)$$

The total of the variable $\underline{x}$ (over the entire population U), denoted $\theta(\underline{x})$, is

$$\theta(\underline{x}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} x_{hij}. \qquad (2.6)$$

The population total in (2.6) can be partitioned into the following subtotals, which we call stratum-totals,

$$\theta_h(\underline{x}) = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} x_{hij}, \quad h=1,2,..,H, \qquad (2.7)$$

and we have

$$\theta(\underline{x}) = \sum_{h=1}^{H} \theta_h(\underline{x}). \qquad (2.8)$$

A domain (sometimes called domain of study) in the household population U, often also referred to as a group in the population, is a (specified) subset of U. In general we denote domains/groups by D.

A basic type of characteristic is a variable total over a domain/group. The following domain/group indicator function will be useful when dealing with domain totals,

$$\underline{1}_D(hij) = \begin{cases} 1 & \text{if household (hij) belongs to the domain D,} \\ 0 & \text{otherwise.} \end{cases} \qquad (2.9)$$

The total for the variable $\underline{x}$ over the domain/group D, denoted $\theta(\underline{x};D)$, is

$$\theta(\underline{x};D) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} x_{hij} \cdot \underline{1}_D(hij). \qquad (2.10)$$

In analogy with the partitioning in (2.8), a domain total can be partitioned into domain/group D totals over strata,

$$\theta_h(\underline{x};D) = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} x_{hij} \cdot \underline{1}_D(hij), \quad h=1,2,\ldots,H. \tag{2.11}$$

We have,

$$\theta(\underline{x};D) = \sum_{h=1}^{H} \theta_h(\underline{x};D). \tag{2.12}$$

Another characteristic of particular interest is the size of a domain/group, denoted g,

$$g(D) = \text{the number of households in the domain } D. \tag{2.13}$$

The size of a domain can be regarded as a domain total, namely the domain total corresponding to the variable

$$\underline{1} = \text{the variable which gives the value 1 to each household in the population.} \tag{2.14}$$

We have,

$$g(D) = \theta(\underline{1};D) = \theta(\underline{1}_D) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \underline{1}_D(hij). \tag{2.15}$$

A further basic type of population characteristics is the mean of the variable $\underline{x}$ over the domain/group D, denoted $\mu(\underline{x};D)$,

$$\mu(\underline{x};D) = \theta(\underline{x};D)/g(D) = \theta(\underline{x};D)/\theta(\underline{1};D). \tag{2.16}$$

If the domain D is the entire population, i.e. D = U, then the corresponding domain mean is called the population $\underline{x}$-mean, denoted $\mu(\underline{x})$.

A special case of mean is proportion. For a (specified) group A in the population U, the proportion of A-households in the domain D, denoted p(A;D), is

$$p(A;D) = g(A \cap D)/g(D). \tag{2.17}$$

This proportion can also be viewed as the domain mean for the variable $\underline{x}=\underline{1}_A$, i.e.

$$p(A;D) = \mu(\underline{1}_A;D). \tag{2.18}$$

When D is set to the whole population, the corresponding population proportion is denoted simply by p(A).

2.2. On the population of individuals.

Each household consists of a number of individuals. We use the letter k to label individuals within a household. Individual no k in household no (hij) is referred to as individual no (hijk). Let

$$K_{hij} = \text{the number of individuals in household (hij).} \tag{2.19}$$

Hence the population of individuals, V, can be written

$$V = \{(hijk);\ k=1,2,..,K_{hij},\ j=1,2,..,M_{hi},\ i=1,2,..,N_h,\ h=1,2,..,H\}. \tag{2.20}$$

The notions of individual-variable $\underline{x}$ and domain/group in the population of individuals are defined in analogy with the definitions of variable and domain for households. An individual-variable $\underline{x}$ means that a number/characteristic is associated with each individual in the population V, and the value associated with individual (hijk) is denoted $x_{hijk}$. Hence we have

$$\underline{x} = \{x_{hijk};\ k=1,2,..,K_{hij},\ j=1,2,..,M_{hi},\ i=1,2,..,N_h,\ h=1,2,..,H\}. \tag{2.21}$$

A domain/group of individuals is a subset of the population V of individuals. The notions of variable totals over the population and over domains/groups, domain/group sizes and domain/-group means are defined in analogy with the corresponding concepts for households. We use the same notation in the individuals case as in the household case. For completeness, and for future use we write down the formulas which are straightforward analogies of the previous formulas for households.

The total of the variable $\underline{x}$ (over the entire population), denoted $\theta(\underline{x})$, is

$$\theta(\underline{x}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \sum_{k=1}^{K_{hij}} x_{hijk}, \tag{2.22}$$

and the corresponding sampling stratum totals are

$$\theta_h(\underline{x}) = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \sum_{k=1}^{K_{hij}} x_{hijk}. \tag{2.23}$$

We have,

$$\theta(\underline{x}) = \sum_{h=1}^{H} \theta_h(\underline{x}) . \tag{2.24}$$

Variable totals over different domains/groups will constitute a basic type of characteristic also in the individuals context,

and domain/group indicator functions will be useful when dealing with domain totals,

$$\underline{1}_D(hijk) = \begin{cases} 1 & \text{if individual } (hijk) \text{ belongs to the domain } D, \\ 0 & \text{otherwise.} \end{cases} \quad (2.25)$$

The total for the variable $\underline{x}$ over the domain/group D, denoted $\theta(\underline{x};D)$, is

$$\theta(\underline{x};D) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \sum_{k=1}^{K_{hij}} x_{hijk} \cdot \underline{1}_D(hijk). \quad (2.26)$$

The domain/group D totals over the sampling strata are

$$\theta_h(\underline{x};D) = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \sum_{k=1}^{K_{hij}} x_{hijk} \cdot \underline{1}_D(hijk). \quad (2.27)$$

We have

$$\theta(\underline{x};D) = \sum_{h=1}^{H} \theta_h(\underline{x};D) \ . \quad (2.28)$$

The size of a domain/group, denoted g, is

$$g(D) = \text{the number of individuals in the domain D.} \quad (2.29)$$

As before, the size of a domain can be regarded as a domain total, namely the domain total corresponding to the variable

$$\underline{1} = \text{the variable which gives the value 1 to each individual in the population.} \quad (2.30)$$

We have

$$g(D) = \theta(\underline{1};D) = \theta(\underline{1}_D) = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \sum_{k=1}^{K_{hij}} \underline{1}_D(hijk). \quad (2.31)$$

The mean of the variable $\underline{x}$ over the domain/group D, denoted $\mu(\underline{x};D)$, is

$$\mu(\underline{x};D) = \theta(\underline{x};D)/g(D) = \theta(\underline{x};D)/\theta(\underline{1};D). \quad (2.32)$$

If the domain D is the entire population V, the corresponding mean is called the population $\underline{x}$-mean, and is denoted by $\mu(\underline{x})$.

A particular case of mean is proportion. For a (specified) group A in the population V, the proportion of A-individuals in the domain/group D, denoted p(A;D), is

$$p(A;D) = g(A \cap D)/g(D). \quad (2.33)$$

This proportion can also be viewed as the mean value for the variable $\underline{x}=\underline{1}_A$, i.e.

$$p(A;D) = \mu(\underline{1}_A;D). \quad (2.34)$$

A proportion in the whole population (i.e. when D = V) is denoted p(A).

## 3. ON THE REVISED MASTER SAMPLE.

Our aim in this section is to describe the Revised Zimbabwe Master Sample frame, in particular the probabilistic structure of household samples from it. In Section 1 we gave a brief description of the operations included in the revision of the original master sample frame, the ZMS. We start here with a somewhat more elaborate description of the ideas and procedures of the revision.

### 3.1. On the construction of the revised master sample.

As stated in Section 1, a vital instrument for the ZNHSC-programme was the Zimbabwe Master Sample (ZMS) which was established in 1983. Already at the outset of the ZNHSCP it was decided that the ZMS would be updated during the period between the 1982 census and the following census in 1992, and that the updating was to be implemented as Round 0 in the Intercensal Demographic Survey (ICDS) to be carried out in the midst of the intercensal period. The main reasons for the updating were the customary ones;

- The segment household lists in the ZMS would become increasingly inaccurate over time.

- The households in the ZMS would suffer an excessive response burden if the master sample remained unchanged during the whole 10-year period between the censuses.

The latter reason led to a desire for an exchange of (at least) the secondary sampling units (SSUs) in the ZMS, i.e. the segments. An aspect which pulled in the direction of making as small changes as possible was the logistical one. In order not to give the enumerators too long travel distances (alternatively to have to fire experienced enumerators and recruit new ones) it was considered desirable that segment exchanges were carried out within comparatively small geographical areas.

From a sampling theoretical point of view, such a constrained SSU-exchange would perhaps most naturally have been carried out by letting the primary sampling units (PSUs) remain fixed, and by selecting new segments within the PSUs. However, for different reasons this was not feasible. We shall not go into details, just refer to the report by Scott (1987). What actually was done in the updating/revision of the sampling frame, was to let (single) EAs be "fix-objects" and then select new segments within the given EAs. This procedure may sound a bit surprising since (single) EAs did not occur as sampling units in the selection of the ZMS. To explain how this type of procedure could lead to the desired goal, we shall first say some words on the construction of the original sampling frame, the ZMS.

The ZMS was a master sample of segments, equipped with segment household lists. The segments in the ZMS were selected as random sub-areas of the PSUs, the latter being called divi-

sions/subdivisions. We shall not enter into the background for the term division/subdivision, only note that it is quite cumbersome. Therefore, from now on we use the simpler term (sub)division. The (sub)divisions were formed by joining a number of adjacent EAs. In Figure 3.1 we illustrate a (sub)-division in the ZMS with its EAs and ZMS-segments (here two). Typical sizes of the different types of areas were; a (sub)-division contained roughly 4000, an EA roughly 800 and a segment close to 100 households.

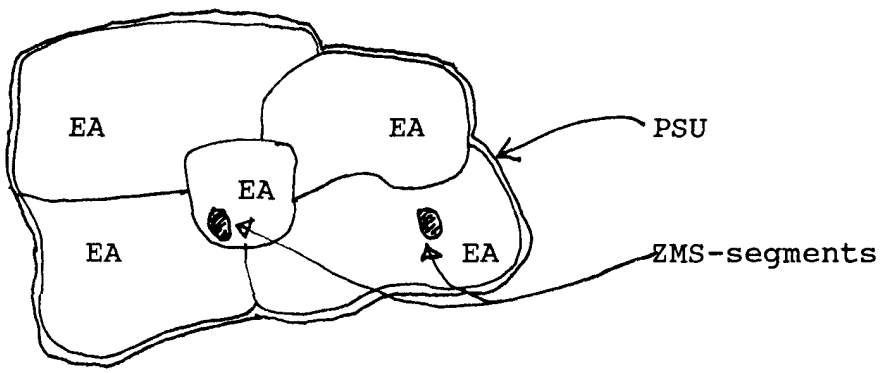


Fig.3.1. A PSU ((sub)division) in ZMS with its EAs and ZMS-segments.

We can make the following observations;

(i). The ZMS-segments do not cross EA-boundaries, i.e. they lie entirely inside one EA.

(ii). The ZMS-segments lie in different EAs.

Properties (i) and (ii) are not particular for just Fig.3.1, they were characteristic for the entire ZMS, a fact which was due to the selection methods used. Firstly, the ZMS-segments in a PSU were selected by systematic sampling within the PSU which "forced" the ZMS-segments to lie well separated. Secondly, it was part of the segmentation instructions that the ZMS-segments should be constructed so as to lie within one EA.

Hence, the segments in ZMS were in one-to-one correspondence with EAs, namely the EAs in which they were located. Accordingly, even if the ZMS was set up as a sample of segments, in view of the one-to-one correspondence between segments and EAs it could also be viewed as a sample of (single) EAs, with the EAs sampled "implicitly" via the segments. This view on the ZMS, as a sample of EAs, was the basis for the updating/revision procedures leading to the Revised Zimbabwe Master Sample (RZMS). By regarding the EAs in ZMS as "pseudo-PSUs", segments could be changed within the EAs instead of within original PSUs (i.e. (sub)divisions).

In Round 0 of the ICDS, household lists for the entire EAs (not only for segments) were made up for all the EAs in the ZMS at the time for the ICDS, i.e. in the first half of 1987. This

set of EAs will be referred to as the full EA-content of the RZMS. Hence, updated household lists for the RZMS EAs are available, and thereby households can be sampled from the RZMS by the following two-stage procedure; In the first stage a sample of EAs is drawn from the full EA-content in the RZMS, and in the second stage households are sampled from the household lists for the selected EAs. In fact this type of sampling procedure, with the full set of EAs, was employed for the ICDS.

However, as an EA often constitutes a fairly large area, a household sample from an EA can be quite dispersed, which in turn leads to transportation difficulties for the interviewers. This was a main reason for the desire to establish a more "compact" version of the RZMS, the segment version. With the aid of the EA household lists, one segment (with close to 100 households) was selected at random in each EA in the RZMS. The segments were visited, mapped and household lists were made up for them. Having these segments with their household lists, samples of households can be also be drawn from the RZMS by the following two-stage procedure; In the first stage a sample of segments is drawn from the full segment-content in the RZMS and in the second stage households are sampled from the household lists for the selected segments. We shall distinguish the two modes for generating household samples from the RZMS by talking of household samples via EAs respectively household samples via segments. However, segments with household lists have not yet been established in all the EAs in the RZMS, only in those EAs which are located in communal lands. The reason for starting there is that the annual Agriculture and Livestock Survey (ALS) concerns communal lands, and communal lands only. In that survey, household samples via segments are used. The idea is that most of the remaining surveys within the ZNHSCP shall use via-segment samples and therefore the full segmentation of the RZMS will soon be completed. Henceforth we presume that the RZMS in fact is fully segmented.

So far the idea of using EAs as pseudo-PSUs works without complications. However, obstacles turn up when one starts to ask about the inclusion probabilities for the sampled units. E.g., for computation of point estimates on the basis of observations from a RZMS-samples one needs to know the inclusion probabilities of the first order to determine the estimation weights. As single EAs did not enter as sampling units in the original ZMS, the inclusion probabilities of the EAs in the RZMS are in principle unknown. However, the pragmatic reasoning suggested by Scott can lead us to numerical values for the first order inclusion probabilities for the EAs as well as for the segments in the RZMS. As long as we are only interested in point estimation, our basic view on household samples from RZMS (via EAs as well as via segments), is that they are two-stage samples with EAs or segments as first-stage sampling units and households as second-stage sampling units.

However, when it comes to second order properties of the estimators, such as their variances, the dependencies between the observations in the sample (or equivalently the second order

inclusion probabilities for the sample) play a crucial role. In such contexts the descriptions of the household samples as two-stage samples no longer suffice. The two-stage descriptions (deliberately) conceal an aspect on the sampling process which generated the RZMS and which is highly relevant for understanding dependencies among the sample observations.

The aspect in question is that the original ZMS was drawn with one stage more than is accounted for in the two-stage descriptions. The segment sample in the original ZMS was selected by first sampling (sub)divisions and then two or three segments from the selected (sub)divisions. As a (sub)division contained around 4000 households, in the perspective of the country it was a small geographical area. Hence for most household and individuals variables, fairly strong (positive) correlations can be expected among observations on households in EAs from the same (sub)division. This correlation does not affect the unbiasedness of point estimators, but in variance estimation contexts it can not be disregarded. As a consequence, when estimating sampling errors it is essential to pay regard to RZMS's "inheritance" from the ZMS, and to keep track of which EAs/segments in the RZMS that emanate from the same (sub)division in the ZMS. For variance estimation purposes one must rely on descriptions of the probabilistic structure of samples from the RZMS which are more elaborate than the previous two-stage descriptions.

The two-stage description is presented in Section 3.2, while the more elaborate description is given in Section 3.4. A further terminological comment; Henceforth we use the term "general household sample" from the RZMS as opposed to the type of household sample that was used for the ICDS, which was "particular" in the following two respects; It was self-weighting, and furthermore it used all the EAs in the RZMS (i.e. no sub-sampling was employed).

## 3.2. Two-stage description of household samples from the RZMS.

From an operational point of view it of course most natural to view the RZMS as the Master Sample. However, when one wants to discuss sampling theoretical details for the RZMS, it is in fact clearest to viewed it as containing two different master samples frames given by a master sample of EAs and a master sample of segments. In the following we distinguish between the two master sample frames by referring to them as RZMS(EA) respectively RZMS(Segm).

The RZMS(EA) consists of a collection of enumeration areas (EAs), with EA as defined in the 1982 census. For each EA in the RZMS(EA) there is available

> a complete, per May/June 1987, list of the households in the EA. (3.1)

> By a general RZMS(EA) household sample, also referred to as a household sample via EAs, we mean a sample of households which is obtained as follows. In the first

stage one draws a sub-sample of the EAs in the full RZMS(EA), and in the second stage one selects households from the household lists for the EAs which were selected in the first round. (3.2)

The RZMS(Segm) consists of a collection of areas called segments. For each segment in the RZMS(Segm) there is available

a complete list of the households in the segment. (3.3)

By a general RZMS(Segm) household sample, also referred to as a household sample via segments, we mean a sample of households which is generated as follows. In the first stage one draws a sub-sample of the segments in the full RZMS(Segm), and in the second stage one selects households from the household lists for the segments which were selected in the first round. (3.4)

As already indicated, the segments in the RZMS(Segm) were generated as follows. By using the EA household lists, the EAs in the RZMS(EA) were segmented i.e. partitioned into subareas, called original segments, containing close to 100 households each. Set

$d_{hi}$ = the number of original segments in EA no (hi). (3.5)

Then, for each EA in the RZMS(EA) one original segment was selected at random (= with equal probabilities). The segments so selected constitute the segments in RZMS(Segm). The selected segments were visited and maps over them and lists of their households were established. As mentioned, the main reason for introducing the segment-frame in addition to the EA-frame is that segments are more "compact" then EAs, and thereby more convenient to survey from the logistic point of view. Note that there is a one-to-one correspondence between the EAs in the RZMS(EA) and the segments in the RZMS(Segm). We will refer to either or both in a "couple" by the term EA/segment. In particular, the two master sample frames contain the same number of primary units, namely 273.

Remark 3.1: In fact it would be possible also to generate household samples with first-stage samples which extend the number of EAs (or segments) over that in the RZMS(EA) (or RZMS(Segm)). (Cf.Lemma A2.2 in Appendix 2.) Such a desire would, however, lead to rather laborious operations. In particular it would require new household listings in the "extra" EAs (segments). We believe that there will be little demand for such "extended" household samples, and therefore we do not pursue the topic. ▒

For specification of the inclusion probabilities for samples from the RZMS we must take into account that, as an "inheritance" from the ZMS, the samples of EAs and segments in the RZMS were obtained by stratified sampling. The corresponding sampling strata were formed as the "cells" in the cross-classification of Zimbabwe's 8 provinces;

- Manicaland
- Mashonaland Central
- Mashonaland East
- Mashonaland West
- Matabeleland North
- Matabeleland South
- Midlands
- Masvingo

with its 6 main types of administrative areas/land use areas, henceforth referred to as sectors;

communal lands areas,
large scale commercial farming areas,
urban & semi-urban areas,
resettlement areas
small scale commercial farming areas,
forests, parks, others.

Hence, in all there were 8x6 = 48 possible strata. However, all of them are not represented by EAs in the RZMS(EA) respectively segments in RZMS(Segm). As the main aims for the ZNHSCP surveys is to provide information on demographic and economic circumstances, no primary sampling units were drawn from the very sparsely populated sampling strata. In particular, the whole sector "forests, parks, others" was excluded. The number of effective sampling strata in the RZMS, i.e. strata with non-zero sample size, is 30. The undercoverage caused by the exclusion of some of the sampling strata is roughly 2% of the total population. In the sequel we disregard this undercoverage problem, and we reason as if it did not exist (or as if Zimbabwe is equivalent with the sub-part of the country which is made up by the effective sampling strata).

Independent PSU-samples were drawn from the (effective) sample strata. The PSU sample sizes in the different sampling strata were for the ZMS decided upon on the basis of the results from the pilot survey which preceded the ZMS. The EA/segment sample sizes in the RZMS reflects the sample allocation in the original ZMS. However, for the present purposes we need not, and shall not go into details on this type of master sample design considerations. The EA/segment sample sizes in the RZMS are exhibited in Table 3.1 below.

| Province | Sector | | | | | | |
|---|---|---|---|---|---|---|---|
| | Communal lands | Commercial farming | Urban, semi-urban | Resettlements areas | Small scale farming | Forest park other | Total |
| Manicaland | 26<br>147281 | 5<br>40521 | 4<br>25736 | 2<br>10123 | 1<br>5170 | 0<br>0 | 38<br>228831 |
| Mashonaland Central | 18<br>62728 | 4<br>34460 | 3<br>12344 | 0<br>1375 | 0<br>2541 | 0<br>0 | 25<br>113448 |
| Mashonaland East | 22<br>96778 | 4<br>25346 | 24<br>232408 | 1<br>3373 | 0<br>2496 | 0<br>0 | 51<br>360401 |
| Mashonaland West | 14<br>51908 | 8<br>79938 | 8<br>43165 | 0<br>2904 | 1<br>3810 | 0<br>535 | 31<br>182260 |
| Matabeleland North | 17<br>60381 | 2<br>8815 | 15<br>116075 | 0<br>1379 | 0<br>341 | 0<br>2510 | 34<br>189501 |
| Matabeleland South | 19<br>67351 | 3<br>25154 | 4<br>6543 | 0<br>524 | 0<br>2460 | 0<br>55 | 26<br>102087 |
| Midlands | 24<br>130401 | 2<br>15526 | 6<br>56180 | 0<br>4745 | 1<br>4263 | 0<br>25 | 33<br>211140 |
| Masvingo | 26<br>148853 | 4<br>29768 | 4<br>19486 | 0<br>2912 | 1<br>5200 | 0<br>322 | 35<br>206541 |
| Total | 166<br>765681 | 32<br>259528 | 68<br>511937 | 3<br>27335 | 4<br>26281 | 0<br>3447 | 273<br>1594209 |

Table 3.1. The first-stage sampling strata for the RZMS. The upper figures in the cells state the number of EAs/segments selected from the sampling stratum. In particular, the 0:s tell which sampling strata that were omitted. The lower figures state the number of households (according to the 1982 census) in the sampling strata.

In line with the notation in Section 2, we use h,i, j and k to label respectively; sampling strata, EAs/segments, households and individuals. However, here we confine the indices to run over the sample, i.e. h runs over the effective sampling strata, i runs within sampling stratum over the selected EAs/segments, j runs within EA/segment over the selected households and k runs within household over the members in the selected households.

Let

$a_{h0}$ = the EA/segment sample size in sampling stratum h in the full RZMS, as specified by the upper figures in Table 3.1, h=1,2,...,H. (3.6)

Set

$M_{hi}$ = the 1987 number of households in EA no (hi). (3.7)

To specify the generation of a general household sample from the RZMS(EA), one should specify firstly the EA sample sizes in the different sampling strata, denoted

$$a_h \quad (\text{where } a_h \leq a_{h0}),\ h=1,2,\ldots,H, \tag{3.8}$$

and secondly the mode for drawing the household samples and the household sample sizes within EAs,

$m_{hi}$ = the size of the household sample from EA no (hi). (3.9)

In case all the $m_{hi}$ are set equal (= m), we say that the household sample has fixed take m in the EAs. (3.10)

Having made these specifications, one first

selects $a_h$ EAs at random ( = with equal probabilities) from the RZMS list of EAs in the sampling stratum h, h=1,2,...,H. (3.11)

Thereafter one draws samples of households from the household lists for the EAs which were selected in the first stage, with sampling mode and sample sizes as prescribed.

The main mode for selecting the second-stage household samples for the ZNHSCP surveys is as follows.

The household sample from EA no (hi) is drawn as a systematic sample of size $m_{hi}$ from the list of households in EA no (hi). The ordering of the household list may vary from survey to survey. (3.12)

Next we turn to household samples via segments. The generation of a general household sample from the RZMS(Segm) is quite analogous to the RZMS(EA) case. In fact, the two cases are so parallel that we allow ourselves to mostly use the same notation for analogous quantities in the two cases. For the sake of completeness we write down the definitions also for the segment case. Set

$Q_{hi}$ = the 1987 number of households in segment no (hi). (3.13)

To specify the generation of a general household sample from RZMS(Segm), one should firstly specify the segment sample sizes from the different sampling strata,

$$a_h \quad (\text{where } a_h \leq a_{h0}),\ h=1,2,\ldots,H, \tag{3.14}$$

and secondly specify the mode for drawing the household samples and the household sample sizes within the selected segments,

$m_{hi}$ = the size of the household sample from segment no (hi). (3.15)

In case all $m_{hi}$ are set equal (= m), we say that the second-stage sample has fixed take m in each segment. (3.16)

Having made these specifications one first

selects $a_h$ segments at random (= with equal probabilities) from the RZMS(Segm) list of segments in the sampling stratum h, h=1,2,...,H. (3.17)

Thereafter one selects samples of households from the segments which were selected in the first round, with sampling mode and sample sizes as specified. The main mode for selecting the second stage samples for the ZNHSCP surveys is as follows.

The household sample from segment (hi) is drawn as a systematic sample of size $m_{hi}$ from the list of households in segment (hi). The ordering of the household list may vary from survey to survey. (3.18)

Remark 3.2: The sample for the ICDS is a RZMS(EA) sample. However, this is an exception among ZNHSCP surveys in the period from 1987 until the next census in 1992. According to the present planning, most of the other surveys will have RZMS(Segm) samples. ▒

Remark 3.3: It is of course a non-trivial optimization problem to find good choices of the sample sizes $a_h$ and $m_{hi}$. However, when the main concern is estimation, as it is in this report, we do not have to bother about optimal/good sample sizes, we just accept the sample sizes that were used. Therefore, we shall not in this context enter into a discussion of how to find good sample sizes. ▒

## 3.3. Inclusion probabilities.

For computation of the estimation weights in point estimators, one has to know the appropriate first order inclusion probabilities, and our aim here is to specify various inclusion probabilities of relevance for the RZMS samples. First we introduce notation for certain size measures associated with the EAs and the sampling strata;

$S_{hi}$ = the size of EA no (hi) (= the number of households in the EA according to the 1982 census). (3.19)

$S_h$ = the size of stratum h (= the number of households in the stratum according to the 1982 census), i.e. (3.20)

$$S_h = \sum_{i=1}^{H} S_{hi}, \quad h=1,2,...,H. \qquad (3.21)$$

We separate the cases with samples from the RZMS(EA) respectively from the RZMS(Segm), and we start with the former case.

For samples from the RZMS(EA).

The inclusion probability for EA no (hi) is

$$\pi_{hi} = a_h \cdot (S_{hi}/S_h),\ i=1,2,\ldots,N_h,\ h=1,2,\ldots,H. \qquad (3.22)$$

Remark 3.4: The inclusion probability values in (3.22) are, for $a_h = a_{h0}$ (cf (3.8)), those which by Scott was lead to by his "pragmatical" reasoning. As already stated, we accept these values without further discussion, and we refer to (3.22) as Scott's assumption. The reader who wants a fuller discussion is referred to Scott's report.

Under Scott's assumption for $a_h = a_{h0}$, the general formula (3.22), i.e. for general $a_h$, follows from (3.11) and Lemma A2.1 in Appendix 2. The inclusion probabilities in (3.23), (3.28) and (3.29) below follow readily from Scott's assumption together with (3.12) (3.17) and (3.18). The details are left to the reader. ▒

The inclusion probability for a (specified) household in EA no (hi) is

$$\pi_{hij} = a_h \cdot (S_{hi}/S_h) \cdot (m_{hi}/M_{hi}). \qquad (3.23)$$

From (3.23) we see that a household sample via EAs is self-weighting on the household level (i.e. each household has the same inclusion probability) if for some constant f we have,

$$a_h \cdot (S_{hi}/S_h) \cdot (m_{hi}/M_{hi}) = f,\ i=1,2,\ldots,N_h,\ h=1,2,\ldots,H. \qquad (3.24)$$

In (the exceptional) case when (3.24) is satisfied for a household sample with fixed take in the EAs (cf.(3.10)), the constant f in (3.24) has the exact interpretation as,

f = the overall sampling rate of households (= the total household sample size as proportion of the total number of households in the population). (3.25)

In the general case, the constant f in (3.24) lies close to the overall sampling rate.

When expressed as a condition on the household sample sizes, (3.24) takes the following form,

$$m_{hi} = f \cdot M_{hi} \cdot (S_h/S_{hi})/a_h. \qquad (3.26)$$

When the household samples are drawn by (circular) systematic sampling, the self-weighting condition (3.24) is met if the following sampling interval is used in the households list for EA no (hi),

$$l_{hi} = a_h \cdot (S_{hi}/S_h)/f. \qquad (3.27)$$

For samples from the RZMS(Segm).

The inclusion probability for segment no (hi) is

$$\pi_{hi} = a_h \cdot (S_{hi}/S_h)/d_{hi}, \qquad (3.28)$$

where $d_{hi}$ is defined in (3.5).

The inclusion probability for a (specified) household in segment no (hi) is

$$\pi_{hij} = a_h \cdot (S_{hi}/S_h) \cdot (m_{hi}/Q_{hi})/d_{hi}. \qquad (3.29)$$

From (3.29) we see that a household sample is self-weighting on the household level (i.e. each household has the same inclusion probability) if for some constant f we have,

$$a_h \cdot (S_{hi}/S_h) \cdot (m_{hi}/Q_{hi})/d_{hi} = f, \quad i=1,2,\ldots,N_h, \quad h=1,2,\ldots,H. \qquad (3.30)$$

Also here the constant f lies close to the overall sampling rate, i.e. we have,

$$f \approx \text{the overall sampling rate of households (= the household sample size as proportion of the total number of households in the population).} \qquad (3.31)$$

Expressed as a condition on the household sample sizes, (3.30) takes the following form,

$$m_{hi} = f \cdot Q_{hi} \cdot d_{hi} \cdot (S_h/S_{hi})/a_h. \qquad (3.32)$$

When the household samples are drawn by (circular) systematic sampling, the self-weighting condition (3.30) is met by using the following sampling interval in the household list in segment (hi),

$$l_{hi} = (a_h/d_{hi}) \cdot (S_{hi}/S_h)/f. \qquad (3.33)$$

## 3.4. Alternative descriptions of EA-samples and segment samples from the RZMS.

As has already been indicated, the descriptions in Sections 3.2 and 3.3 of household samples from the RZMS are not satisfactory for all purposes, in particular not when it comes to variance estimation. We shall here present alternative descriptions of the probabilistic structure of EA-samples and segment samples from the RZMS. These descriptions are adapted to the variance estimation method which we shall employ later on, namely the so called "ultimate clusters method". More precisely we shall rely on "The UC-procedure in stratified version" as formulated at the end of Appendix 1. The material in the present sub-section will not be needed until we come to estimation of sampling errors in Section 6, and as it is somewhat sophisticated and also interlacing with Appendices 1 and 2, we sug-

gest that reading of this sub-section is postponed until the results are needed in Section 6.

The logical order of our presentation of the ultimate clusters method a bit involved, as a consequence of our strive to refer pronouncedly theoretical stuff to the appendices. Broadly, the logical order goes as follows. In Appendix 2 we introduce various notions and results concerning general probability samples, in particular the concept of "inclusion proportionates". Notions and results from Appendix 2 are then used when formulating the ultimate clusters method in Appendix 1. The contents of these appendices are prerequisites for the following. Our main aims here are two-fold;

(i) To state clearly that we mean that the underlying assumptions for the "UC-procedure in stratified version" in fact are satisfied for general samples from the RZMS.

(ii) To "transform" the previous inclusion probabilities to the corresponding inclusion proportionates and to present specific values for some other parameters in the general UC-procedure.

Here we confine ourselves to the samples of EAs or segments which constitute the first step in the drawing of general household samples from the RZMS. The following presentation as well as the presentations in Appendices 1 and 2 perhaps make the matter look more complicated than it is. As an attempt to ease the flavour of sophistication, we give below a "pedestrian's version" of the core of the condition (iv) in the following Assumptions 3.1 and 3.2.

> If the observed values from a RZMS household sample are grouped so as to bring observations from the same (sub)-division together, then observations in different groups are (or at least can be regarded as) independent of each other. Note that observations within the same group may be correlated, though. (3.34)

We now turn to the more formal considerations an we start with samples via EAs. As stated above, we confine the interest to the EA-sample per se. We use the terminology that an EA-sample is composed by a collection of EA-(sub)samples from the different sampling strata. Let

$R_h$ = the number of (sub)divisions which are represented in the EA-sample from sampling stratum h. (3.35)

The letter r is used to label the (sub)divisions which are represented in the EA-sample (and r runs inside the sample strata). Hence, inside stratum h, r runs over r =1,2,...,$R_h$. Furthermore, let

$b_{hr}$ = the number of sampled EAs from (sub)division r in stratum h. (3.36)

Note the following relation.

$$\sum_{r=1}^{R_h} b_{hr} = a_h \ . \qquad (3.37)$$

In Assumption 3.1 below, we collect various aspects on the probabilistic structure of an EA-sample. The claims are motivated afterwards.

ASSUMPTION 3.1 Probabilistic structure of a general EA-sample from the RZMS(EA).

(i) The EA-samples from different sampling strata are independent of each other.

(ii) From each sampling stratum, the EA-sample is drawn without replacement and with prescribed sample size. The sample size in sampling stratum h is $a_h$.

(ii) The inclusion proportionates (see Appendix 2) for the EA-sample from stratum h are

$$\beta_{hi} = S_{hi}/S_h, \ i=1,2,\ldots,N_h. \qquad (3.38)$$

(iv) Let $G_{h1}$, $G_{h2}$, ..., $G_{hR_h}$ denote the (sub)divisions which are represented in the EA-sample from stratum h. Then the contents of the following EA-groups

$$\{i;i\epsilon G_{h1}\}, \{i;i\epsilon G_{h2}\}, \ \ldots \ , \{i;i\epsilon G_{hR_h}\} \qquad (3.39)$$

can with good approximation be regarded as being independent of each other.

Comments on the justification of the assumptions (i)-(iv): The assumptions (i) and (ii) are only repetitions of facts which were stated in Section 3.2. The claim in (iii) is essentially only a reformulation of Scott's assumption (3.22). Links between (3.22) and (iii) are provided by Lemma A2.1 and formula (A2.6) in Appendix 2. The most direct way to make the assumption (iv) plausible is to refer to Remark A1.4 in Appendix 1. We let that be justification enough, but we comment on it below.

The description of sample structure which is given in the above Assumption is sort of "backwards" compared with how stage-wise samples usually are specified. The "normal" route is to specify the draw characteristics for the successive stages. In the above description we view the EA-sample as a two-stage sample with (sub)divisions as PSUs and EAs as SSUs. However, we are specific only on inclusion characteristics after the second stage, while we are vague on draw characteristics relating to the first stage.The reason for this type of "backwards"

description is that we want to keep Scott's assumption (3.22) as the central assumption on inclusion characteristics. Given this, it would only complicate matters to strive for being more specific on draw characteristics for the first stage and we would also run into the problem, which is discussed at length by Scott, concerning the adequate documentation of the selection of the original ZMS. However, although the above description is partly vague, it will suffice for the purpose of deriving estimates for sampling errors.

We conclude the justification by stressing that (i)-(iv) is a set of assumptions. We claim that it reasonable to believe that the assumptions are met in the practical situations. However, they may and can be questioned on very much the same grounds as those which Scott use in his discussion of the inclusion probabilities. However, on essentially the same grounds as for the inclusion probabilities we regard the assumptions to be "pragmatically reasonable". ▒

From the one-to-one correspondence between EAs and segments in the RZMS(EA) and the RZMS(Segm), it should be clear that the probabilistic structure of segment samples from the RZMS(Segm) can be described in a very similar way. For the sake of completeness we write down the analog for segment samples in Assumption 3.2 below. This Assumption is of course as much of a pragmatically reasonable assumption as the previous one. Furthermore, justifications can be given along the same lines as above, and we leave the details to the reader.

ASSUMPTION 3.2 Probabilistic structure of a general segment sample from the RZMS(Segm):

(i) The segment samples from different sampling strata are independent of each other.

(ii) From each sampling stratum, the segment sample is drawn without replacement and with prescribed sample size. The sample size in sampling stratum h is $a_h$.

(ii) The inclusion proportionates (see Appendix 2) for the segment sample from sampling stratum h are

$$\beta_{hi} = (S_{hi}/S_h)/d_{hi},\ i=1,2,\ldots,N_h. \qquad (3.40)$$

(iv) Let $G_{h1}$, $G_{h2}$, ..., $G_{hR_h}$ denote the (sub)divisions which are represented in the sample of segments from sampling stratum h. Then the contents of the following "segment groups"

$$\{i;i\epsilon G_{h1}\},\ \{i;i\epsilon G_{h2}\},\ \ldots,\ \{i;i\epsilon G_{hR_h}\} \qquad (3.41)$$

can with good approximation be regarded as independent of each other.

Remark 3.5: Here we make a final comment, which concerns the empirical values of the sizes of the (sub)division groups, i.e. the b:s in (3.36). According to Tin's manual, CSO (December 1986), in the selections for the ZMS, two segments were sampled from the selected (sub)divisions in most of the sampling strata. However, in the sampling strata in the sector "urban & semi-urban areas" three segments were drawn from the sampled (sub)divisions. Therefore, ideally the b-values for the RZMS should be either 2 or 3. This is not the case, though; in fact b-values = 1,2 and 3 are represented in the RZMS. The reason for this must again be the fact which has been mentioned before; Various undocumented alterations took place in the ZMS in its initial years. Fortunately though, for all EAs in the RZMS the original (sub)division is recorded. As a consequence of this it is possible to determine which EAs/segments in the RZMS that come from the same (sub)division. ▒

## 4. ORGANIZATION OF THE OBSERVATIONS FROM A SURVEY WITH A RZMS SAMPLE.

The execution of a RZMS survey starts with the drawing of a household sample along the lines presented in the previous section. The next step is to collect the desired observations from the sampled households.

In the data collection process, non-responses may occur (and only in exceptional cases do they not occur). Non-response always causes problem, which can be handled in different ways depending on the topic of the survey. Possibilities which are considered and used for the ZNHSCP surveys are;

- To accept non-responses and adjust for them (as well as one can) in the estimation phase.
- To use substitute respondents at non-response.
- To impute for non-responses in some appropriate way.

In the subsequent discussion of estimation procedures we shall disregard non-response for the following reason. When nonresponse is handled along some of the lines mentioned above, one can usually bring the estimation procedures back on the "ideal" situation (i.e. without non-response). In the substitution and imputation cases, this is done so as to say straight away. In the case when non-responses are accepted, one usually handles the non-response problem along the following lines. The quantity $m_{hi}$, which was defined as the number of sampled households in (3.9) and (3.15) is instead interpreted as the number of responding households and after this modification, estimation formulas for the ideal situation are applied. This procedure is justified at least if non-responses can be regarded to occur "at random" (relative to the values of the variable under interest). Note, though, that if non-responses are accepted then a self-weighting property for the original sample may be violated which, however, is not a really serious obstacle.

Henceforth we assume that the households are to be interviewed about values of household variables as well as individual-variables for members of the household. Furthermore, we assume that the survey interest concerns groups of households as well as groups of individuals.

For the registration of the collected data (at least) two files are set up, one households file and one individuals file. We start by discussing the former.

The households file contains one record for each observed household, every record containing information on

- identification of the household,
- sampling design parameters,
- observed values of (household) variables.

The contents of the three parts of the records are discussed below.

Identification of the household. This part of the record gives unambiguous information on the "label" (hrij) of the household, i.e. (see Section 3) on the sampling stratum h which the EA/-segment belongs to, the number i of the EA/segment within the sampling stratum, the number j of the household within the EA/segment and also the label r of the (sub)division that the EA/segment belongs to.

Sampling design parameters. The sampling procedure is assumed to be as specified in Section 3. We separate the cases with household samples via EAs respectively via segments.

For samples from RZMS(EA).

The relevant sampling design parameters are;

$S_h$ = the 1982 number of households in the sampling stratum,
$a_h$ = the EA sample size in the sampling stratum,
$R_h$ = the number of (sub)divisions which are represented in the sample from the sampling stratum,
$b_{hr}$ = the number of sampled EAs from the (sub)division,
$S_{hi}$ = the 1982 number of households in the EA,
$M_{hi}$ = the 1987 number of households in the EA,
$m_{hi}$ = the size of the household sample from the EA.

For samples from RZMS(Segm).

The relevant sampling design parameters are;

$S_h$ = the 1982 number of households in the sampling stratum,
$a_h$ = the segment sample size in the sampling stratum,
$R_h$ = the number of (sub)divisions which are represented in the sample from the sampling stratum,
$b_{hr}$ = the number of sampled segments from the (sub)-division,
$S_{hi}$ = the 1982 number of households in the EA to which the segment belongs,
$Q_{hi}$ = the 1987 number of households in the segment,
$d_{hi}$ = the number of original segments in the EA, which the segment belongs to,
$m_{hi}$ = the size of the household sample from the segment.

The needs for information on the above sampling design parameters will become clear when we come to point estimation and estimation of sampling errors in Sections 5 and 6.

Remark 4.1: In the files listed below, we have included information on all the sampling design parameters specified above. In practical situations, though, one often confines to some "condensed" version the parameters (i.e. some function of them), which gives enough information for the specific estima-

tion purposes; Inclusion of only "estimation weights" is a typical example. However, for the sake of generality we continue to include all sampling design parameters in the data files. ▒

Observed variable values. Should need no further explanation.

The individuals file is organized quite analogously; the minor differences being that here identification concerns individuals, i.e. the identification part should specify (hrijk), and variables stand for individual-variables.

In Tables 4.1 - 4.4 we present the information content, and its structure, in the households and individuals files. Again we separate the cases with household samples via EAs respectively via segments.

For samples from the RZMS(EA).

Tables 4.1 and 4.2 indicate how the households file respectively the individuals file should be organized.

| Identification | | | | Sampling design parameters | | | | | | | Variable values | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | Hh | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $M_{hi}$ | $m_{hi}$ | $\underline{x}$ | $\underline{y}$ | $\underline{z}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| h | r | i | j | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $M_{hi}$ | $m_{hi}$ | $x_{hij}$ | $y_{hij}$ | $z_{hij}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Table 4.1. The households file with a general record.

| Identification | | | | | Sampling design parameters | | | | | | | Variable values | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | Hh | Ind | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $M_{hi}$ | $m_{hi}$ | $\underline{x}$ | $\underline{y}$ | $\underline{z}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| h | r | i | j | k | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $M_{hi}$ | $m_{hi}$ | $x_{hijk}$ | $y_{hijk}$ | $z_{hijk}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Table 4.2. The individuals file with a general record.

Note that $\underline{x},\underline{y},\underline{z},\ldots$ in Table 4.1 denote household variables while they in Table 4.2 denote individual-variables. The sampling design parameters are the same in both tables, though.

For samples from the RZMS(Segm).

Tables 4.3 and 4.4 indicate how the households file respectively the individuals file should be organized.

| Identification | | | | Sampling design parameters | | | | | | | | Variable values | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | Hh | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $Q_{hi}$ | $d_{hi}$ | $m_{hi}$ | $\underline{x}$ | $\underline{y}$ | $\underline{z}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| h | r | i | j | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $Q_{hi}$ | $d_{hi}$ | $m_{hi}$ | $x_{hij}$ | $y_{hij}$ | $z_{hij}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Table 4.3. The households file with a general record.

| Identification | | | | | Sampling design parameters | | | | | | | | Variable values | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | Hh | Ind | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $Q_{hi}$ | $d_{hi}$ | $m_{hi}$ | $\underline{x}$ | $\underline{y}$ | $\underline{z}$ |
| . | . | . | . | . | . | . | . | . | . | . | | . | . | . | . |
| h | r | i | j | k | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $Q_{hi}$ | $d_{hi}$ | $m_{hi}$ | $x_{hijk}$ | $y_{hijk}$ | $z_{hijk}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Table 4.4. The individuals file with a general record.

# 5. POINT ESTIMATION ON THE BASIS OF OBSERVATIONS ON HOUSEHOLD SAMPLES FROM THE RZMS.

## 5.1. Some generalities on estimation on the basis of probability samples.

We start with some general words on estimation on the basis of observations on a general probability sample. For simplicity we assume that the objects in the population are labeled as in the household population U in (2.4). We confine ourselves to sampling procedures without replacement. Then, in the course of the sampling process an object in the population is either sampled or it is not, and we describe this duality by the sample inclusion indicators,

$$I_{hij} = \begin{cases} 1 & \text{if object (hij) is included in the sample,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

The inclusion probability for household (hij) is defined as (P denotes probability)

$$\pi_{hij} = P(I_{hij}=1). \tag{5.2}$$

Let D denote a domain in U. The Horvitz-Thompson estimator for the domain total $\theta(\underline{x};D)$ is

$$\hat{\theta}(\underline{x};D) = \sum_{(hij)\epsilon U} I_{hij} \cdot \underline{1}_D(hij) \cdot (x_{hij}/\pi_{hij}). \tag{5.3}$$

As is well known, the estimator in (5.3) yields unbiased estimation of the true domain total $\theta(\underline{x};D)$.

By (2.15), estimation of a domain size can be regarded as a special case of estimation of a domain total. This observation together with (5.3) leads to the following unbiased domain size estimator,

$$\hat{g}(D) = \sum_{(hij)\epsilon U} I_{hij} \cdot \underline{1}_D(hij)/\pi_{hij}. \tag{5.4}$$

By virtue of (2.16), to estimate the mean of the variable $\underline{x}$ over the domain D we employ the following "natural" (and usual) estimator

$$\hat{\mu}(\underline{x};D) = \hat{\theta}(\underline{x};D)/\hat{g}(D) = \hat{\theta}(\underline{x};D)/\hat{\theta}(\underline{1};D). \tag{5.5}$$

The estimator in (5.5) is not exactly, but approximately unbiased, with good approximation at least if the sample size is fairly large.

The above estimation formulas are general and they hold for any probability sampling procedure without replacement. In the sequel we shall restrict to the type of sampling procedures which were considered in Section 3. As before, we let h,i,j and k label sampling strata, EAs/segments, households and individuals respectively. From now on we let the indices

run only over the sample and not, as in Section 2, over the entire population. Hence i and j run over i = 1,2,... ,$a_h$ and j = 1,2,...,$M_{hi}$. We let H, which h runs over, denote the number of effective strata, i.e. strata with positive EA sample size. So far we have not mentioned the division/subdivision label r, and that label will in fact be "superfluous" in this section on point estimation, but it will be needed when we come to estimation of sampling errors in Section 6.

In the following we present point estimation formulas in "algebraic versions". In Section 6, where the main theme is estimation of sampling errors, the approach will be more "computation oriented", and in that section we also shed some further light on computational aspects of point estimates.

## 5.2. Estimation for groups of households.

We continue to use terminology and notation which has been introduced so far. In particular, U is the population of households, D a group of households (in U) and $\underline{x} = \{x_{hij}\}$ a household variable. We separate the cases with household samples via EAs and via segments.

### Samples from the RZMS(EA).

The sample is assumed to be a general household sample via EAs as described in Section 3, and we continue to use the notation for sampling parameters introduced there.

The general estimation formulas (5.3)-(5.5), the partitioning formula (2.24) and the formula (3.23) for inclusion probabilities lead to the following estimators for group totals, group sizes, group means and group proportions. (Note that the previously introduced convention to let indices run over the sample only has the effect that the inclusion indicators in (5.3) and (5.4) are absorbed into the summations.)

$$\hat{\theta}(\underline{x};D) = \sum_{h=1}^{H}(1/a_h) \sum_{i=1}^{a_h}(S_h/S_{hi})\cdot(M_{hi}/m_{hi})\cdot\sum_{j=1}^{m_{hi}} x_{hij}\cdot 1_D(hij), \quad (5.6)$$

$$\hat{g}(D) = \sum_{h=1}^{H}(1/a_h) \sum_{i=1}^{a_h}(S_h/S_{hi})\cdot(M_{hi}/m_{hi}) \sum_{j=1}^{m_{hi}} 1_D(hij), \quad (5.7)$$

$$\hat{\mu}(\underline{x};D) = \hat{\theta}(\underline{x};D)/\hat{g}(D), \quad (5.8)$$

$$\hat{p}(A;D) = \hat{g}(A\cap D)/\hat{g}(D). \quad (5.9)$$

In the particular case with a self-weighting household sample, i.e. when the relation (3.24) is satisfied, the formulas (5.6)-(5.9) simplify as follows;

$$\hat{\theta}(\underline{x}:D) = (1/f)\cdot\sum_{h=1}^{H} \sum_{i=1}^{a_h} \sum_{j=1}^{m_{hi}} x_{hij}\cdot 1_D(hij), \quad (5.10)$$

$$\hat{g}(D) = (1/f)\cdot\sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}} 1_D(hij). \tag{5.11}$$

$$\hat{\mu}(\underline{x}:D) = \sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}} x_{hij}\cdot 1_D(hij) \Big/ \sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}} 1_D(hij). \tag{5.12}$$

$$\hat{p}(A;D) = \sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}} 1_{A\cap D}(hij) \Big/ \sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}} 1_D(hij). \tag{5.13}$$

Remark 5.1: The formulas (5.10)-(5.13) can be verbalised as follows. Under the self-weighting condition (3.24) we have;

- Group means and proportions in the household population are estimated by the corresponding group means and proportions in the sample.

- Group totals and group sizes are estimated by multiplying the corresponding totals and sizes in the sample by the inverse sampling rate 1/f (cf.(3.25)). ▒

Samples from the RZMS(Segm).

Here we assume that the sample is a general household sample via segments as described in Section 3, and we continue to use the notation for sampling parameters introduced there.

The general estimation formulas (5.3)-(5.5), the partitioning formula (2.24) and the formula (3.29) for inclusion probabilities lead to the following estimators for group totals, group sizes, group means and group proportions.

$$\hat{\theta}(\underline{x};D) = \sum_{h=1}^{H}(1/a_h)\sum_{i=1}^{a_h}(S_h/S_{hi})\cdot(d_{hi}\cdot Q_{hi}/m_{hi})\cdot\sum_{j=1}^{m_{hi}} x_{hij}\cdot 1_D(hij), \tag{5.14}$$

$$\hat{g}(D) = \sum_{h=1}^{H}(1/a_h)\sum_{i=1}^{a_h}(S_h/S_{hi})\cdot(d_{hi}\cdot Q_{hi}/m_{hi})\sum_{j=1}^{m_{hi}} 1_D(hij), \tag{5.15}$$

$$\hat{\mu}(\underline{x};D) = \hat{\theta}(\underline{x};D)/\hat{g}(D), \tag{5.16}$$

$$\hat{p}(A;D) = \hat{g}(A\cap D)/\hat{g}(D). \tag{5.17}$$

In the particular case with a self-weighting household sample, i.e. when (3.30) is satisfied, the formulas (5.14)-(5.17) simplify to those in (5.10)-(5.13). Hence, the claims in Remark 5.1 hold also in this case.

5.3 Estimation for groups of individuals.

In this sub-section we consider the population V of all individuals (see (2.20)), D denotes a group of individuals (in V) and $\underline{x} = \{x_{hijk}\}$ an individual-variable. As in the previous sub-section we let the indices h,i,j and k run over the sample only. We assume that observations are made on all individuals in a selected household. Hence, the index k runs over k = 1,2, ..., $K_{hij}$ (cf.(2.19)). As in the households case, the index r will be irrelevant in this point estimation context.

The estimation formulas to be considered can be regarded as special cases of Horvitz-Thompson estimators corresponding to the probability samples of individuals. However, it is probably easier to view the formulas as special cases of the estimation formulas for households, with household variable defined by

$$x_{hij}(D) = \sum_{k=1}^{K_{hij}} x_{hijk} \cdot \underline{1}_D(hijk), \qquad (5.18)$$

i.e. the households variable is obtained by summing the values of the individual-variables over the members in the household who also belong to the group D under consideration. We leave the details of the "identification" between the households and individuals cases to the reader.

Again we separate the cases with household samples via EAs and via segments.

Samples from the RZMS(EA).

The estimation formulas for estimation of group totals, group sizes, group means and group proportions here take the following forms.

$$\hat{\theta}(\underline{x}:D) =$$

$$= \sum_{h=1}^{H} (1/a_h) \sum_{i=1}^{a_h} (S_h/S_{hi}) \cdot (M_{hi}/m_{hi}) \cdot \sum_{j=1}^{m_{hi}} \sum_{k=1}^{K_{hij}} x_{hijk} \cdot \underline{1}_D(hijk), \qquad (5.19)$$

$$\hat{g}(D) = \sum_{h=1}^{H} (1/a_h) \sum_{i=1}^{a_h} (S_h/S_{hi}) \cdot (M_{hi}/m_{hi}) \cdot \sum_{j=1}^{m_{hi}} \sum_{k=1}^{K_{hij}} \underline{1}_D(hijk), \qquad (5.20)$$

$$\hat{\mu}(\underline{x};D) = \hat{\theta}(\underline{x};D)/\hat{g}(D), \qquad (5.21)$$

$$\hat{p}(A;D) = \hat{g}(A \cap D)/\hat{g}(D). \qquad (5.22)$$

Under the assumption that the household sample is self-weighting, i.e. that (3.24) is satisfied, the formulas (5.19)-(5.22) simplify as follows.

$$\hat{\theta}(\underline{x};D) = (1/f)\cdot\sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}}\sum_{k=1}^{K_{hij}} x_{hijk}\cdot 1_D(hijk), \tag{5.23}$$

$$\hat{g}(D) = (1/f)\cdot\sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}}\sum_{k=1}^{K_{hij}} 1_D(hijk), \tag{5.24}$$

$$\hat{\mu}(\underline{x};D) =$$

$$= \sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}}\sum_{k=1}^{K_{hij}} x_{hijk}\cdot 1_D(hijk) \Big/ \sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}}\sum_{k=1}^{K_{hij}} 1_D(hijk), \tag{5.25}$$

$$\hat{p}(A;D) = \sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}}\sum_{k=1}^{K_{hij}} 1_{A\cap D}(hijk) \Big/ \sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{m_{hi}}\sum_{k=1}^{K_{hij}} 1_D(hijk). \tag{5.26}$$

Remark 5.2: From (5.23)-(5.26) we see that the following counterpart of Remark 5.1 holds true. Under the self-weighting condition (3.24) we have;

- Group means and proportions in the individuals population are estimated by the corresponding means and proportions in the sample.
- Group totals and group sizes in the individuals population are estimated by multiplying the corresponding totals and sizes in the sample by the inverse sampling rate 1/f (cf (3.26)). ▒

Samples from the RZMS(Segm).

By applying the formula (3.29) for inclusion probabilities instead of (3.23) we get the following estimation formulas.

$$\hat{\theta}(\underline{x}:D) =$$

$$= \sum_{h=1}^{H}(1/a_h)\sum_{i=1}^{a_h}(S_h/S_{hi})\cdot(d_{hi}\cdot Q_{hi}/m_{hi})\cdot\sum_{j=1}^{m_{hi}}\sum_{k=1}^{K_{hij}} x_{hijk}\cdot 1_D(hijk), \tag{5.27}$$

$$\hat{g}(D) = \sum_{h=1}^{H}(1/a_h)\sum_{i=1}^{a_h}(S_h/S_{hi})\cdot(d_{hi}\cdot Q_{hi}/m_{hi})\cdot\sum_{j=1}^{m_{hi}}\sum_{k=1}^{K_{hij}} 1_D(hijk), \tag{5.28}$$

$$\hat{\mu}(\underline{x};D) = \hat{\theta}(\underline{x};D)/\hat{g}(D), \tag{5.29}$$

$$\hat{p}(A;D) = \hat{g}(A\cap D)/\hat{g}(D). \tag{5.30}$$

As is readily realised, under the self-weighting condition (3.30) the above formulas simplify to those in (5.23)-(5.26). Hence, the claims in Remark 5.2 apply also in the segment case.

## 5.4. An alternative way to estimate the number of households and the number of individuals in an area.

Scott suggests (see his report p.23) that when evaluating the ICDS, more accurate estimates of group sizes than those given by (5.7) and (5.20) could be obtained, at least for certain types of groups, by using ratio type estimators. We believe this is true, although it is difficult (at least at present) to be precise about the variance reduction that could be achieved. Below we introduce the ratio type estimators which are of interest. We presume that a household sample via EAs has been selected as specified in Section 3, and that observations on certain EA-variables (to be specified) have been gathered.

The estimation interest concerns estimation of the number of households and the number of individuals in an area B of the following type.

B is a geographical area which is a union of (1982) census EAs. (5.31)

As usual we separate the households and the individuals cases, and we start with the former.

### 5.4.1 Estimation of the number of households in an area.

Set

$g(B;87)$ = the 1987 number of households in the area B, (5.32)

$g(B;c)$ = the (1982) census number of households in the area B. (5.33)

Define $\tau(B)$ by the relation

$$g(B;87) = g(B;c)\cdot\tau(B), \tag{5.34}$$

i.e.

$$\tau(B) = g(B;87)/g(B;c). \tag{5.35}$$

In virtue of (5.34), if $g(B;c)$ is known one can estimate $g(B;87)$ by first estimating $\tau(B)$. We regard the census values for the number of households in (all) EAs as known. In particular, under the assumption (5.31) $g(B;c)$ is also known.

When (5.31) holds, we have the following estimate of $g(B;87)$ where, as before, $M_{hi}$ denotes the (1987) number of households in EA no (hi) (cf.(2.3)),

$$\hat{g}(B;87) = \sum_{h=1}^{H}(1/a_h)\cdot\sum_{i=1}^{a_h}(S_h/S_{hi})\cdot M_{hi}\cdot\delta((hi);B), \tag{5.36}$$

where $\delta$ is the following indicator for area B inclusion,

$$\delta((hi);B) = \begin{cases} 1 & \text{if EA no (hi) belongs to the area B,} \\ 0 & \text{otherwise.} \end{cases} \qquad (5.37)$$

Note that under (5.31), an EA either lies entirely inside or entirely outside the area B.

We can estimate g(B;c) analogously (even if g(B;c) in fact is assumed to be known) by the right hand side of (5.36) with $M_{hi}$ changed to $S_{hi}$, i.e. by the estimator,

$$\hat{g}(B;c) = \sum_{h=1}^{H} (1/a_h) \cdot \sum_{i=1}^{a_h} (S_h/S_{hi}) \cdot S_{hi} \cdot \delta((hi);B). \qquad (5.38)$$

Now (5.36) and (5.38) enable the following estimation of $\tau(B)$,

$$\hat{\tau}(B) = \hat{g}(B;87)/\hat{g}(B;c). \qquad (5.39)$$

By combining (5.39) and (5.34) we are led to the following ratio estimator for the 1987 number of households in the area B,

$$\widehat{g(B;87)}_R = g(B;c) \cdot \hat{\tau}(B), \qquad (5.40)$$

where $\hat{\tau}(B)$ is defined by (5.39), (5.36) and (5.38).

5.4.2 Estimation of the number of individuals in an area.

We stick to the assumptions about a household sample via EAs and an area B which satisfies (5.31). However, we shift the interest from estimation of the number of households in the area B to estimation of the number of individuals in it. We use the same notation as above also in this individuals case, but change the definitions in the following "natural" way,

$$g(B;87) = \text{the 1987 number of individuals in the area B,} \qquad (5.41)$$

$$g(B;c) = \text{the (1982) census number of individuals in the area B.} \qquad (5.42)$$

We regard the census values for the number of individuals in (all) EAs to be known. In particular, under (5.31) g(B;c) is known. Furthermore we assume that the 1987 number of individuals in the selected EAs are known. Also here we define $\tau(B)$ by (5.35). Let

$$T_{hi}(87) = \text{the 1987 number of individuals in EA no (hi).} \qquad (5.43)$$

The following relation holds,

$$T_{hi} = \sum_{j=1}^{M_{hi}} K_{hij}, \qquad (5.44)$$

where $K_{hij}$ is the number of individuals in household no (hij) (cf.(2.19)). The corresponding number for the census is denoted

$$T_{hi}(c) = \text{the census number of individuals in EA no (hi).} \tag{5.45}$$

Under (5.31), g(B;87) and g(B;c) can be estimated as follows where $\delta$ is as in (5.37),

$$\hat{g}(B;87) = \sum_{h=1}^{H}(1/a_h)\cdot\sum_{i=1}^{a_h}(S_h/S_{hi})\cdot T_{hi}(87)\cdot\delta((hi);B), \tag{5.46}$$

$$\hat{g}(B;c) = \sum_{h=1}^{H}(1/a_h)\cdot\sum_{i=1}^{a_h}(S_h/S_{hi})\cdot T_{hi}(c)\cdot\delta((hi);B). \tag{5.47}$$

Hence, we are led to the following ratio estimator for the 1987 number of individuals in the area B,

$$\widehat{g(B;87)}_R = g(B;c)\cdot\hat{\tau}(B), \tag{5.48}$$

where $\hat{\tau}(B)$ now is defined by (5.39), (5.46) and (5.47).

Remark 5.3: In order to apply the above estimation methods, one must know the values of $M_{hi}$ and $T_{hi}$ for the EAs in the sample. In the ICDS, Round 0, these values were collected, and hence the method can be applied to the ICDS data. ▒

## 6. ESTIMATION OF SAMPLING ERRORS AND COMPUTATION ALGORITHMS.

In this section we shall exhibit procedures for estimation of the sampling errors for the estimators which were considered in Section 5. Throughout, our method for construction of sampling error estimates will be the so called ultimate clusters method, which is discussed in Appendix 1.

We shall make the usual division into cases with household samples via EAs (treated in Sub-section 6.1) respectively via segments (treated in Sub-section 6.2). Within each of the two cases we subdivide after the following targets for the estimator; "totals for groups of households", "means for groups of households", "totals for groups of individuals" and "means for groups of individuals". However, many of the situations are very analogous, and therefore we give detailed treatments of only some of them. In Sub-section 6.3 we consider estimation of the sampling errors for the ratio type estimators which were introduced in Section 5.4.

### 6.1 Estimation of the sampling errors for estimates based on household samples from the RZMS(EA).

The sample is assumed to be a general household sample via EAs as described in Section 3, and we continue to use the notation for sampling parameters introduced there.

The main instrument for derivation of sampling error estimates will be the "UC-procedure in stratified version", which is formulated at the end of Appendix 1. Therefore, acquaintance with Appendix 1 is a prerequisite for the following considerations. So is also acquaintance with the material in Section 3.4. By virtue of Assumption 3.1 (see Section 3.4) we can apply the "UC-procedure in stratified version" (see Section A1.2). In view of (iii) and (iv) in Assumption 3.1 we apply the UC-procedure with the following particular specifications,

$$\beta_{hi} = S_{hi}/S_h, \quad i=1,2,\ldots,N_h, \quad h=1,2,\ldots,H. \tag{6.1}$$

The groups $G_{h1}, G_{h2}, \ldots, G_{hR_h}$ are those generated by the (sub)divisions. The corresponding group sizes are denoted $b_{hr}$ (cf. (3.36)). (6.2)

#### 6.1.1 For estimates of totals for groups of households.

We assume that a household variable $\underline{x}$ and a group D of households are specified, and fixed. Our concern will be estimation of the sampling error of the estimate (5.6) of the group total $\Theta(\underline{x};D)$. As already stated, our main tool will be the UC-procedure. Note that in our formulation this procedure leads to a point estimate, given by (A1.30), as well as to an estimate of the variance estimator, given by (A1.31). We shall soon see that the UC-procedure point estimate in fact coincides with that in (5.6). To give a comprehensive idea of the compu-

tations needed, and also to provide background for EDP-programming we shall give a fairly detailed presentation of the organization of the successive steps in computations. We believe that a lucid presentation is obtained in terms of successive "reductions" of the households file and we shall use that terminological approach.

As point of departure we take the households file (see Section 4), "reduced" to contain only the variables $\underline{x}$ and $\underline{q}$ as illustrated in Table 6.1. The variable $\underline{q}$ (which might be vector valued) is "group-D-specifying", i.e. it contains information which enables determination of whether a household belongs the group D or not.

| Identification | | | | Sampling design parameters | | | | | | | Variables | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | Hh | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $M_{hi}$ | $m_{hi}$ | $\underline{x}$ | $\underline{q}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . |
| h | r | i | j | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $M_{hi}$ | $m_{hi}$ | $x_{hij}$ | $q_{hij}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . |

Table 6.1. The households file (reduced to $\underline{x}$ and $\underline{q}$).

The variables $y_{hi}$ and $Y_{hi}$ in (A1.26) and (A1.29) are chosen according to (6.3) and (6.4) below. In accordance with previous conventions, the index j runs over the population in (6.3) while it runs over the sample in (6.4).

$$y_{hi} = \sum_{j=1}^{M_{hi}} x_{hij} \cdot \underline{1}_D(hij), \tag{6.3}$$

$$Y_{hi} = (M_{hi}/m_{hi}) \cdot \sum_{j=1}^{m_{hi}} x_{hij} \cdot \underline{1}_D(hij). \tag{6.4}$$

It is readily seen that the unbiasedness condition in (A1.29) is fulfilled with y and Y as above.

The next reduction step consists in computation of the quantities in (6.5), which in terms of the UC-procedure mean computation of $Y_{hi}/\beta_{hi}$,

$$\hat{\theta}_{hi}(\underline{x};D) = (S_h/S_{hi}) \cdot (M_{hi}/m_{hi}) \cdot \sum_{j=1}^{m_{hi}} x_{hij} \cdot \underline{1}_D(hij). \tag{6.5}$$

By virtue of Assumption 3.1 and Lemma A2.4 the following holds.

> The variables $\hat{\theta}_{hi}(\underline{x};D)$, $i = 1,2,\ldots,N_h$, can with good approximation be viewed as independent EA-wise estimates of the group D $\underline{x}$-total for the sampling stratum h. (6.6)

The collection of $\hat{\theta}$-estimates is illustrated by the file in Table 6.2, which means a reduction of the file in Table 6.1, to the effect that households have been reduced away.

| Identification | | | Sampling design parameters | | | Variable |
|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | $a_h$ | $R_h$ | $b_{hr}$ | $\hat{\theta}(\underline{x};D)_{hi}$ |
| . | . | . | . | . | . | . |
| h | r | i | $a_h$ | $R_h$ | $b_{hr}$ | $\hat{\theta}_{hi}(\underline{x};D)$ as in (6.5) |
| . | . | . | . | . | . | . |

Table 6.2. File with EA-wise estimates of group D $\underline{x}$-totals for the sampling strata.

At this junction the reductions leading to the point estimate respectively to the variance estimate take different routes. First we follow the route to the point estimate in (A1.30). The file in Table 6.2 is then reduced by computation of estimates of the group D $\underline{x}$-totals for the sampling strata, by averaging the $\theta$-estimators in Table 6.2 within sampling strata,

$$\hat{\theta}_h(\underline{x};D) = (1/a_h) \cdot \sum_{i=1}^{a_h} \hat{\theta}_{hi}(\underline{x};D), \quad h=1,2,\ldots,H. \tag{6.7}$$

By (6.6), $\hat{\theta}_h(\underline{x};D)$ is an unbiased estimator of $\theta_h(\underline{x};D)$ (see (2.11)).

The resulting estimates are shown in the file in Table 6.3, which is a reduced version of the file in Table 6.2.

| Identification | Variable |
|---|---|
| Stratum | Estimates of group D $\underline{x}$-total for the strata |
| 1 | $\hat{\theta}_1(\underline{x};D)$ |
| 2 | $\hat{\theta}_2(\underline{x};D)$ |
| . | . |
| h | $\hat{\theta}_h(\underline{x};D)$ as in (6.7) |
| . | . |
| H | $\hat{\theta}_H(\underline{x};D)$ |

Table 6.3. File with stratum-wise estimates of group D $\underline{x}$-totals for the sampling strata.

Finally along this route, the estimate (A1.30) of the group total $\theta(\underline{x};D)$ is obtained by summing the estimates in Table 6.3,

$$\hat{\theta}(\underline{x};D) = \sum_{h=1}^{H} \hat{\theta}_h(\underline{x};D). \qquad (6.8)$$

It is readily checked that the estimate in (6.8) in fact is identical with that in (5.6). We leave the details to the reader.

We now return to the file in Table 6.2 to follow the route which leads to the variance estimate in (A1.31). Then, the next reduction step is (cf. (A1.33)) to average the estimates in Table 6.2 within (sub)divisions,

$$\hat{\theta}_{h(r)}(\underline{x};D) = (1/b_{hr}) \cdot \sum_{i \epsilon G_{hr}} \hat{\theta}_{hi}(\underline{x};D). \qquad (6.9)$$

The quantities in (6.9) can be viewed as (sub)division-wise estimates of group D $\underline{x}$-totals for the sampling strata. They are collected in the file in Table 6.4, which is a reduction of the file in Table 6.2 to the effect that EAs are averaged out.

| Identification | | Variable |
|---|---|---|
| Stratum | Subdiv | (Sub)division-wise estimates of group D $\underline{x}$-totals in the strata |
| . | . | . |
| h | 1 | $\hat{\theta}_{h(r)}(\underline{x};D)$ |
| h | . | . |
| h | r | $\hat{\theta}_{h(r)}(\underline{x};D)$ as in (6.9) |
| h | . | . |
| h | $R_h$ | $\hat{\theta}_{hR_h}(\underline{x};D)$ |
| . | . | . |

Table 6.4. File with (sub)division-wise estimates of group D $\underline{x}$-totals for the sampling strata.

Next, in accordance with (A1.32), compute the sample variances within the sampling strata for the $\theta$-estimates in Table 6.4, i.e. compute for h=1,2,...,H,

$$S^2_h(\hat{\underline{\theta}}) = (1/(R_h-1)) \cdot \sum_{r=1}^{R_h} [\hat{\theta}_{h(r)}(\underline{x};D) - \hat{\theta}^*_h(\underline{x};D)]^2, \qquad (6.10)$$

where (cf. (A1.34))

$$\hat{\theta}^*_h(\underline{x};D) = (1/R_h) \cdot \sum_{r=1}^{R_h} \hat{\theta}_{h(r)}(\underline{x};D). \qquad (6.11)$$

| Identification | Variable |
|---|---|
| Stratum | Sample variance |
| 1 | $S^2_1(\hat{\underline{\theta}})$ |
| . | . |
| h | $S^2_h(\hat{\underline{\theta}})$ as in (6.10) |
| . | . |
| H | $S^2_H(\hat{\underline{\theta}})$ |

Table 6.5. File with sample variances within sampling strata.

In accordance with (A1.31), the final step in the computation of the estimate of the estimator variance is given by

$$\hat{V}[\hat{\theta}(\underline{x};D)] = \sum_{h=1}^{H} S^2_h(\hat{\underline{\theta}})/R_h. \qquad (6.12)$$

Remark 6.1: When computing the entries in Table 6.5, one may meet the following problem, and even in the case with the full collection of EAs in the RZMS one does meet the problem.

> An (effective) stratum may contain just one (sub)division. If so, the variance in (6.10) becomes questionable from an algebraic point of view (being an expression of the type 0/0) as well as from a more intrinsic point of view.
>
> Let us first remove the "algebraic problem" by adopting the convention that a variance based on just one observation always is set to 0. Under this convention one still has a problem, though. In (6.12) some variance components, which all are positive quantities, are estimated by 0, which means sure under-estimation.

A common way to meet the last problem is;

> To "collapse" strata, i.e. to create new strata by joining original sampling strata (with believed similar mean and variation structures) so that each new stratum contains at least two sampled (sub)divisions, thereby admitting non-degenerate variance estimates. (Option $\alpha$)

Another possibility is;

> To use the previous convention that "one-observation variances" are set to zero, also in (6.12). (Option $\beta$)

An intermediate possibility is;

> To estimate the stratum variance for a "degenerate" stratum by "borrowing" an appropriate variance estimate from another (non-degenerate) stratum which is judged to have similar variation characteristics. (Option $\tau$)

Option $\beta$ yields a reasonable approach if the estimate for which one wants to estimate the sampling error receives only minor contributions from the degenerate sampling strata (i.e. strata with just one (sub)division). Then the estimate (6.12) gives only slight underestimation of the true estimator variance.

In the full RZMS(EA), there are in all 30 effective sampling strata. Of these, six contain only one (sub)division and all of them lie in the sector "small scale farming areas". These sampling strata carry only some 1.5% of the "effective" population. Therefore, for most domains of study the degenerate strata contribute only little.

We therefore recommend the use of Option $\beta$ when estimating sampling errors for estimates to which the degenerate sampling strata contribute little, and to sustain from computation sampling errors for estimates to which the "degenerate" strata contribute considerably. Most estimates will be of the former type, and estimates of the latter type can be questioned on quite general grounds because they will be very unreliable. As an extra support for Option $\beta$, we note that other approximations for variances go in the conservative direction so we can "afford" the Option-$\beta$ approximation which goes in the opposite direction.

A main reason for advocating Option $\beta$ rather than $\alpha$ is that Option $\alpha$ leads to a more involved handling of the data files. Under Option $\alpha$ one would have to add information about which collapsed stratum a record belongs to. In our belief the extra efforts would lead to only marginal changes and the efforts would not pay. Option $\alpha$ would be feasible, though. Similar objections can be raised vis-a-vi Option $\tau$, even if that option would be less complicated to handle than Option $\alpha$. ▒

Remark 6.2: Variance estimation for the group size estimator in (5.7) can of course be regarded as a special case of the above procedure, namely the special case which is obtained by setting $\underline{x} = \underline{1}$ (cf.(2.14) and (2.15)). ▒

Remark 6.3: For a self-weighting household sample, the general estimator (5.6) takes the form (5.10). It is readily seen that in this case, the variables in (6.5) can be computed "directly" as follows,

$$\hat{\theta}_{hi}(\underline{x};D) = (a_h/f)\cdot\sum_{j=1}^{m_{hi}} x_{hij}\cdot\underline{1}_D(hij). \qquad (6.13)$$

The previous variance estimation procedure can then be applied with $\theta_{hi}$-estimates as in (6.13).

For self-weighting samples, the estimation procedure can be carried out with less sampling design information than that in the file in Table 6.1. It is readily seen that a starting file of the following appearance suffices.

6.7

| Identification | | | | Sampling design parameters | | | | Variables | |
|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | Hh | $a_h$ | $R_h$ | $b_{hr}$ | 1/f | $\underline{x}$ | $\underline{q}$ |
| . | . | . | . | . | . | . | . | . | . |
| h | r | i | j | $a_h$ | $R_h$ | $b_{hr}$ | 1/f | $x_{hij}$ | $q_{hij}$ |
| . | . | . | . | . | . | . | . | . | . |

Table 6.6. Possible starting form for the households file (reduced to $\underline{x}$ and $\underline{q}$) in the case with a self-weighting household sample.

Thereby we have carried through estimation of estimator variances by the UC-procedure in the case when the estimator concerns the "total for a group of households", and we turn to the case "mean for a group of households".

6.1.2. For estimates of means for groups of households.

Also here we let $\underline{x}$ and D denote the household variable and the households group under interest. We shall consider estimation of the sampling error of the estimator of $\mu(\underline{x};D)$ in (5.8). Again, the UC-procedure in stratified version will be our main tool, here together with Lemma A3.1 in Appendix 3.

As before, the point of departure is the (reduced) households file in Table 6.1. In the next reduction step, we compute the quantities in (6.14) and (6.15) which both are of the type $Y_{hi}/\beta_{hi}$, in (6.14) with Y as in (6.4) and in (6.15) with Y in (6.4) specialized by setting $\underline{x}=\underline{1}$,

$$\hat{\theta}_{hi}(\underline{x};D)=(S_h/S_{hi})\cdot(M_{hi}/m_{hi})\cdot\sum_{j=1}^{m_{hi}} x_{hij}\cdot\underline{1}_D(hij), \qquad (6.14)$$

$$\hat{g}_{hi}(D) = (S_h/S_{hi})\cdot(M_{hi}/m_{hi})\cdot\sum_{j=1}^{m_{hi}} \underline{1}_D(hij). \qquad (6.15)$$

From Assumption 3.1 and Lemma A2.4 we conclude the following.

The pairs $(\hat{\theta}_{hi}(\underline{x};D),\hat{g}_{hi}(D))$ can, at least approximately, be viewed as independent random vectors, (6.16)

the components of which give EA-wise, unbiased estimates of group D $\underline{x}$-totals respectively group D sizes in the strata. (6.17)

| Identification | | | Sampling design parameters | | | Variables | |
|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | $a_h$ | $R_h$ | $b_{hr}$ | $\hat{\theta}_{hi}(\underline{x};D)$ | $\hat{g}_{hi}(D)$ |
| . | . | . | . | . | . | . | . |
| h | r | i | $a_h$ | $R_h$ | $b_{hr}$ | $\hat{\theta}_{hi}(\underline{x};D)$ as in (6.14) | $\hat{g}_{hi}(D)$ as in (6.15) |
| . | . | . | . | . | . | . | . |

Table 6.7. EA-wise estimates of group D $\underline{x}$-totals and of group D sizes in the sampling strata.

The next step towards the estimate $\hat{\mu}(\underline{x};D)$ is to average within strata in the file in Table 6.7, i.e. to compute

$$\hat{\theta}_h(\underline{x};D) = (1/a_h) \cdot \sum_{i=1}^{a_h} \hat{\theta}_{hi}(\underline{x};D), \tag{6.18}$$

$$\hat{g}_h(D) = (1/a_h) \cdot \sum_{i=1}^{a_h} \hat{g}_{hi}(D). \tag{6.19}$$

The collection of $\hat{\theta}_h(\underline{x};D)$- and $\hat{g}_h(D)$-values is illustrated in Table 6.8.

| Identificat. | Variables | |
|---|---|---|
| Sampling stratum | Estimate of group D $\underline{x}$-total in stratum | Estimate of group D size in stratum |
| 1 | $\hat{\theta}_1(\underline{x};D)$ | $\hat{g}_1(D)$ |
| . | . | . |
| h | $\hat{\theta}_h(\underline{x};D)$ as in (6.18) | $\hat{g}_h(D)$ as in (6.19) |
| . | . | . |
| H | $\hat{\theta}_H(\underline{x};D)$ | $\hat{g}_H(D)$ |

Table 6.8. File with estimates of group D $\underline{x}$-totals and of group D sizes in the sampling strata.

Thereafter, summations over the sampling strata yield estimates of $\theta(\underline{x};D)$ and $g(D)$ (cf. (6.8)). The estimate of $\mu(\underline{x};D)$ is then computed as,

$$\hat{\mu}(\underline{x};D) = \sum_{h=1}^{H} \hat{\theta}_h(\underline{x};D) / \sum_{h=1}^{H} \hat{g}_h(D). \tag{6.20}$$

It is readily seen that the estimates (6.20) and (5.8) agree.

To compute an estimate of the sampling error for the group mean

estimator, first average within (sub)divisions,

$$\hat{\theta}_{h(r)}(\underline{x};D) = (1/b_r)\cdot\sum_{i\epsilon G_r}\hat{\theta}_{hi}(\underline{x};D), \tag{6.21}$$

$$\hat{g}_{h(r)}(D) = (1/b_r)\cdot\sum_{i\epsilon G_r}\hat{g}_{hi}(D). \tag{6.22}$$

| Identification | | Variables | |
|---|---|---|---|
| Sampl. stratum | Subdiv | $\hat{\theta}_{h(\cdot)}(\underline{x};D)$ | $\hat{g}_{h(\cdot)}(D)$ |
| . | . | . | . |
| . | . | . | . |
| h | 1 | $\hat{\theta}_{h(1)}(\underline{x};D)$ | $\hat{g}_{h(1)}(D)$ |
| h | . | . | . |
| h | r | $\hat{\theta}_{h(r)}(\underline{x};D)$ as in (6.21) | $\hat{g}_{h(r)}(D)$ as in (6.22) |
| h | . | . | . |
| h | $R_h$ | $\hat{\theta}_{h(R_h)}(\underline{x};D)$ | $\hat{g}_{hR_h}(D)$ |
| . | . | . | . |

Table 6.9. File with (sub)division-wise estimates of group D $\underline{x}$-totals and group D sizes in the sampling strata.

In view of (6.16) and (6.17), $\hat{\theta}_{h(r)}(\underline{x};D)$ and $\hat{g}_{h(r)}(D)$ can be viewed as (sub)division-wise, unbiased estimates of $\theta_h(\underline{x};D)$ and $g_h(D)$. Hence, alternative estimates of $\theta(\underline{x};D)$ and $g(D)$ are,

$$\hat{\theta}^*(\underline{x};D) = \sum_{h=1}^{H}(1/R_h)\cdot\sum_{r=1}^{R_h}\hat{\theta}_{h(r)}(\underline{x};D), \tag{6.23}$$

$$\hat{g}^*(D) = \sum_{h=1}^{H}(1/R_h)\cdot\sum_{r=1}^{R_h}\hat{g}_{h(r)}(D). \tag{6.24}$$

The estimates in (6.23) and (6.24) lead to the following alternative estimator for $\mu(\underline{x};D)$,

$$\hat{\mu}^*(\underline{x};D) = \sum_{h=1}^{H}(1/R_h)\cdot\sum_{r=1}^{R_h}\hat{\theta}_{h(r)}(\underline{x};D) \Big/ \sum_{h=1}^{H}(1/R_h)\cdot\sum_{r=1}^{R_h}\hat{g}_{h(r)}(D). \tag{6.25}$$

In accordance with the general philosophy in the UC-procedure,

we use an estimate of $V[\hat{\mu}^*(\underline{x};D)]$ to estimate $V[\hat{\mu}(\underline{x};D)]$. To derive such a variance estimate, first apply Lemma A3.1 on (6.25) and use the fact that the samples from the different sampling strata are assumed to be independent, to get

$$V[\hat{\mu}^*(\underline{x};D)] \approx$$

$$\approx V\Big[\sum_{h=1}^{H}(1/R_h)\cdot\sum_{r=1}^{R_h}\{\hat{\theta}_{h(r)}(\underline{x};D)-\mu^*(\underline{x};D)\cdot\hat{g}_{h(r)}(D)\}\Big]/(E[\hat{g}^*(D)])^2 =$$

$$= \sum_{h=1}^{H} V[(1/R_h) \cdot \sum_{r=1}^{R_h} \{\hat{\theta}_{h(r)}(\underline{x};D) - \mu^*(\underline{x};D) \cdot \hat{g}_{h(r)}(D)\}]/(E[\hat{g}^*(D)])^2, \quad (6.26)$$

where

$$\mu^*(\underline{x};D) = E[\hat{\theta}^*(\underline{x};D)]/E[\hat{g}^*(D)]. \quad (6.27)$$

As a consequence of (6.16) and (6.17), the r-summations in (6.26) run over terms which are independent and which all have the same expectation. Hence, (6.26) and Lemma A1.1 yield, with notation in accordance with (A3.7),

$$V[\hat{\mu}^*(\underline{x};D)] \approx \sum_{h=1}^{H} S^2(\hat{\underline{\theta}}_h(\underline{x};D) - \mu^*(\underline{x};D) \cdot \hat{\underline{g}}_h(D))/(R_h \cdot E[\hat{g}^*(D)])^2. \quad (6.28)$$

However, in (6.28) the quantities $\mu^*(\underline{x};D)$ and $E[\hat{g}^*(D)]$ are unknown. To obtain a "computable" estimate, we exchange these quantities by their observed counterparts, i.e. by

$\hat{\mu}^*(\underline{x};D)$ and $\hat{g}^*(D)$. Thereby we arrive at the following formula for variance estimation,

$$\hat{V}[\hat{\mu}^*(\underline{x};D)] = \sum_{h=1}^{H} S^2(\hat{\underline{\theta}}_h(\underline{x};D) - \hat{\mu}^*(\underline{x};D) \cdot \hat{\underline{g}}_h(D))/(R_h \cdot \hat{g}^*(D)^2). \quad (6.29)$$

As a preparation for the computation of the right hand side in (6.29), we compute the following quantities, the collection of which is illustrated in Table 6.10 below,

$$S^2_h(\hat{\underline{\theta}}) = S^2(\hat{\underline{\theta}}_h(\underline{x};D)) \text{ in accordance with (A3.7)}, \quad (6.30)$$

$$S^2_h(\hat{\underline{g}}) = S^2(\hat{\underline{g}}_h(D)) \text{ in accordance with (A3.7)}, \quad (6.31)$$

$$C_h(\hat{\underline{\theta}},\hat{\underline{g}}) = C(\hat{\underline{\theta}}_h(\underline{x};D)),\hat{\underline{g}}_h(D)) \text{ in accordance with (A3.8)}. \quad (6.32)$$

| Identif | Variables | | |
|---|---|---|---|
| Strat | Variance of $\theta$ | Variance of g | Covariance of $\theta$ and g |
| 1 | $S^2_1(\hat{\underline{\theta}})$ | $S^2_1(\hat{\underline{g}})$ | $C_1(\hat{\underline{\theta}},\hat{\underline{g}})$ |
| . | . | . | . |
| h | $S^2_h(\hat{\underline{\theta}})$ as in (6.30) | $S^2_h(\hat{\underline{g}})$ as in (6.31) | $C_h(\hat{\underline{\theta}},\hat{\underline{g}})$ as in (6.32) |
| . | . | . | . |
| H | $S^2_H(\hat{\underline{\theta}})$ | $S^2_H(\hat{\underline{g}})$ | $C_H(\hat{\underline{\theta}},\hat{\underline{g}})$ |

Table 6.10: Variances and covariances for $\hat{\theta}_{hi}$ and $\hat{g}_{hi}$.

The final step in the computation of the estimator variance is to apply the formula (A3.9) in (6.29), which leads to the following estimate of the estimator variance,

$$\hat{V}[\hat{\mu}(\underline{x};D)] =$$

$$=\sum_{h}^{H}[S^2(\hat{\underline{\theta}}_h)+\hat{\mu}^*(\underline{x};D)^2\cdot S^2(\hat{\underline{g}}_h)-2\hat{\mu}^*(\underline{x};D)\cdot C(\hat{\underline{\theta}}_h,\hat{\underline{g}}_h)]/(R_h\cdot\hat{g}^*(D))^2. \quad (6.33)$$

Remark 6.4: The contents in Remark 6.1 is applicable also here. For completeness we add to the convention in Remark 6.1 that covariances based on just one observation should be set to 0. Option $\beta$ is recommended also in this case, on essentially the same grounds as before. ▒

Remark 6.5: Variance estimation for the estimator (5.9) of a group proportion can of course be regarded as a special case of the above procedure, namely the case which corresponds to $\underline{x} = \underline{1}_A$. ▒

Remark 6.6: For a self-weighting household sample, the general estimator (5.8) takes the form (5.12). It is readily seen that in this case the variables in (6.14) and (6.15) can be computed as,

$$\hat{\underline{\theta}}_{hi}(\underline{x};D) = (a_h/f)\cdot\sum_{j=1}^{m_{hi}} x_{hij}\cdot\underline{1}_D(hij), \quad (6.34)$$

$$\hat{\underline{g}}_{hi}(D) = (a_h/f)\cdot\sum_{j=1}^{m_{hi}} \underline{1}_D(hij). \quad (6.35)$$

The previous procedure is then carried out with values as in (6.34) and (6.35). As before, for self-weighting samples the variance estimation procedure can be carried out with less sampling design information than in Table 6.1. It is readily seen that the file in Table 6.6 suffices also in this case. ▒

### 6.1.3. For estimates of totals for groups of individuals.

Here $\underline{x}$ and D denote an individual-variable respectively a group of individuals. Our aim is to derive an estimate of the variance of the estimator of $\Theta(\underline{x};D)$ in (5.19). This case can be handled along very much the same lines as the case in Subsection 6.1.1. The point of departure is the individuals file as illustrated in Table 6.11 below.

| Identification | | | | | Sampling design parameters | | | | | | | Variables | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | Hh | Ind | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $M_{hi}$ | $m_{hi}$ | $\underline{x}$ | $\underline{q}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| h | r | i | j | k | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $M_{hi}$ | $m_{hi}$ | $x_{hijk}$ | $q_{hijk}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Table 6.11. The individuals file (reduced to $\underline{x}$ and $\underline{q}$).

The variables $y_{hi}$ and $Y_{hi}$ are here defined as follows,

$$y_{hi} = \sum_{j=1}^{M_{hi}} \sum_{k=1}^{K_{hij}} x_{hijk} \cdot \underline{1}_D(hijk), \tag{6.36}$$

$$Y_{hi} = (M_{hi}/m_{hi}) \cdot \sum_{j=1}^{m_{hi}} \sum_{k=1}^{K_{hij}} x_{hijk} \cdot \underline{1}_D(hijk). \tag{6.37}$$

After the above changes, proceed as in Sub-section 6.1.1. The details are left to the reader.

In the case with a self-weigting sample, the quantity $\hat{\theta}_{hi}$ (cf. (6.5)) can be computed as

$$\hat{\theta}_{hi} = (1/f) \cdot \sum_{j=1}^{m_{hi}} \sum_{k=1}^{K_{hij}} x_{hijk} \cdot \underline{1}_D(hijk). \tag{6.38}$$

6.1.4. For estimates of means for groups of individuals.

Here the concern is to derive an estimate of the variance of the estimator of $\mu(\underline{x};D)$ in (5.21). This case can be treated along almost exactly the same lines as in Sub-section 6.1.2. The only modifications are firstly that $y_{hi}$ and $Y_{hi}$ should be defined by (6.36) and (6.37) and secondly that the definition of the quantity in (6.15) should be changed to

$$\hat{g}_{hi}(D) = (S_h/S_{hi}) \cdot (M_{hi}/m_{hi}) \cdot \sum_{j=1}^{m_{hi}} \sum_{k=1}^{K_{hij}} \underline{1}_D(hijk). \tag{6.39}$$

After these changes, proceed as in Sub-section 6.1.2. Again the details are left to the reader. Previous comments on self-weighting samples apply also here (cf. Remarks 6.1 and 6.4).

## 6.2 Estimation of the sampling errors for estimates based on household samples from the RZMS(Segm).

In the essentials the procedures for RZMS(Segm) samples parallel those which are presented in Sub-sections 6.1.1 and 6.1.2, now justified by Assumption 3.2. One modification, which is a consequence of the formula (3.28), is that the $\beta_{hi}$:s here should be as follows,

$$\beta_{hi} = (S_{hi}/S_h)/d_{hi}. \tag{6.40}$$

### 6.2.1. For estimates of totals for groups of households.

The task is to present an estimate of the variance of the estimator in (5.14). As in Sub-section 6.1.1, the point of departure is the households file, here in its RZMS(Segm) version. As before, $\underline{q}$ denotes a variable (possibly vector valued) which is group-D-specifying.

| Identification | | | | Sampling design parameters | | | | | | | | Variables | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | Hh | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $Q_{hi}$ | $d_{hi}$ | $m_{hi}$ | $\underline{x}$ | $\underline{q}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| h | r | i | j | $S_h$ | $n_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $Q_{hi}$ | $d_{hi}$ | $m_{hi}$ | $x_{hij}$ | $q_{hij}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Table 6.12. The households file (reduced to $\underline{x}$ and $\underline{q}$).

The variables $y_{hi}$ and $Y_{hi}$ are defined as follows.

$$y_{hi} = \sum_{j=1}^{Q_{hi}} x_{hij} \cdot \underline{1}_D(hij),$$

$$Y_{hi} = (Q_{hi}/m_{hi}) \cdot \sum_{j=1}^{m_{hi}} x_{hij} \cdot \underline{1}_D(hij). \tag{6.41}$$

Note that (6.40) and (6.41) lead to the following counterpart of the variable in (6.5),

$$\hat{\theta}_{hi}(\underline{x};D) = d_{hi} \cdot (S_h/S_{hi}) \cdot (Q_{hi}/m_{hi}) \cdot \sum_{j=1}^{m_{hi}} x_{hij} \cdot \underline{1}_D(hij). \tag{6.42}$$

Then, proceed as in Sub-section 6.1.1. The details are left to the reader.

6.2.2. For estimates of means for groups of households.

The task is to estimate the variance of the estimator in (5.16). For this we can use the procedures in Sub-section 6.1.2 with some minor modifications. The definition (6.42) is used and the counterpart of the quantity in (6.15) is

$$\hat{g}_{hi}(D) = d_{hi} \cdot (S_h/S_{hi}) \cdot (Q_{hi}/m_{hi}) \cdot \sum_{j=1}^{m_{hi}} \underline{1}_D(hij). \qquad (6.43)$$

After these modifications, the procedure in Sub-section 6.1.2 can be employed. The details are left to the reader.

6.2.3 For estimates of totals for groups of individuals.

Here the desire is to estimate the variance of the estimator of $\theta(\underline{x};D)$ in (5.27). The algorithm to be used is the one in Sub-section 6.1.3, after some preliminary modifications. The modification of $\beta$ is stated in (6.40). The point of departure is the individuals file (in its RZMS(Segm) version), as illustrated in Table 6.13.

| Identification | | | | | Sampling design parameters | | | | | | | | Variables | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | Hh | Ind | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $Q_{hi}$ | $d_{hi}$ | $m_{hi}$ | $\underline{x}$ | $\underline{q}$ |
| . | . | . | . | . | . | . | . | . | . | . | | . | . | . |
| h | r | i | j | k | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $Q_{hi}$ | $d_{hi}$ | $m_{hi}$ | $x_{hijk}$ | $q_{hijk}$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Table 6.13. The individuals file (reduced to $\underline{x}$ and $\underline{q}$).

The variables $y_{hi}$ and $Y_{hi}$ are introduced as follows

$$y_{hi} = \sum_{j=1}^{Q_{hi}} \sum_{k=1}^{K_{hij}} x_{hijk} \cdot \underline{1}_D(hijk), \qquad (6.44)$$

$$Y_{hi} = (Q_{hi}/m_{hi}) \cdot \sum_{j=1}^{m_{hi}} \sum_{k=1}^{K_{hij}} x_{hijk} \cdot \underline{1}_D(hijk). \qquad (6.45)$$

Thereafter the procedure in Sub-section 6.1.3 is applied.

6.2.4 For estimates of means for groups of individuals.

Here the procedure in Sub-section 6.1.4 can be applied after modifications along the lines in the previous sub-section. The details are left to the reader.

## 6.3 Variance estimation for the ratio type estimators.

In this section we shall consider estimation of the sampling error for the estimators in (5.40) and (5.48). We start with that in (5.40), and we adhere to the notation in Sub-section 5.4. In particular B denotes an area which is a union of EAs (cf. (5.31)).

In view of (5.40) we have,

$$V[\widehat{g(B;87)}_R] = g(B;c)^2 \cdot V[\hat{\tau}(B)]. \qquad (6.46)$$

From (6.46) it is seen that we are through if we can exhibit an estimate of $V[\hat{\tau}(B)]$, and that will be our next goal.

Also here we give the description in terms of reductions of a basic file; which in this context is the RZMS EA/segment file which is discussed in Section 7 (see Table 7.1), and which is repeated in Table 6.14 below.

| Identification | | | Sampling design parameters | | | | | Variables | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $s_{hi}$ | $M_{hi}$ | $T_{hi}(82)$ | $T_{hi}(87)$ |
| . | . | . | . | . | . | . | . | . | . | . | . |
| h | r | i | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $s_{hi}$ | $M_{hi}$ | $T_{hi}(82)$ | $T_{hi}(87)$ |
| . | . | . | . | . | . | . | . | . | . | . | . |

Table 6.14. The RZMS EA/segment file.

The present variance estimation problem is in fact only a variation of the problem which was considered in Sub-section 6.1.2. For the sake of completeness we shall write down the computation algorithm, even if it does not contain any new ideas. However, this time we leave the justifications to the reader; they are obtained by parallelling the reasoning in Sub-section 6.1.2.

The first reduction step consists in deriving the following EA-wise estimates of the number of B-households in the sampling strata, in 1987 respectively at the (1982) census,

$$\hat{g}_{hi}(B;87) = (S_h/S_{hi}) \cdot M_{hi} \cdot \delta((hi);B), \qquad (6.47)$$

$$\hat{g}_{hi}(B;c) = S_h \cdot \delta((hi);B). \qquad (6.48)$$

The following procedure is simply a "copy" of the algorithm in Sub-section 6.1.2 with

$\hat{\theta}_{hi}(\underline{x};D)$ changed to $\hat{g}_{hi}(B;87)$, (6.49)

$\hat{g}_{hi}(D)$ changed to $\hat{g}_{hi}(B;c)$. (6.50)

| Identification | | | Sampling design parameters | | | Variables | |
|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | $a_h$ | $R_h$ | $b_{hr}$ | $\hat{g}_{hi}(B;87)$ | $\hat{g}_{hi}(B;c)$ |
| . | . | . | . | . | . | . | . |
| h | r | i | $a_h$ | $R_h$ | $b_{hr}$ | $\hat{g}_{hi}(B;87)$ as in (6.47) | $\hat{g}_{hi}(B;c)$ as in (6.48) |
| . | . | . | . | . | . | . | . |

Table 6.15. File with EA-wise estimates of the number of area B households in the strata.

The next step towards the estimate $\hat{\tau}(B)$ is to average within strata in the file in Table 6.15, i.e. to compute

$$\hat{g}_h(\underline{x};87) = (1/a_h) \cdot \sum_{i=1}^{a_h} \hat{g}_{hi}(B;87), \quad (6.51)$$

$$\hat{g}_h(B;c) = (1/a_h) \cdot \sum_{i=1}^{a_h} \hat{g}_{hi}(B;c). \quad (6.52)$$

| Identif. | Variables | |
|---|---|---|
| Stratum | Estimate of #(hh in B) in stratum,1987 | Estimate of #(hh in B) in stratum,census |
| 1 | $\hat{g}_1(B;87)$ | $\hat{g}_1(B;c)$ |
| . | . | . |
| h | $\hat{g}_h(B;87)$ as in (6.51) | $\hat{g}_h(B;c)$ as in (6.52) |
| . | . | . |
| H | $\hat{g}_H(B;87)$ | $\hat{g}_H(B;c)$ |

Table 6.16. File with stratum-wise estimates of the number of area B households in the sampling strata.

Next, summations over strata in Table 6.16 yield estimates of $g(B;87)$ and $g(B;c)$, and the estimate of $\tau(B)$ is then computed as the ratio between these two estimates,

$$\hat{\tau}(B) = \sum_{h=1}^{H} \hat{g}_h(B;87) / \sum_{h=1}^{H} \hat{g}_h(B;c). \quad (6.53)$$

It is readily checked that the estimate in (6.53) is the same as that given by the estimator in (5.39).

To compute an estimate of the sampling error of $\hat{\tau}(B)$ we first average within (sub)divisions,

$$\hat{g}_{h(r)}(B;87) = (1/b_{hr}) \cdot \sum_{i \epsilon G_r} \hat{g}_{hi}(B;87), \tag{6.54}$$

$$\hat{g}_{h(r)}(B;c) = (1/b_{hr}) \cdot \sum_{i \epsilon G_r} \hat{g}_{hi}(B;c). \tag{6.55}$$

| Identification | | Variables | |
|---|---|---|---|
| Stratum | Subdiv | $\hat{g}_{h(\cdot)}(B;87)$ | $\hat{g}_{h(\cdot)}(B;c)$ |
| . | . | . | . |
| h | 1 | $\hat{g}_{h(1)}(B;87)$ | $\hat{g}_{h(1)}(B;c)$ |
| h | . | . | . |
| h | r | $\hat{g}_{h(r)}(B;87)$ as in (6.54) | $\hat{g}_{h(r)}(B;c)$ as in (6.55) |
| h | . | . | . |
| h | $R_h$ | $\hat{g}_{h(R_h)}(B;87)$ | $\hat{g}_{h(R_h)}(B;c)$ |
| . | . | . | . |

Table 6.17. File with (sub)division-wise estimates of the number of area B households in the strata.

Alternative estimates of g(B;87) and g(B;c) are,

$$\hat{g}^*(B;87) = \sum_{h=1}^{H} (1/R_h) \cdot \sum_{r=1}^{R_h} \hat{g}_{h(r)}(B;87), \tag{6.56}$$

$$\hat{g}^*(B;c) = \sum_{h=1}^{H} (1/R_h) \cdot \sum_{r=1}^{R_h} \hat{g}_{h(r)}(B;c). \tag{6.57}$$

The estimates in (6.56) and (6.57) lead to the following alternative estimate of $\tau(B)$,

$$\hat{\tau}^*(B;87) = \sum_{h=1}^{H} (1/R_h) \cdot \sum_{r=1}^{R_h} \hat{g}_{h(r)}(B;87) / \sum_{h=1}^{H} (1/R_h) \cdot \sum_{r=1}^{R_h} \hat{g}_{h(r)}(B;c). \tag{6.58}$$

In accordance with the general philosophy in the UC-procedure we use an estimate of $V[\hat{\tau}^*(B)]$ to estimate $V[\hat{\tau}(B)]$. To obtain such an estimate, apply Lemma A3.1 and use the fact that the samples from the different strata are assumed to be independent. Then proceed as in (6.27)-(6.33), i.e. compute

$$S^2_h(87) = S^2(\hat{g}_h(B;87)) \text{ in accordance with (A3.7)}, \tag{6.59}$$

$$S^2_h(c) = S^2(\hat{g}_h(B;c)) \text{ in accordance with (A3.7)}, \tag{6.60}$$

$$C_h(87,c) = C((\hat{g}_h(B;87)),\hat{g}_h(B;c)) \text{ as in (A3.8)}. \tag{6.61}$$

| Identif | Variables | | |
|---|---|---|---|
| Strat | Variance of $\hat{g}(87)$ | Variance of $\hat{g}(c)$ | Covariance of $\hat{g}(87)$ and $\hat{g}(c)$ |
| 1 | $S^2_1(87)$ | $S^2_1(c)$ | $C_1(87,c)$ |
| . | . | . | . |
| h | $S^2_h(87)$ as in (6.59) | $S^2_h(c)$ as in (6.60) | $C_h(87,c)$ as in (6.61) |
| . | . | . | . |
| H | $S^2_H(87)$ | $S^2_H(c)$ | $C_H(87,c)$ |

Table 6.18: File with sample variances and covariances for g(87) and g(c).

We come to the following variance estimate.

$$\hat{V}[\hat{\tau}(B)] = \tag{6.62}$$

$$= \sum_{h=1}^{H} [S_h^2(87) + \hat{\tau}^*(B)^2 \cdot S_h^2(c) - 2 \cdot \hat{\tau}^*(B) \cdot C_h(87,c)] / (R_h \cdot \hat{g}^*(B;c))^2,$$

By combining (6.46) and (6.62) we arrive at the following final formula for estimation of the variance of the estimate in (5.40).

$$\hat{V}[\widehat{g(B;87)}_R] = g(B;c)^2 \cdot \hat{V}[\hat{\tau}(B)], \tag{6.63}$$

where the last factor is given by (6.62).

Remark 6.7: The contents of Remark 6.1 have relevance also here. ▒

The case with estimation of the number of individuals in the area B can be treated quite analogously. The only difference is that $\hat{g}_{hi}(B;87)$ and $\hat{g}_{hi}(B;c)$ are defined as follows,

$$\hat{g}_{hi}(B;87) = (S_h/S_{hi}) \cdot T_{hi}(87) \cdot \delta((hi);B), \tag{6.64}$$

$$\hat{g}_{hi}(B;c) = (S_h/S_{hi}) \cdot T_{hi}(c) \cdot \delta((hi);B). \tag{6.65}$$

Then use the previous algorithm.

# 7. ON THE EVALUATION OF THE INTERCENSAL DEMOGRAPHIC SURVEY.

## 7.1. Objectives and execution of the Intercensal Demographic Survey.

The objectives for the Intercensal Demographic Survey (ICDS) were formulated as follows;

1. Updating of the area master sample.

2. To give statistics on population and on demographic and socio-economic variables.

3. To serve as a pilot survey for the 1992 Census.

The first step in the execution of the ICDS, consisted in the "drawing" of, or rather deciding upon, the ZMS EAs to be included in the revised master sample, the RZMS, as described in Section 3. The subsequent ICDS field work, which was carried out in 1987 and 1988, comprised three distinct rounds, the listing round (also called Round 0), Round 1 and Round 2.

The listing round was performed in May/June 1987. Each EA in the RZMS was visited and enumerated, leading to;

- a complete lists of the households in the EA,

- information on the number of individuals in the households in the EA.

The Round 1 was carried out in August/September of 1987. From the EA household lists established in Round 0, systematic samples of households were drawn so as to obtain a self-weighting household sample, which was accomplished by using the sampling interval in (3.27) with sampling fraction f=1/113. The total ICDS sample consisted of around 16 000 households.

As indicated before, the self-weighting property could have been violated if different response rates would have been obtained in the different EAs. However, this was avoided by substituting for households which did not respond, with households having similar characteristics according to the information gathered in the listing round. Thereby, in all EAs the response rates were exactly or very close to 100%.

The selected households were interviewed, and the following type of information was gathered;

- the head of household was identified,

- lists of the "usual members" in the household as well as of visitors in the household (during the night of the interview day) were established.

Furthermore, for each usual member and visitor various characteristics were recorded, as;

- age
- sex
- marital status
- education
- economic activity (occupation)
- and others.

Moreover, for the individuals, information was collected concerning change of place of residence during the last 12 months and for women of age 12 years or more there were asked various questions relating to child bearing, in particular the date of last birth. Questions were also asked concerning deaths in the household during the last 12 months. The main aims with these questions were to provide data for estimation of migration rates, birth rates and death rates.

From experience one knows that estimates of rates of migration, birth and death which are based on "retrospective" longitudinal data, as in Round 1, often are quite unreliable (chiefly because of memory weaknesses). As birth, death and migration rates are vital parameters in population projections, it was considered beneficial to make special efforts in order to get presumably better estimates for these demographic characteristics than could be expected from Round 1. The achievement of better rate estimates was the chief aim for Round 2.

The field work of Round 2 was carried out in August 1988, one year after Round 1. Exactly the same households as in Round 1 were to be interviewed by using a questionnaire which was similar to the one used in Round 1. The interviewers also brought information on the 1987 composition of the households (name and sex), thereby having a good basis for measuring changes due to in- and out-moving, births and deaths that had taken place during the last year. It should be noted that non-response becomes a more intricate problem for Round 2, because in that round household substitutions could not be made in cases with non-response.

Although there is close connections between Rounds 1 and 2 of the ICDS, Round 1 can be regarded as a separate survey in its own right, which perfectly well can be evaluated without access to Round 2 data. It is also the intention of CSO to evaluate ICDS, Round 1 separately. A main reason for this is that the Round 1 data were collected in 1987 while the Round 2 data were collected a year later. Accordingly, the Round 1 data will be available for processing a year ahead of the Round 2 data. Evaluation of the Round 2 data and combined evaluation of Rounds 1 and 2 are of course planned, but the estimation problems of those evaluations will be a later story, which we shall not enter into in this report.

Next some general words on the tabulation plans for the ICDS. The main objective for the listing round was that it should provide findings to be used in the revision of the ZMS. Accordingly, no "tabulation plan" was formulated for the listing round. However, the listing round provides some valuable infor-

mation on population and numbers of households, which complement the demographical data collected in Round 1. Therefore, the "population information" obtained in the listing round should be entered into the RZMS-documentation in an appropriate way, so that it can be used as valuable "auxiliary information" in the evaluation of the data from Round 1. More about this later on.

Concerning the tabulation plans for Round 1, there is an extensive material available, notably the CSO report of March 1987 and the reports by Arvidsson (1987) and Lagerlöf (1988). Lagerlöf presents tabulation specifications in a "close-to-EDB-processing" language. We shall not go into details on this point, just indicate the main lines. The planned tables should yield information on;

- population (i.e. the number of individuals) in various groups specified by conditions on variables as "geographical area", age, sex, marital status, etc,
- household numbers for different categories of head of household,
- educational conditions,
- economic activity,
- migration conditions,
- birth rates.
- death rates.

As already indicated, the main objectives for Round 2 is to render improved estimates of birth, death and migration rates. As the discussion of estimation problems will be confined to Round 1, we do not say more about tabulation plans for and the evaluation of Round 2. A more detailed discussion can be found e.g. in the report by Johansson (1988).

## 7.2 On the organization of the data from the listing round.

Although the material from the listing round primarily should serve the sampling technical purpose of establishing the revised master sample, the RZMS, it also contained data which are of value for estimation of population figures (see Sub-section 5.4). These data from the listing round should preferably be saved in the file which documents the RZMS. Extensive suggestions for the documentation of the RZMS are given in Annex 4 in ZIMSTAT:5, Part 1 (1989). We refer to that report for a detailed discussion, and here we confine ourselves to just a brief outline of the central file in the RZMS-documentation, called the RZMS EA/segment file, which is illustrated in Table 7.1.

First we repeat some notation which has been introduced earlier.

$S_h$ = the 1982 number of households in the sampling stratum,
$a_h$ = the EA sample size in the sampling stratum,
$R_h$ = the number of (sub)divisions which are represented in the sample from the sampling stratum,
$b_{hr}$ = the number of sampled EAs from the (sub)division,
$S_{hi}$ = the 1982 census number of households in the EA,
$M_{hi}$ = the 1987 number of households in the EA,
$T_{hi}(82)$ = the 1987 number of individuals in th EA.
$T_{hi}(87)$ = the 1987 number of individuals in th EA.

| Identification | | | Sampling design parameters | | | | | Variables | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Stratum | Subdiv | EA | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $S_{hi}$ | $M_{hi}$ | $T_{hi}(82)$ | $T_{hi}(87)$ |
| . | . | . | . | . | . | . | . | . | . | . | . |
| h | r | i | $S_h$ | $a_h$ | $R_h$ | $b_{hr}$ | $S_{hi}$ | $S_{hi}$ | $M_{hi}$ | $T_{hi}(82)$ | $T_{hi}(87)$ |
| . | . | . | . | . | . | . | . | . | . | . | . |

Table 7.1. (Part of) the RZMS EA/segment file with a general record.

Remark 7.1: As discussed in Sub-section 6.3, the file in Table 7.1 contains the data which are needed for application of the ratio estimators in Sub-section 5.4 and for estimating their sampling errors. ▒

Remark 7.2: In the file in Table 7.1, we have put $S_{hi}$ at two places, which of course is unnecessary in the practical implementation. The reason is that we want to stress the fact that $S_{hi}$ plays a double role in the estimation context which is mentioned in the previous remark, it serves as information on the sampling design as well as a "variable". ▒

## 7.3. On the evaluation the data from the ICDS, Round 1.

The file organization for the Round 1 data was decided upon after an infological and datalogical analysis of the ICDS, which is presented in the report by Lagerlöf (1988). The following basic files are used;

- a households file,
- an individuals file,
- a file for deceased,
- a file for fertile women.

Lagerlöf (1988) gives a thorough presentation of the files, their interrelations, the contents of their records, etc. The structures of the files are in accordance with the suggestions in Section 4, but for the fact that no sampling design parameters are included in the files. Nevertheless, tabulations have been carried out.

This was possible due to the special feature of the ICDS, that it has a self-weighting household sample. Then, as can be seen in Section 5, no information on the sampling design is needed but for the fact that the sampling rate was f = 1/113. However, when it comes to estimation of the sampling errors for the values tabulated so far, it will be inevitable to complement the files with appropriate information on sampling design parameters (see Section 6). The needed information will be contained in the RZMS EA/segment file if it is established in accordance with the suggestions in ZIMSTAT:5, Part 1 (1989). As documentation of the RZMS should have very high priority, we recommend that the RZMS-documentation is carried out before entering on the problem of estimating sampling errors. The RZMS-documentation will ensure that no vital sapling design information for the ICDS is lost.

# APPENDIX 1. ON THE ULTIMATE CLUSTERS/RANDOM GROUP METHOD FOR ESTIMATION OF SAMPLING ERRORS.

The task of estimating the sampling errors for estimates based on RZMS samples is somewhat intricate, and an entirely rigourous treatment of the problem would require more precise knowledge of the sampling characteristics than is available. We shall circumvent the obstacles by relying on a procedure which is somewhat approximate, and also perhaps somewhat inefficient compared with sampling error estimates that could have been used had all desired sampling characteristics been known. The procedure to be used belongs to the class of methods which goes under the name "the method of ultimate clusters", also referred to by the term "the random groups method". However, the efficiency loss i most likely small and the ultimate clusters method has the great merit of leading to comparatively simple computations.

As detailed discussions of the method of ultimate clusters/-random groups can be found in the literature, e.g. in Chapter 2 in Wolter (1986), we do not aim at a complete presentation here. Our aim is to provide enough background for a fairly easy understanding and checking of the procedures suggested for RZMS surveys and notably for the ICDS.

Henceforth, $E[\cdot]$ and $V[\cdot]$ denote expectation and variance of the random quantity within the brackets.

## A1.1. On the ultimate clusters method.

The following well-known result (see e.g. Wolter (1985), Theorem 2.2.1) will be fundamental.

LEMMA A1.1: Let $Z_1, Z_2, \ldots, Z_R$ be independent random variables which all have the same expected values. The corresponding sample mean is denoted

$$\bar{Z} = (1/R) \cdot \sum_{r=1}^{R} Z_r \ . \qquad \text{(A1.1)}$$

Let $S^2$ denote the "ordinary" sample variance for the variables, i.e.

$$S^2 = 1/(R-1) \cdot \sum_{r=1}^{R} (Z_r - \bar{Z})^2 \ . \qquad \text{(A1.2)}$$

Then,

$S^2/R$ is an unbiased estimator of $V[\bar{Z}]$. (A1.3)

Remark A1.1: Note that no assumption is made about equal variances for the Z-variables. ▒

Let U=(1,2,...,N) be a finite population and $\underline{y} = (y_1, y_2, \ldots, y_N)$ a variable on U. The $\underline{y}$-total over the population is denoted

$$\theta(\underline{y}) = \sum_{i=1}^{N} y_i \ . \qquad \text{(A1.4)}$$

Such totals will be the main "targets" for our subsequent estimation efforts.

We consider a general probability sample from U, and we adhere to the terminology and notation which is introduced in Appendix 2. The sample is assumed to be drawn without replacement and with predetermined sample size n. The corresponding sample inclusion indicators are denoted by $I_1, I_2, \ldots, I_N$, and the inclusion probabilities, $\pi$, and inclusion proportionates, $\beta$, by

$$\pi_i = n \cdot \beta_i = P(I_i=1) \ , \ i = 1,2,\ldots,N. \qquad \text{(A1.5)}$$

As is well known, the following estimator $\hat{\theta}(\underline{y})$, the Horwitz-Thompson estimator, yields unbiased estimation of the population total $\theta(\underline{y})$,

$$\hat{\theta}(\underline{y}) = (1/n) \cdot \sum_{i=1}^{N} (y_i/\beta_i) \cdot I_i \ . \qquad \text{(A1.6)}$$

If the second order inclusion probabilities for the sampling procedure are known, an exactly unbiased estimator of the estimator variance $V[\hat{\theta}(\underline{x})]$ can be written down. However, we shall not pursue such a route. Instead of making very specific assumptions about second order inclusion probabilities we shall confine ourselves to situations where some general assumptions on approximate independence are judged to be satisfied. We start by considering the following assumption.

> ASSUMPTION A1.1: The sampled items can, with good approximation, be regarded as the outcomes of n independent selections $T_1, T_2, \ldots, T_n$ of items from the population U. In each selection we have
>
> $$P(T_v=i) = \beta_i, \ i=1,2,\ldots,N, \ v=1,2,\ldots,n. \qquad \text{(A1.7)}$$

Under the Assumption A1.1, the estimator in (A1.6) can be viewed as follows,

$$\hat{\theta}(\underline{y}) = (1/n) \cdot \sum_{v=1}^{n} (y_{T_v}/\beta_{T_v}), \qquad \text{(A1.8)}$$

where the summation in (A1.8) goes over terms which are independent random variables. Furthermore, as is discussed in more detail in Lemma A2.5 in Appendix 2, we have

$$E[y_{T_v}/\beta_{T_v}] = \theta(\underline{y}), \ v=1,2,\ldots,n. \qquad \text{(A1.9)}$$

Combination of the above with Lemma A1.1 leads to the following

result, which is the ultimate clusters method in its very simplest version.

RESULT A1.1: Under Assumption A1.1, the formula (A1.10) below yields an approximately unbiased estimator of the estimator variance $V[\hat{\theta}(\underline{y})]$,

$$\hat{V}[\hat{\theta}(\underline{y})] = 1/n(n-1)\cdot\sum_{v=1}^{n}[(y_{T_v}/\beta_{T_v})-\hat{\theta}(\underline{y})]^2. \qquad \text{(A1.10)}$$

Remark A1.2: The variance estimator in (A1.10) is simply the sample variance for the observed values of $y_i/\beta_i$ divided by the number of observations in the sample. ▒

Remark A1.3: In the particular case when the sample is drawn by simple random sampling (i.e. when $\beta_1=\beta_2=\ldots=\beta_N$), the approximation in the variance estimator in (A1.10) consists in disregarding the finite population correction. This is a mild approximation if the sampling fraction is small. Note that the approximation goes in the conservative direction, i.e. the actual variance is overestimated.

For general probability samples, the approximation is of a similar nature provided the observations are drawn in a well "mixing" way (i.e. so as to be close to independent). ▒

Variance estimators of the type (A1.10) will, however, be too crude in the type of situations we are concerned with in the RZMS context, because the assumption about approximate independence among all the observations will not be sufficiently well fulfilled. We shall therefore consider variance estimation in a somewhat more elaborate frame-work. In particular we shall change Assumption A1.1 to the following one.

ASSUMPTION A1.2: For some grouping $G_1,G_2,..,G_R$ of the sampled items, the item groups can, with good approximation, be regarded as being drawn independently of each other.

Remark A1.4: One type of situations where Assumption A1.2 is applicable is as follows. Assume that the sample under consideration is drawn by a two-stage procedure. Let the index i label the second-stage sampling units, while indication for first-stage sample units is suppressed. As regards the first sampling stage, we assume that we only know which observations that come from the same first-stage sampling units. Let $G_1,G_2,\ldots,G_R$ denote the grouping generated by bringing together the observations from the same sampled first-stage unit. Then, provided that the first stage sample has a fairly small sampling fraction, Assumption A1.2 is satisfied.(Cf. Lemma A2.5.)

In fact, the above concretization lies behind the notion of

"ultimate cluster". The first-stage sampling units are regarded to be the "ultimate" (= the very largest) clusters. Note, that for sampling units in stage-wise sampling, one usually counts in the opposite direction; The "ultimate" sampling units are the very smallest ones. ▒

Remark A1.5: Although second order inclusion probabilities are not mentioned explicitly, Assumption A1.2 can be regarded as an approximation assumption concerning (at least some of) the second order inclusion probabilities. Note, though, that the assumption about knowledge of first order inclusion quantities is still in force, i.e. the quantities $\pi_i$ and $\beta_i$ in (A1.5) are regarded to be known. ▒

When Assumption A1.2 is in force, we set

$$b_r = \#(G_r) \quad (=\text{the number of items in the group } G_r). \tag{A1.11}$$

Against the above background and Lemmas A2.4 and A2.5 we can formulate the following approximation result.

RESULT A1.2: Under Assumption A1.2, the random variables

$$Z_r(\underline{y}) = (1/b_r)\cdot\sum_{i\epsilon G_r}(y_i/\beta_i) \ , \ r=1,2,\ldots,R, \tag{A1.12}$$

can, with good approximation, be viewed as independent random variables which all have expected value $\theta(\underline{y})$.

Hence, by applying Lemma A1.1 to the Z-variables in (A1.12) we arrive at the following result.

RESULT A1.3: Let Assumption A1.2 be in force and let the $Z(\underline{y})$-variables be as in (A1.12) and let $S^2$ be defined in accordance with (A1.2).

Then $S^2/R$ is an approximately unbiased estimator of the variance $V[\hat{\theta}(\underline{y})^*]$, where

$$\hat{\theta}(\underline{y})^* = (1/R)\cdot\sum_{r=1}^{R} Z_r(\underline{y}). \tag{A1.13}$$

Next we formulate a prototype for the estimation procedure, including estimation of sampling errors, which we shall for use in the RZMS contexts. For easy reference we give the procedure a name and we call it the UC-procedure (where UC emanates from "ultimate cluster").

THE UC-PROCEDURE (for estimation of a population total together with an estimate of the sampling error): Consider a probability sample with fixed size n, for which the inclusion proportionates $\underline{\beta}=(\beta_1,\beta_2,\ldots,\beta_N)$ are known.

Assumption A1.2 is in force with grouping $\underline{G}=(G_1,G_2,..,G_R)$, and group sizes $b_r$.

(i) Estimate the population total $\theta(\underline{y})$ by

$$\hat{\theta}(\underline{y})=(1/n)\cdot\sum_{v=1}^{n}(y_{T_v}/\beta_{T_v}). \qquad (A1.14)$$

(ii) Estimate the estimator variance $V[\hat{\theta}(\underline{y})]$ by

$$\hat{V}[\hat{\theta}(\underline{y})] = S^2(\underline{y};\underline{\beta};\underline{G})/R, \qquad (A1.15)$$

where

$$S^2(\underline{y};\underline{\beta};\underline{G}) =(1/(R-1))\cdot\sum_{r=1}^{R}(Z_r(\underline{y})-\hat{\theta}(\underline{y})^*)^2, \qquad (A1.16)$$

and the $Z_r$:s are defined in (A1.12) and $\theta(\underline{y})^*$ in (A1.13).

Remark A1.6: Note that the above UC-procedure contains the follwing "inconsistency". It states that $\hat{\theta}(\underline{y})$ (see (A1.14)) should be used to give the point estimate for $\theta(\underline{y})$, while the variance of this estimator should be estimated by an estimate of the variance for the estimator $\hat{\theta}(\underline{y})^*$ (see (A1.13)), which also is an unbiased estimator of $\theta(\underline{y})$ but not exactly the same as $\hat{\theta}(\underline{y})$.

The reasons for this inconsistency are mainly practical. The estimate $\hat{\theta}(\underline{y})$ is simpler to compute than $\hat{\theta}(\underline{y})^*$ to the effect that it does not require any information on the grouping, while $\hat{\theta}(\underline{y})^*$ is the estimator on which the ultimate clusters method can be applied. ▒

Remark A1.7: From what has been said so far, it should be clear that the UC-procedure gives a variance estimate which is more or less biased. Bias sources are the assumption about approximate independence (neglection of finite population correction) and the inconsistency mentioned in the previous remark. However, these bias sources are in most cases negligible, and in the sequel we presume that so is the case. ▒

A1.2. Extensions of the UC-procedure.

First we extend the UC-procedure to situations where the y-values are not observed exactly, only estimated. A case which leads to this type of situation is as follows. Assume that the units in the population, i.e. the i:s, in fact are clusters of sub-units. Let $y_i$ denote the total of sub-unit y-values over cluster i. Assume that the sampled clusters are not totally inspected, only samples of sub-units from them are observed. Then, the observations do not lead to exact information on $y_i$, only to an estimate of it. Introduce the following assumption.

> The sub-samples from the clusters are drawn independently of each other and independently of what happened in previous sampling stages. (A1.17)

For each sampled i, $Y_i$ is an unbiased estimator of $y_i$ based on the sub-sample observations. (A1.18)

Then, we have the following variation of Result A1.2, and again Lemmas A2.4 and A2.5 provide background.

RESULT A1.4 Under Assumption A1.2,(A1.17) and (A1.18), the random variables

$$Z_r(\underline{y}) = (1/b_r)\cdot\Sigma(Y_i/\beta_i),\ r=1,2,\ldots,R, \qquad (A1.19)$$
$$i\epsilon G_r$$

can be viewed, with good approximation, as independent random variables which all have expected value $\theta(\underline{y})$.

As a consequence of the above result; the UC-procedure can be applied if $y_i$ is changed to $Y_i$, provided that the assumptions in Result A1.4 are met.

We conclude by presenting an extension of the UC-procedure to situations where the "basic sample" is stratified. First we formulate various assumptions and notation.

We presume that the population U is partitioned into H (disjoint and exhaustive) sampling strata, of sizes $N_1,N_2,\ldots,N_H$,

$$U = A_1 \quad A_2 \quad \ldots \quad A_H, \qquad (A1.20)$$

and that random samples are drawn from the sampling strata.

(i) The samples from the different sampling strata are independent of each other. (A1.21)

(ii) The samples $\underline{I}_h=(I_{h1},I_{h2},\ldots,I_{hN_h})$, h=1,2,...,H from the different sampling strata are drawn without replacement and with prescribed sample sizes; $a_1,a_2,\ldots,a_H$. (A1.22)

(iii) The inclusion proportionates for the sample from sampling stratum h are denoted $\underline{\beta}_h=(\beta_{h1},\beta_{h2},\ldots,\beta_{hN_h})$. (A1.23)

(iv) The Assumption A1.2 holds for each sample from the different sampling strata. For stratum h, the grouping is denoted by $\underline{G}_h=(G_{h1},G_{h2},\ldots,G_{hR_h})$ and the group sizes by $\underline{b}_h=(b_{h1},b_{h2},\ldots,b_{hR_h})$. (A1.24)

(v) The population variable of interest is specified by

$$\underline{y} = \{\underline{y}_h;\ h=1,2,\ldots,H\}, \qquad (A1.25)$$

where

$$\underline{y}_h=\{y_{h1},y_{h2},\ldots,y_{hN_h}\}. \qquad (A1.26)$$

The corresponding sampling stratum totals are denoted

$$\theta_h(\underline{y}) = \sum_{i=1}^{N_h} y_{hi}. \qquad (A1.27)$$

For the population total $\theta(\underline{y})$ we then have,

$$\theta(\underline{y}) = \sum_{h=1}^{H} \theta_h(\underline{y}). \qquad \text{(A1.28)}$$

(vi) For sampled items i, the variables

$$\{Y_{hi};\ i=1,2,\ldots,a_h,\ h=1,2,\ldots,H\}$$

yield unbiased estimation of the corresponding y-values. (A1.29)

Under the above assumptions, the UC-procedure can be applied in each sampling stratum to estimate stratum totals and variances of the estimators of the stratum totals. By combining this with the formula (A1.28) and the assumption (A1.21) that the samples from the different strata are independent we arrive at the following extension of the UC-procedure. Some of the details are left to the reader.

THE UC-PROCEDURE IN STRATIFIED VERSION: Let assumptions and notation be as in (i)-(vi) above.

(i) Estimate the population total $\theta(\underline{y})$ by

$$\hat{\theta}(\underline{y}) = \sum_{h=1}^{H} (1/a_h) \cdot \sum_{i=1}^{a_h} (Y_{hi}/\beta_{hi}). \qquad \text{(A1.30)}$$

(ii) Estimate the variance of the estimator by

$$\hat{V}[\hat{\theta}(\underline{y})] = \sum_{h=1}^{H} S^2(\underline{y}_h;\underline{\beta}_h;\underline{G}_h)/R_h, \qquad \text{(A1.31)}$$

where

$$S^2(\underline{y}_h;\underline{\beta}_h;\underline{G}_h) = (1/(R_h-1)) \cdot \sum_{r=1}^{R_h} (Z_{hr}(\underline{y}) - \hat{\theta}_h(\underline{y}))^2. \qquad \text{(A1.32)}$$

and

$$Z_{hr}(\underline{y}) = (1/b_{hr}) \cdot \sum_{i \in G_{hr}} (Y_{hi}/\beta_{hi}), \qquad \text{(A1.33)}$$

and

$$\hat{\theta}_h(\underline{y})^* = (1/R_h) \cdot \sum_{r=1}^{R_h} Z_{hr}(\underline{y}). \qquad \text{(A1.34)}$$

## APPENDIX 2. *SOME NOTIONS AND RESULTS CONCERNING GENERAL PROBABILITY SAMPLES.*

Our main aim in this appendix is to present results on the effects of certain general operations on probability samples. In particular we shall consider random thinning (or sub-sampling), random extension and random permutation of samples.

Let $U=(1,2,\ldots,N)$ be a finite population. In the first round we chose to define a general probability sample (random sample) from U, drawn without replacement as a collection $\underline{I}=(I_1,I_2,\ldots,I_N)$ of random indicators, i.e. random variables which only take the values 0 or 1, together with the interpretation

$$I_i = \begin{cases} 1 & \text{if item } i \text{ is sampled,} \\ 0 & \text{if item } i \text{ is not sampled.} \end{cases} \qquad \text{(A2.1)}$$

We refer to $I_1,I_2,\ldots,I_N$ as the sample inclusion indicators. The corresponding inclusion probabilities are

$$\pi_i = P(I_i=1) = E[I_i],\ i=1,2,\ldots,N. \qquad \text{(A2.2)}$$

In the sequel we shall mostly assume that

$$\text{the sample has prescribed (or fixed) size} = n, \qquad \text{(A2.3)}$$

i.e. that

$$I_1 + I_2 + \ldots + I_N = n. \qquad \text{(A2.4)}$$

Under (A2.3) we have,

$$\pi_1 + \pi_2 + \ldots + \pi_N = n. \qquad \text{(A2.5)}$$

For samples with prescribed sample size, it is sometimes more convenient to work with re-scaled inclusion probabilities, and we introduce the following inclusion proportionates,

$$\beta_i = \pi_i/n,\ i=1,2,\ldots,N. \qquad \text{(A2.6)}$$

From (A2.5) and (A2.6) we see that the $\beta$:s are standardized in the sense that,

$$\sum_{i=1}^{N} \beta_i = 1. \qquad \text{(A2.7)}$$

We now turn to random thinning of samples, also referred to as sub-sampling. Let $\underline{I}=(I_1,I_2,\ldots,I_N)$ be a random sample from U, and let $\underline{J}=(J_1,J_2,\ldots,J_N)$ be random indicators. Define $\underline{L}=(L_1,L_2,\ldots,L_N)$ by

$$L_i = I_i \cdot J_i,\ i=1,2,\ldots,N. \qquad \text{(A2.8)}$$

It is readily seen that the $L_i$:s are random indicators, and hence that $\underline{L}$ can be interpreted as a random sample from U. This sample is called the $\underline{J}$-thinning of $\underline{I}$, or the sub-sample of $\underline{I}$ given by $\underline{J}$.

The inclusion probabilities for the sub-sample $\underline{L}$ are given by

$$\pi(\underline{L})_i = \pi_i \cdot P(J_i=1 | I_i=1), \quad i=1,2,\ldots,N. \tag{A2.9}$$

Assume that $\underline{I}$ and $\underline{J}$ are related so that for some $0 \leq n' \leq n$ we have,

$$P(J_i=1 | I_i=1) = n'/n. \tag{A2.10}$$

Then (A2.9) yields

$$\pi(\underline{L})_i = \pi_i \cdot (n'/n), \quad i=1,2,\ldots,N. \tag{A2.11}$$

When (A2.3) is met and $n'$ is an integer, (A2.10) is satisfied by the following procedure (which gives an implicit definition of $\underline{J}$);

> From the n items in the $\underline{I}$-sample, select independently of "everything else", a simple random sample of size $n'$. (A2.12)

Under (A2.12), the sub-sample $\underline{L}$ has the prescribed sample size $n'$, and hence its inclusion proportionates are well-defined. From (A2.11) it is readily seen that the sub-sample has the same inclusion proportionates as the original $\underline{I}$-sample. We summarize the reasoning in the following lemma.

> LEMMA A2.1: Let $\underline{I}$ be a general probability sample with prescribed size n, and let $\underline{I}$ be a sub-sample according as in (A2.12). Then the sub-sample, which has the prescribed size $n'$, has the same inclusion proportionates as the original sample $\underline{I}$.

Next we turn to extensions of random samples. Let $\underline{I}$ and $\underline{J}$ be random samples from the population U. The combined sample consists of the items which belong to at least one of $\underline{I}$ and $\underline{J}$. More formally, the combined sample $\underline{L}$ is given by the inclusion indicators

$$L_i = \min(1, I_i+J_i), \quad i=1,2,\ldots,N. \tag{A2.13}$$

Note that even if both $\underline{I}$ and $\underline{J}$ have fixed sizes, $\underline{L}$ does not have fixed size in general because the number of common objects in the two samples is in general a random number.

If the samples $\underline{I}$ and $\underline{J}$ are independent, the inclusion probabilities for the combined sample are,

$$\pi(\underline{L})_i = \pi(\underline{I})_i + \pi(\underline{J})_i - P(I_i=J_i=1), \quad i=1,2,\ldots,N. \tag{A2.14}$$

At this stage we introduce, for future use, two conditions on a random sample with prescribed size n, conditions which are somewhat vague in nature,

The sampling fraction n/N is "fairly small". (A2.15)

The inclusion proportionates for the items in the population are of "equal order of magnitude", i.e. the quantity $\max_{i\in U} \beta_i / \min_{i\in U} \beta_i$ is "moderate". (A2.16)

If $\underline{I}$ and $\underline{J}$ both satisfy (A2.15) and (A2.16) and if the samples are independent, the last term in (A2.14) is negligible compared with the others. Hence we have with good approximation,

$$\pi(\underline{L})_i \approx \pi(\underline{I})_i + \pi(\underline{J})_i, \quad i=1,2,\ldots,N. \tag{A2.17}$$

Furthermore;

The size of the combination of the samples $\underline{I}$ and $\underline{J}$ will be close to size($\underline{I}$)+size($\underline{J}$), (A2.18)

Against this background we formulate the following result.

LEMMA A2.2: Let $\underline{I}$ be a random sample with prescribed size and with inclusion proportionates $\underline{\beta}=(\beta_1,\beta_2,\ldots,\beta_N)$. Extend $\underline{I}$ by combining it with an independently drawn sample with prescribed size and with the same inclusion proportionates $\underline{\beta}$.

If (A2.15) and (A2.16) are met, the extended sample can be regarded as a fixed size (with size = obtained size) random sample with inclusion proportionates $\underline{\beta}$.

A more "naive" approach to the notion of a random sample $\underline{I}$ is to say that it is the collection of items which are obtained in the sampling process, i.e. to view the sample as the set

$$\underline{T} = \{i;\ I_i=1\}, \tag{A2.19}$$

and we shall do so in the following.

Under (A2.3), the set $\underline{T}$ in (A2.19) contains exactly n objects, and we write it,

$$\underline{T} = (T_1,T_2,\ldots,T_n). \tag{A2.20}$$

In the representation (A2.20) we meet the following problem; Which ordering principle is used in (A2.20) when the sampled items are labelled by 1,2,...,n? Unless otherwise stated we assume that the following ordering principle is employed.

The sampled items are labelled by a totally random permutation of 1,2,...,n, which is independent of "everything else". (A2.21)

Let $\underline{y}=(y_1,y_2,\ldots,y_N)$ be a variable on the population U. Set

$$Y_v = y_{T_v} \ , \ v=1,2,\ldots,n, \tag{A2.22}$$

i.e. $Y_1,Y_2,\ldots,Y_n$ are the sampled y-values in random order. A variable $Y_v$ can be viewed as a "randomly chosen" sampled y-observation.

LEMMA A2.3: If the sample has fixed prescribed size n and (A2.21) is used, we have,

$$E[Y_1] = E[Y_2] = \ldots = E[Y_n] = (1/n)\cdot\sum_{i=1}^{N} y_i\cdot\pi_i. \tag{A2.23}$$

Proof: As a consequence of the random ordering in (A2.21), the Y-variables are so called exchangeable random variables. As such they have the same (marginal) distributions, and in particular the same expected values. Hence we have proved (A2.23) but for the computation of the common value for the expectations. To compute that, we first note the following straightforward relation,

$$\sum_{v=1}^{n} Y_v = \sum_{i=1}^{N} y_i\cdot I_i \ . \tag{A2.24}$$

By taking expectation in (A2.24) we get, in view of (A2.2),

$$\sum_{v=1}^{n} E[Y_v] = \sum_{i=1}^{N} y_i\cdot\pi_i \ . \tag{A2.25}$$

By remembering that the Y:s have the same expected values, (A2.20) readily leads to (A2.23) and the lemma is proved. ▒

The following consequence of (A2.23) is obtained by exchanging y to $y/\beta$ and by recalling (A2.6).

LEMMA A2.4: Consider a general probability sample $\underline{I}$ with prescribed size n and with inclusion proportionates $(\beta_1,\beta_2,\ldots,\beta_N)$. Let $\underline{T}$ be defined by (A2.19)-(A2.21). Set

$$Z_v = y_{T_v}/\beta_{T_v} \ , \ v=1,2,\ldots,N. \tag{A2.26}$$

Then,

$$E[Z_v] = \Theta(\underline{y}) = \sum_{i=1}^{N} y_i, \ v=1,2,\ldots,n. \tag{A2.27}$$

Under the additional assumptions that (A2.15) and (A2.16) are in force, we can add to the claims in the above lemma. Then, the Z-variables can, with good approximation, be regarded as

independent random variables. The full claim is formulated below.

LEMMA A2.5: Consider a general probability sample $\underline{I}$ with prescribed size n and with inclusion proportionates $(\beta_1,\beta_2,\ldots,\beta_N)$. Let $\underline{T}$ be defined by (A2.19)-(A2.21). Set

$$Z_v = y_{T_v}/\beta_{T_v} \ , \ v=1,2,\ldots,N. \tag{A2.28}$$

If also (A2.15) and (A2.16) are met, we have;

(i) $Z_1,Z_2,\ldots,Z_n$ can, with good approximation, be regarded as independent random variables.

(ii) $Z_1,Z_2,\ldots,Z_n$ all have the same expected value,

$$E[Z_v] = \Theta(\underline{y}) = \sum_{i=1}^{N} y_i, \ v=1,2,\ldots,n. \tag{A2.29}$$

## APPENDIX 3. ON RATIO VARIABLES.

Estimators of means and proportions are random variables of ratio type, i.e. of the type Z/W where Z and W are joint random variables. As is well known, ratio variables are a bit unpleasant in the sense that there are no simple exact formulas for their means and variances. The usual way to circumvent this obstacle is to apply so called Taylor approximation. The topic is certainly well-known and we shall not give proofs, just write down some well-known formulas as memory aids for the reader who wants to check the derivations of estimators of the sampling errors for the estimators.

TAYLOR APPROXIMATION OF A RATIO VARIABLE: Let Z and W be joint random variables with expected values $\Omega_Z$ and $\Omega_W$ respectively. Assume that the standard deviations of Z and W are "small" compared with $\Omega_W$. Then the random variable Z/W is well approximated as follows,

$$Z/W \approx \Omega_Z/\Omega_W + (1/\Omega_W)\cdot[(Z-\Omega_Z)-(\Omega_Z/\Omega_W)\cdot(W-\Omega_W)]. \quad (A3.1)$$

By taking expectation and variance in the formula (A3.1) we are led to the following approximation formulas for the mean and variance of the ratio variable Z/W,

$$E[Z/W] \approx \Omega_Z/\Omega_W, \quad (A3.2)$$

$$V[Z/W] \approx V[Z-(\Omega_Z/\Omega_W)\cdot W]/(\Omega_W)^2. \quad (A3.3)$$

The following result is only a particular case of formula (A3.3), in case when Z and W are summation variables.

LEMMA A3.1: Let $Z_1,W_1,Z_2,W_2,\ldots$ be joint random variables. Set, for an arbitrary summation set Q,

$$Z = \sum_{q\in Q} Z_q \quad \text{and} \quad W = \sum_{q\in Q} W_q. \quad (A3.4)$$

As above, let $\Omega_Z$ and $\Omega_W$ denote the expected values of Z and W. Then we have,

$$V[Z/W] \approx V[\Sigma_{q\in Q}(Z_q-(\Omega_Z/\Omega_W)\cdot W_q)]/(\Omega_W)^2. \quad (A3.5)$$

When combining the above result with that in Lemma A1.1, one usually meets a special type of computation problems, and below we list some formulas which are useful in untangling that type of computation problems.

For paired observations $(X_1,Y_1),(X_2,Y_2),\ldots(X_R,Y_R)$, sample means, sample variances and sample covariances are defined and denoted along the following lines,

$$\bar{X} = (1/R) \cdot \sum_{r=1}^{R} X_r \ . \qquad \text{(A3.6)}$$

$$S^2(\underline{X}) = 1/(R-1) \cdot \sum_{r=1}^{R} (X_r - \bar{X})^2 \ . \qquad \text{(A3.7)}$$

$$C(\underline{X},\underline{Y}) = 1/(R-1) \cdot \sum_{r=1}^{R} (X_r - \bar{X}) \cdot (Y_r - \bar{Y}) \ . \qquad \text{(A3.8)}$$

Then the following well-known formula holds,

$$S^2(a \cdot \underline{X} + b \cdot \underline{Y}) = a^2 \cdot S^2(\underline{X}) + b^2 \cdot S^2(\underline{Y}) + 2ab \cdot C(\underline{X},\underline{Y}) \ . \qquad \text{(A3.9)}$$

Remark A3.1: For numerical computations the following formulas are usually more convenient than (A3.7) and (A3.8),

$$S^2(\underline{X}) = (1/(R-1)) \cdot [\sum_{r=1}^{R} X_r^2 - (\sum_{r=1}^{R} X_r)^2/R] \ . \qquad \text{(A3.10)}$$

$$C(\underline{X},\underline{Y}) = (1/(R-1)) \cdot [\sum_{r=1}^{R} X_r \cdot Y_r - (\sum_{r=1}^{R} X_r) \cdot (\sum_{r=1}^{R} Y_r)/R] \ . \qquad \text{(A3.11)}$$

▒

REFERENCES

ARVIDSSON, A.(May 1987). On the 1987 Intercensal Demographic Survey.

CSO (December 1986). Document on development of a survey design, a master sample, analysis and quality control, and questionnaires in Zimbabwe National Household Survey Capability programme (ZNHSCP).

CSO (March 1987). Intercensal Demographic Survey. Working document. (Includes a tabulation plan.)

CSO (March 1987). Intercensal Demographic Survey, Round 1. Interview Manual.

CSO (October 1987). Zimbabwe National Household Survey Capability Programme.

CSO (January 1988) Intercensal Demographic Survey, Editing and Coding Manual.

CSO (February 1988). Intercensal Demographic Survey, Second Round. (Working document)

CSO (May 1988). Intercensal Demographic Survey, Second Round. Interviewer's Manual.

JOHANSSON, L.(June 1988) On the 1987/88 Intercensal Demographic Survey, Round Two.

LAGERLÖF, B.(January 1988). Development of system design for National Household Surveys.

SCOTT, C.(May 1987). Report on a mission to Zimbabwe National Household Survey Capability Programme.

ZIMSTAT:5, Part 1 (1989). Sampling and estimation in the Zimbabwe Household Survey Programme, II.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports, Statistics Sweden, are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown (beige) covers).

Reports published earlier during 1989 are:

1989:1 (grön) Går det att mäta produktivitetsutvecklingen för SCB? **(Rune Sandström)**

1989:2 (grön) Slutrapporter från U-avdelningens översyn av HINK och KPI **(flera författare)**

1989:3 (grön) A Cohort Model for Analyzing and Projecting Fertility by Birth Order **(Sten Martinelle)**

1989:4 (beige) **Abstracts** I - Sammanfattningar av metodrapporter från SCB

1989:5 (gul) On the use of Semantic Models for specifying Information Needs **(Erik Malmborg)**

1989:6 (grön) On Testing for Symmetry in Business Cycles **(Anders Westlund och Sven Öhlén)**

1989:7 (grön) Design and quality of the Swedish Family Expenditure Survey **(Håkan L Lindström, Hans Lindkvist och Hans Näsholm)**

1989:8 (grön) Om utnyttjande av urvalsdesignen vid regressionsanalys av surveydata **(Lennart Nordberg)**

1989:9 (grön) Variations in the Age-Pattern of Fertility in Sweden Around 1986 **(Michael Hartmann)**

1989:10 (grön) An Application of Generalized Precision Functions in the 1985 Swedish Family Expenditure Survey **(Håkan L Lindström and Peter Lundquist)**

1989:11 (gul) Statistics production in the 90's - decentralization without chaos **(Bo Sundgren)**

1989:12 (grön) Översyn av urvalen i de objektiva skördeuppskattningarna **(Erling Andersson)**

1989:13 (gul) ADBs roll i statistiken **(Bo Sundgren)**

1989:14 (grön) Krisgruppsarbetet och räknarexperimentet i HUT 88 **(Håkan L. Lindström)**

Kvarvarande BEIGE och GRÖNA exemplar av ovanstående promemorior kan rekvireras från Elisabet Klingberg, U/STM, SCB, 115 81 Stockholm, eller per telefon 08-783 41 78.

Dito GULA exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 Stockholm, eller per telefon 08-783 41 47.