Variance estimation for systematic pps-sampling

Bengt Rosén

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1991:15

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 1991:15. Variance estimation for systematic pps-sampling / Bengt Rosén.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1991-X101OP9115

Variance estimation for systematic pps-sampling

Bengt Rosén

R&D Report
Statistics Sweden
Research - Methods - Development
1991:15

Från trycket	September 1991
Producent	Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare	Åke Lönnqvist
Förfrågningar	Bengt Rosén, tel. 08-783 44 90

ISSN 0283-8680
Garnisonstryckeriet, Stockholm

Variance estimation for systematic pps-sampling

by

Bengt Rosén
Statistics Sweden

Abstract: We present a novel approach to computation of confidence intervals based on linear statistics, notably Horvitz-Thompson estimators, when observations come from a systematic pps-sample. The new method is believed to work well also in situations with large sampling fractions. Like already existing methods the new one is based on approximation arguments, and the crucial idea is to approximate systematic pps-sampling (under total ignorance of frame order) by successive pps-sampling. We report on some simulation findings, which support the conjecture that the proposed variance estimation method works quite satisfactorily over a very wide range of sampling fractions. Moreover, it is numerically simple. Accordingly we want to recommend it for practical use, in particular in situation with systematic pps-samples with high sampling rates.

Contents

1. Introduction and outline of the paper

The main task of this paper is to present a novel approach to computation of confidence intervals from observations on a systematic sample. Systematic sampling procedures are widely employed in survey practice, and their popularity depend generally on the fact that they provide means for exploiting auxiliary information to improve estimation precision (relative to e.g. simple random sampling). Systematic sampling procedures enable two different kinds of variance reducing forces to be set in action, namely "variation of selection probabilities", accomplished by **systematic pps-sampling**, respectively "appropriate ordering of the sampling frame". The latter method is especially relevant in connection with ordinary (equal selection probabilities) systematic sampling, henceforth referred to as **simple systematic sampling**. The two forces can also be combined, thereby hopefully leading to added positive effects. Below we give a brief discussion of merits and drawbacks of systematic sampling. Extensive overviews can be found in e.g. the papers by Bellhouse (1988) and Murthy & Rao (1988).

Varying selection probabilities is employed in list sampling situations (i.e. study units are sampled "directly") as well as in multistage sampling. In the former type of situation, considerable reduction of estimator variances can be achieved by sampling so that inclusion probabilities become (fairly) proportional to the values of an auxiliary variable, usually called the **size measure variable**, which in turn is known (or believed) to be fairly proportional to the values of the study variable. In multi-stage sampling situations, varying probabilities procedures are usually employed in the initial selection stages for reasons which rather are opposite to the previous ones; To render fairly equal inclusion probabilities to the (ultimate) study units. However, whatever be the rationale for desiring varying inclusion probabilities, one meets the problem of advising a sampling scheme which leads to desired inclusion probabilities. A chief reason for the popularity of systematic pps-sampling is that it offers an operationally very simple solution to this problem. As regards "frame ordering", it is well known that substantial reductions of estimator variances can be achieved if the sampling frame can be arranged so that the values of the study variable exhibit a "smooth trend" (preferably linear) in their frame order. The accomplishment of a beneficial frame ordering, requires (good) auxiliary information.

The method of systematic sampling, be it pps or simple, is associated with certain difficulties, though, as regards estimation of estimator variances, for subsequent use in confidence intervals, and the difficulties become particularly pronounced when the sampling fraction is "large" (say 25% or more). The pertinent feature of the proposed new method is that it is believed to work satisfactorily over a wide range of sampling fractions, notably large ones. Even if very high sampling fractions are fairly unusual in sampling practice, they do occur. In fact, the present investigation was initiated by confidence interval problems in Statistics Sweden's Labour Force Barometer, which surveys some 100 different populations (many of them quite small) with systematic pps-samples, of which quite many have sampling rates of 50% or more. Confidence intervals based on systematic samples is, however, certainly not new problem area, and a wealth of methods have been suggested. All of them are of a more or less approximate nature, and this holds also for the method to be presented. An extensive overview of variance estimation for systematic samples can be found in the book by Wolter (1985), Chapter 7.

Next we give a brief indication of the central idea in the new approach. To provide background, we first note that the variance estimation methods which are most widely used in present practice, are based on the following heuristic approximation argument;

> A systematic pps-sample behaves very much like a "corresponding" pps-sample drawn with replacement. (1.1)

For with-replacement sampling, there is a standard solution to the variance estimation problem, and (1.1) provides a link for using this solution also for systematic pps-sampling. However, since systematic pps-sampling is a without-replacement procedure, the solution provided by (1.1) is only approximative. The approximation works quite well, though, as long as the sampling fraction is negligible (and thereby, the difference between with- and without-replacement is little), but the larger the sampling fraction becomes, the cruder the approximation is. A natural thought is that "remedy" for cases with non-negligible sampling fractions could be achieved by employing some finite population correction (fpc) to the estimator, (the simplest fpc-candidate being the "traditional" one, (1-n/N)). In fact, correction with fpc's can make the estimators work satisfactorily over a wider range of sampling fractions, as is discussed in more detail in e.g. Wolter (1985). A common feature of the suggested fpc's is, however, that they all are considerably ad hoc. The approach to be studied in this paper is based on an approximation which differs from (1.1) to the effect that systematic pps-sampling is, already from the beginning, approximated by a without-replacement procedure, namely so-called **successive pps-sampling**. The basic approximation argument, still heuristic though, runs as follows;

> A systematic pps-sample behaves very much like a "corresponding" successive pps-sample. (1.2)

We conclude this section with an outline of the rest of the paper. When systematic sampling is employed, the variance of a sample statistic depends crucially on the order in the sampling frame. Therefore, when investigating estimator variances one inevitably has to handle "frame order" as a formal ingredient. In practice the frame order is usually not generated by an "objective" random mechanism, but typically by the surveyor's efforts (for which there are good reasons) to impose an order, which renders as much "smooth trend" as possible to the (most important) study variables. As a consequence of this, the formal modelling of frame order will concern codification of "opinions" on the frame order, which more or less inevitably leads to probability models with "subjective" ingredients. This pronounced subjectivity aspect distinguishes inference for systematic sampling from that for the majority of other sampling procedures. In Sections 2 and 3 we formulate and discuss a precise and coherent formal frame-work for finite population inference, with special regard to aspects that are pertinent to systematic sampling. We also review some well-known results, for subsequent use in the theoretical justifications and the evaluations of the new approach. The main contents of these sections is in essence not new but, to the best of our understanding, from a more technical point of view they contain some novelties.

As can be surmised from (1.2), the notion of successive pps-sampling will be instrumental, and this sampling procedure is specified and discussed in Section 4. In particular we derive a variance estimation procedures for it, based on asymptotic results by the author (Rosén, 1971a,b). This method is of interest in its own right, maybe mainly from a theoretical point of view, though, since successive pps-sampling is fairly seldom used in survey sampling practice. Anyway, the variance estimation method for successive pps-sampling provides the fundamental instrument for implementing the approximation argument in (1.2), which then leads to the new variance estimation method for systematic pps-sampling. The details of the implementation are given in Section 5, which is of a fairly theoretical nature. More practically oriented consequences are formulated and discussed in Section 6.

The proposed method is based on theoretical considerations as well as on its intuitive appeal. The theoretical reasoning contains various approximation arguments, though, notably asymptotic results, and intuition is always unreliable. Therefore, it is not possible to make general, precise statements about the accuracy of the method in practical applications. To provide some insight into the matter, simulation studies have been carried out, which are reported on in Section 7. In that section we also carry out some discussion of the validity of normal distribution approximations, and draw the main conclusion which runs as follows. The numerical findings support the conjecture that the proposed variance estimation method works quite satisfactorily over a very wide range of sampling fractions and, moreover, the method is computationally very simple. Towards this background we want to recommend it for practical use, notably in situations where fairly high sampling fractions prevail.

2. A general frame-work for finite population inference

As stated in Section 1, inference from systematic samples is associated with the special complication that, at least in most cases, it contains a crucial "subjective" ingredient. To lay ground for pregnant technical discussions of this aspect, we settle in this section some general notions, terminology and notation on inference from finite populations, with special regard to notions of relevance for inference from systematic samples.

2.1 Populations, variables and characteristics

U denotes a (finite) **population** with N distinguishable **units**, which are assumed to be labelled by the first N integers; $U=\{1,2,...,N\}$. A **variable** $\mathbf{x}=(x_1,x_2,..,x_N)$ on U associates a value to each unit in U, x_i denoting the value associated with unit i. If not stated otherwise, a variable is assumed to take numerical values. Arithmetic operators on variables on the same population should be interpreted as unit-wise operations; For variables $\mathbf{x}=(x_1,x_2,..,x_N)$, $\mathbf{y}=(y_1, y_2,...,y_N)$ and numbers α and β, $\alpha\cdot \mathbf{x}+\beta\cdot \mathbf{y}$ takes the value $\alpha\cdot x_i+\beta\cdot y_i$ for unit i, while $\mathbf{x}\cdot \mathbf{y}$ and $\mathbf{x}/\mathbf{y}$ take the values $x_i\cdot y_i$ respectively x_i /y_i. **1** denotes the variable with value 1 for each unit in U. A **domain**, in general denoted by D, is a subset of the units in the population. The **domain** D **indicator** (variable) $\mathbf{1}_D$ is; $\mathbf{1}_D(i)=1$ if unit i belongs to domain D and $\mathbf{1}_D(i)=0$ otherwise.

For a variable $\mathbf{x}=(x_1,x_2,...x_N)$, an **x-characteristic** $\chi(\mathbf{x})$ is defined by $\chi(\mathbf{x})=\phi(x_1,x_2,...x_N)$, where ϕ is a symmetric function (i.e. invariant under permutation of the arguments). Analogously, an (x,y,..)-**characteristic**, $\chi(\mathbf{x},\mathbf{y},...)$ is a function of $x_1,x_2,...x_N$, $y_1,y_2,...y_N$,.., which is invariant under permutation of the x's, y's, .. . The population/domain characteristics which are specified in (2.1) and (2.2) below, are of particular interest in survey contexts; the **population x-total**, $\tau(\mathbf{x})$; the **population x-mean**, $\mu(\mathbf{x})$; the **domain** D **size**, $\kappa(D)$; the **domain** D **x-total**, $\tau(\mathbf{x};D)$; the **domain** D **x-mean**, $\mu(\mathbf{x};D)$. In (2.1) and (2.2) we also emphasize that population total can be viewed as the central type of characteristic.

$$\tau(x)=\sum_{i\in U} x_i,\quad \mu(x)=\frac{\tau(x)}{N}. \tag{2.1}$$

$$\tau(x;D)=\sum_{i\in D} x_i=\tau(x\cdot\mathbf{1}_D),\quad \kappa(D)=\tau(\mathbf{1};D)=\tau(\mathbf{1}_D),\quad \mu(x;D)=\frac{\tau(x;D)}{\kappa(D)}=\frac{\tau(x\cdot\mathbf{1}_D)}{\tau(\mathbf{1}_D)}. \tag{2.2}$$

Henceforth, terms as "x-characteristic" and "population characteristic" are meant to cover "domain characteristic" as well as "(entire) population characteristic". As is readily checked, all characteristics in (2.1) and (2.2) can, for appropriate choices of **x** and **y**, be viewed as special cases of the following general type of (x,y)-characteristic, $\mu(\mathbf{x},\mathbf{y})$, called a **ratio of totals** characteristic;

$$\mu(x,y)=\tau(x)/\tau(y). \tag{2.3}$$

2.2 Models for inference from sample observations to population characteristics

In the following we introduce a frame-work for inference from sample observations to population characteristics. The frame-work could be made more general, but for the sake of brevity we make simplifying assumptions whenever it is in accordance with the main aim, which is to formulate a frame-work that is general enough to allow pregnant discussion of inference aspects for systematic samples. The following question is taken as the point of departure.

For a specific **study variable**, call it **x**, a sample of values, an **x-sample**, is (to be) collected. On basis of the **x**-sample, one wants to make statements about one or more **x**-characteristic. What procedures are rational and efficient for making such statements? (2.4)

When making statements about population characteristics on the basis of a sample of values, one practically always confronts a situation of making statements "under uncertainty", where the crucial uncertainty emanates from the fact that one lacks a unique relationship between the sampled **x**-values and the **x**-values in the population (and an **x**-characteristic is determined by the latter). The common way to bridge this kind of "information gap", is to introduce a probability model which relates the mentioned two sets of **x**-values. We pursue this route in the following basic notion of "univariate inference model". We comment on multivariate cases later on.

A **univariate inference model for sample values** for a specific study variable, call it **x**, on the population U=(1,2,...,N), is a triple $(\Omega,\{\$(\mathbf{x});\ \mathbf{x}\in\Omega\},\{P_{\mathbf{x}};\ \mathbf{x}\in\Omega\})$ as specified below.

(i) Ω is a collection of ordered N-tuples $\mathbf{x}=(x_1,x_2,...,x_N)$ where $x_1,x_2,...,x_N$ are presumed to be reals. Hence Ω is some subset of R^N (= the space of ordered N-tuples of reals).

(ii) For each $\mathbf{x}=(x_1,x_2,\ ...x_N)$ in Ω, $\$(\mathbf{x})$ is a collection of tuples whose components are selected among the values offered by $x_1,x_2,...x_N$. The tuples in $\$(\mathbf{x})$ may have differing numbers of components.

(iii) For each **x** in Ω, $P_{\mathbf{x}}$ is a probability distribution on $\$(\mathbf{x})$.

The concrete interpretations of the objects in an inference model are as follows. Ω specifies the range for x:s that are regarded to be **apriori possible** (or maybe rather that x:s outside Ω are regarded as apriori impossible). Hence, in a practical situation the choice of Ω should reflect the surveyor's (prior) knowledge of the population variable **x**.

$\$(\mathbf{x})$ specifies the tuples of **x**-values that are regarded as possible outcomes of the data collection, given the true population variable $\mathbf{x}=(x_1,x_2,...x_N)$. $\$(\mathbf{x})$ is referred to as the collection of **conceivable x-samples** given **x**, alternatively as the **x-sample space**. In a practical situation, the choice of $\$(\mathbf{x})$ should reflect the surveyor's knowledge of the data collection process. In the above set-up it is implicitly assumed that **measurements are carried out without errors**, since sampled **x**-values are assumed to be chosen from the true ones. If one wants to cover situations with measurement errors, the model has to be extended. However, if not stated otherwise, we presume in the following that inference models are "without measurement errors".

Next we turn to the family $\{P_{\mathbf{x}};\ \mathbf{x}\in\Omega\}$. Usually the distributions $P_{\mathbf{x}}$ can be chosen in a fairly objective/unanimous way. However, in the general setting, and notably in systematic sampling contexts, $P_{\mathbf{x}}$ may very well be interpreted as the surveyor's "subjective" probability distribution for the different conceivable **x**-samples under the premise that he/she knows the true **x** and also given his/her knowledge of the data collection process, but without any knowledge of the actual outcome of the **x**-sample. Thereby we adopt the view that "opinions under uncertainty" (always) can be represented in a meaningful and "personally unique" way by personalistic probability distributions. We shall not dwell on these foundational matters, which often are quite controversial, we just refer to the literature on foundation of statistics for a fuller discussion, e.g. to the book by Savage (1954). $P_{\mathbf{x}}$ is referred to as the **x-sample distributions** given **x**, and $\{P_{\mathbf{x}};\ \mathbf{x}\in\Omega\}$ is called **the family of x-sample distribution.** The choice of the family $\{P_{\mathbf{x}};\ \mathbf{x}\in\Omega\}$) is (of course) the most intricate part of the specification of an inference model for a practical situation. We postpone the discussion of this matter for a while, and right now we only emphasize that we have adopted the view that it is possible to specify a unique (personalistic) probability distribution for the conceivable **x**-samples given **x**. We also make the following comment. By allowing personalistic probability distributions we have opened up for a "totally" Bayesian approach to inference problems, but we shall not pursue that route any further. In particular, we view **x** as a **parameter** (in the sense of general statistical theory), not as a random quantity (in the Bayesian sense).

Under additional assumptions on the data collection process, the space $\$(\mathbf{x})$ of conceivable **x**-samples can be chosen with more specific structure. For sample selection there are various general options as e.g. "with or without replacement", "predetermined or random sample size", "whether the sampled **x**-values carry a relevant order or not", etc. For the purposes of this paper it will suffice to consider sample selections **without replacement,** with **predetermined sample size**, denoted n, and which render the sampled **x**-values an **order of relevance**. In view of these assumptions we restrict to **x**-sample spaces of the form $\$(\mathbf{x})=\mathbf{x}^{(n)}$, where $\boldsymbol{x}^{(n)} = \{(x_{i_1},x_{i_2},\dots,x_{i_n}) : (i_1,i_2,\dots,i_n)$ runs over the ordered n-tuples of distinct elements from U}.

When addressing a problem of the type (2.4), we presume that an inference model $(\Omega,\{\mathbf{x}^{(n)};\ \mathbf{x}\in\Omega\},\{P_{\mathbf{x}};\ \mathbf{x}\in\Omega\})$ is specified. Then, the **x**-sample can be viewed as a random vector $\mathbf{X}=(X_1, X_2,\dots,X_n)$, in which the component order ($\nu=1,2,\dots,n$) agrees with the relevant sample order. For each **x** in Ω, the (simultaneous) distribution of $X_1,X_2,\dots,X_n$ is determined by $P_{\mathbf{x}}$. Expectation and variance relative to $P_{\mathbf{x}}$ are denoted by $E_{\mathbf{x}}$ and $V_{\mathbf{x}}$ respectively. An **X-statistic**, say K(X), is a function of the (random) sample values, $K(X)= k(X_1,X_2,\dots,X_n)$. (For statistics we make no assumption about symmetric functions, though.)

We now turn to the problem of making statements about a characteristic $\chi(\mathbf{x})$ on the basis of an **x**-sample. A statistic K(X) is said to yield **unbiased** (point) **estimation** of $\chi(\mathbf{x})$ if;

$$E_{\mathbf{x}}[K(X)] = \chi(\mathbf{x}),\quad \mathbf{x}\in\Omega\,. \tag{2.5}$$

As regards point estimation we resort on the following standard principles; (i) To be admissible, an estimator should be at least approximately unbiased, preferably unbiased. (ii) Among (approximately) unbiased estimators, an estimator is the more preferred, the smaller its variance is, or is believed to be. For the sake of brevity we leave to the reader to make precise, according to own preferences, the admittedly vague notion "approximately unbiased".

A good survey should not only present point estimates for characteristics of interest, but also report on "uncertainty bounds" for the estimates. Also on this matter we follow the standard route, i.e. to give uncertainty bounds in terms of confidence intervals. Say that the statistics K(X) and Q(X), where Q is non-negative, are **λ-pivotal** for the characteristic $\chi(\mathbf{x})$, if the following relation holds with "good enough" approximation (Φ denoting the standard normal distribution function);

$$P_x(\,-\lambda \le [K(X)-\chi(x)]/\sqrt{Q(X)} \le \lambda\,) \approx \Phi(\lambda/2)\,,\quad x\in\Omega\,. \tag{2.6}$$

Hence, if K and Q are λ-pivotal for $\chi(\mathbf{x})$, the (random) interval $[K-\lambda\sqrt{Q},\ K+\lambda\sqrt{Q}]$ is a confidence interval for $\chi(\mathbf{x})$, with approximate confidence level $\Phi(\lambda/2)$. When speaking of pivotal without specific "λ-qualification" we assume that $\lambda=2$, which makes the (random) interval $[K-2\cdot\sqrt{Q}, L+2\cdot\sqrt{Q}]$ to an **approximately 95% confidence interval** for $\chi(x)$. (Thereby we adhere to the survey sampling practice of regarding ≈95% as the "canonical" confidence level.)

Hence, the chief problem of (estimation) inference is to find quantities K and Q which are pivotal for a given **x**-characteristic $\chi(\mathbf{x})$, including justification of the "pivotality". The perfect justification would be a rigorously proved **normal distribution approximation result**, to the effect that the following inequality holds for an ε that is small enough to comply with prevailing approximation requests;

$$\sup_{x\in\Omega}\ \sup_{-\infty<t<\infty}\ | P_x([K(X)-\chi(x)]/\sqrt{Q(X)}\le t)-\Phi(t)\,|\le\varepsilon.$$

However, it is usually very difficult to give rigorous justifications of normal approximation results of the above type. It is a good deal easier, but still difficult, to prove analogous results on **asymptotic normality**. In the sequel, we shall partly rely on asymptotic results, but we shall also resort to a combination of the following arguments (A) and (B) (and the reader can chose the mixture of them, that he/she prefers). (A) There is an, although unproven, "meta-theorem" which says that any "decent" statistic is asymptotically normally distributed, with its expected value and variance as normal distribution parameters, when the sample size increases "beyond all bounds". (B) Whatever asymptotic results one has at disposal, when it comes to the practical, and thereby "finite", situation, an amount of approximation will always remain. Therefore when establishing pivotality, at some stage one has to rely on empirical evaluations even if one was guided to the pivotal quantities by theoretical results.

We shall expand somewhat on the contents of the meta-theorem which is alluded to in (A). A more precise, but still quite vague, version of it runs as follows. Let K be a "decent", in particular (approximately) unbiased, estimator for $\chi(x)$ and let Q be a "decent", in particular (approximately) unbiased, estimator of the variance $V_x(K)$ i.e. the following relation is assumed to be satisfied;

$$E_x[Q] = V_x[K] \quad (\text{or } E_x[Q] \approx V_x[K]),\ \ x\in\Omega. \tag{2.7}$$

Then K and Q satisfy (2.6), at least under "general conditions", provided that the sample size is "fairly large". In particular, K and Q are pivotal for $\chi(x)$. A further aspect of "decency" should be that the estimator is **consistent**, i.e. that with increasing sample size it converges (in probability) to the characteristic it estimates. Also the term "consistent" is admittedly vague, but we sustain from a more detailed discussion of it, and again we leave to the reader to fill in details according to own preferences.

So far we have only considered univariate inference models. Extensions to multivariate situations can be made along different lines. One concerns inference about a multivariate characteristic $\chi(x,y,...)$ (e.g. a correlation coefficient). Extension in this direction can be made by "straightforward analogy", letting the (multivariate) variable (x,y,...) play the role of the univariate variable **x** in the hitherto considerations. The details are left to the reader.

In conclusion the above discussion amounts to the following. In the search for good procedures for (estimation) inference for a characteristic $\chi(x)$, the following steps should be run through. (i) An inference model for **x**-samples is specified. (ii) Look for (approximately) unbiased estimators for $\chi(\mathbf{x})$, where unbiased is relative to the family $\{P_x;\ x\in\Omega\}$ of **x**-sample distributions. If more than one estimator candidate is available, select the one which has, or is believed to have, the smallest variance. Let K be an estimator of special interest. (iii) Seek an (approximately) unbiased estimator Q of the **inferential variance** $V_x(K)$, again unbiasedness is relative to the family $\{P_x;\ x\in\Omega\}$. Such an estimator Q is referred to as an **estimator of the inferential variance** for K. (iv) Then K and Q are pivotal for $\chi(\mathbf{x})$.

2.3 Inference models with unit-sample distributions

So far we have not required that the family $\{P_x;\ x\in\Omega\}$ of **x**-sample distributions in an inference model $(\Omega,\{x^{(n)};\ x\in\Omega\},\{P_x;\ x\in\Omega\})$ should have any particular structure, but for the assumption that P_x should be a probability distribution on $\mathbf{x}^{(n)}$ for every $\mathbf{x}\in\Omega$. However, in most practical survey situations there is additional structure to the family $\{P_x;\ x\in\Omega\}$, which emanates from the fact that an **x**-sample usually is generated by first selecting a **unit-sample** from the population and then observing **x** for the sampled units. Next we settle some notions in this type of context. Then we work under the presumption that **missing values do not occur**, i.e. that **x**-values actually can be collected for all members of the unit-sample. In particular we presume that non-response does not occur. As is well known, this is an exceptional situation in practice. Its modelling requires extension, though, of the model which we are about to introduce, and we sustain from making that extension in the first round.

A **unit-sample distribution** for the population U=(1,2,...,N) is a probability distribution/-measure **P** on a specified collection $\$(U)=\{\sigma\}$ of subsets of U, say for definiteness ordered subsets. In terms of the previous notion of "**x**-sample distribution", a unit-sample distribution can be viewed as an $\boldsymbol{\ell}$-sample distribution, where $\boldsymbol{\ell}$ is the "label variable" $\boldsymbol{\ell}$=(1,2,...,N). The elements, σ, of $\$(U)$ are referred to as **conceivable unit-samples**. The **sample inclusion indicators**, $I_1,I_2,...,I_N$, are the following random variables on $\$(U)$; I_i=1 if i is contained in s and I_i=0 otherwise. The **inclusion probabilities** of the first and second orders for **P** are defined as follows, where **E** denotes expectation with respect to **P**;

$$\pi_i := \mathbf{P}(I_i=1) = \mathbf{E}(I_i),\ \ i=1,2,...,N, \tag{2.8}$$

$$\pi_{ij} := \mathbf{P}(I_i=I_j=1) = \mathbf{E}[I_i\cdot I_j],\ \ i,j=1,2,...,N. \tag{2.9}$$

Under additional assumptions on the sampling process, the set $\$(U)$ can be chosen with more structure, than being just a collection of subsets of U. Unless otherwise stated, we presume that a unit-sample is selected **without replacement**, has **prescribed sample size** n (i.e. that $I_1+I_2+...+I_N$ = n) and that it carries **a relevant order**. Then $\$(U)$ can, and will, be chosen as $N^{(n)}$ = the collection of all ordered n-tuples of distinct elements from (1,2,...,N).

For a population U with a variable $\mathbf{x}=(x_1,x_2,...,x_N)$ and a unit-sample distribution **P**, the corresponding **induced x-sample distribution**, denoted $\mathbf{P}\circ\mathbf{x}$, is defined as follows (as before $X=(X_1,X_2,...,X_n)$ is the **x**-sample viewed as a random vector);

$$\boldsymbol{P}\circ\boldsymbol{x}(X_1=u_1, X_2=u_2,..., X_n=u_n) = \sum_{\{(i_1,i_2,...,i_n)\in N^{(n)}:\ x_{i_k}=u_k,\ k=1,2,...,n\}} \boldsymbol{P}[(i_1,i_2,...,i_n)]\,. \tag{2.10}$$

If all units in U have the same inclusion probability, the unit-sample is said to be **self-weighting**. An example is provided by the "prototype" for all sampling procedures, simple random sampling. Size n (ordered) **simple random sampling** (without replacement) from U ($n\leq N$), is specified by the unit-sample distribution **P** which has $N^{(n)}$ as the set of conceivable samples and gives each element in $N^{(n)}$ the same probability. Simple random sampling with sample size n will be referred to by the notation SRS(n).

We now return to inference. An inference model is said to **have unit-sample distribution,** if it is of the form $(\Omega,\{\mathbf{x}^{(n)};\ \mathbf{x}\in\Omega\},\{\mathbf{P}_{\mathbf{x}}\circ\mathbf{x};\ \mathbf{x}\in\Omega\})$. The family $\{\mathbf{P}_{\mathbf{x}};\ \mathbf{x}\in\Omega\}$ is then called the **family of unit-sample distributions.** An inference model with unit-sample distribution is said to have **canonical** unit-sample distribution $\mathbf{P}$, if $\mathbf{P}_{\mathbf{x}}=\mathbf{P}$ for all $\mathbf{x}\in\Omega$. The most common situations in sample survey practice lead to inference models with canonical unit-sample distributions, which therefore can be regarded as sort of "standard" model. However, as we shall see, models with non-canonical unit-sample distributions are highly relevant in connection with systematic samples (and also in other cases). The main reason why models with canonical unit-sample distributions are most common in sample survey practice is as follows. On good grounds there is a strong recommendation for sample survey practice, which is also widely followed, that variable samples, say **x**-samples, should be collected by first selecting a unit-sample, and then observing the **x**-values for the sampled units. Moreover, it is recommended that the unit-sample should be generated by an "objective" random mechanism, which operates independently of the specific **x**-values (and of the values of other study variables). When this type of **x**-sample selection is employed, it is usually non-controversial, whatever be the surveyor's views on foundation of probability and statistical inference, that the relevant inference model is, at least in essence, $(\Omega,\{\mathbf{x}^{(n)};\ \mathbf{x}\in\Omega\},\ \{\mathbf{P}\circ\mathbf{x};\ \mathbf{x}\in\Omega\})$, where $\mathbf{P}$ follows from the random selection mechanism. Thereby we have also indicated at least one general answer to the question which was posed earlier: "How to select the family of **x**-sample distributions in a practical situation?" Another, but much more technical, reason for the interest in the model $(\Omega,\{\mathbf{x}^{(n)};\ \mathbf{x}\in\Omega\},\ \{\mathbf{P}\circ\mathbf{x};\ \mathbf{x}\in\Omega\})$ is the following. Under an inference model with canonical unit-sample distribution, there is a general, and "smooth", estimation theory available, known as Horvitz-Thompson theory, which is briefly reviewed in the next sub-section.

We conclude this section with some further comments on multivariate situations. In most practical sample surveys, the estimation interest concerns characteristics for many different study variables, $\mathbf{x}(1),\mathbf{x}(2),...,\mathbf{x}(m)$ (e.g. sex, age,..,income for individuals), the values of which are observed for the units in a (common) unit-sample. We shall not consider the estimation problems for the different variables as "simultaneous", though, but as "separate". Then, one specifies a collection of univariate inference models; $(\Omega_1,\{\mathbf{x}(1)^{(n)};\ \mathbf{x}(1)\in\Omega_1\},\{\mathbf{P}_{\mathbf{x}(1)};\ \mathbf{x}(1)\in\Omega_1\})$, $(\Omega_2,\{\mathbf{x}(2)^{(n)};\ \mathbf{x}(2)\in\Omega_2\},\{\mathbf{P}_{\mathbf{x}(2)};\ \mathbf{x}(2)\in\Omega_2\})$, ..., $(\Omega_m,\{\mathbf{x}(m)^{(n)};\ \mathbf{x}(m)\in\Omega_m\},\{\mathbf{P}_{\mathbf{x}(m)};\ \mathbf{x}(m)\in\Omega_m\})$. In general, these models need not be related in any particular way, but for the fact that the same population size and the same sample size occur in all of them. Even under the assumption that each separate model has canonical unit-sample distribution, the unit-sample distributions $\mathbf{P}_1,\mathbf{P}_2,..\ .,\mathbf{P}_m$ may be taken to be different, i.e. the collection of inference models may be of the form $(\Omega_1,\{\mathbf{x}(1)^{(n)};\ \mathbf{x}(1)\in\Omega_1\},\{\mathbf{P}_1\circ\mathbf{x}(1);\ \mathbf{x}(1)\in\Omega_1\})$, $(\Omega_2,\{\mathbf{x}(2)^{(n)};\ \mathbf{x}(2)\in\Omega_2\},\{\mathbf{P}_2\circ\mathbf{x}(2);\ \mathbf{x}(2)\in\Omega_2\})$, ..., $(\Omega_m,\{\mathbf{x}(m)^{(n)};\ \mathbf{x}(m)\in\Omega_m\},\{\mathbf{P}_m\circ\mathbf{x}(m);\ \mathbf{x}(m)\in\Omega_m\})$. In many (not to say most) situations, though, it is natural to employ the same unit-sample distribution $\mathbf{P}$ in all models; $\mathbf{P}_1= \mathbf{P}_2=...=\mathbf{P}_m =\mathbf{P}$. If so, we say that $\mathbf{P}$ is **uniformly canonical** for the variables $\mathbf{x}(1),\mathbf{x}(2),..\ .,\mathbf{x}(m)$.

The reader may mean that the above frame-work for multivariate situations gives bit of "overkill capacity" as regards practice. We admit that one only seldom goes outside the frame-work with uniformly canonical unit-sample distributions. However, systematic sampling, the main concern of this paper, is one of the "exceptions". In Section 3, notably in Examples 3.1 and 3.2, we present more concrete examples of the various possibilities which are discussed above.

2.4 Horvitz-Thompson theory

We consider (estimation) inference from an **x**-sample under an inference model with unit-sample distribution, $(\Omega,\{x^{(n)};\ x\in\Omega\},\ \{P_x\circ x;\ x\in\Omega\})$. As usual let $I_1,I_2,...,I_N$ be the inclusion indicators and $X=(X_1,X_2,...,X_N)$ the **x**-sample viewed as a random vector. A **linear X-statistic** with **weight variable** $w=(w_1,w_2,...,w_N)$, where $w_1,w_2,...,w_N$ are regarded as known for all units in U, is a statistic of the form;

$$L(X;w)=\sum_{i\in U} x_i\cdot w_i\cdot I_i\,. \tag{2.11}$$

Under the assumed model, the inclusion probabilities (of first as well as second order) depend on **x**, and we use the notation $\pi(x)=(\pi_1(x),\pi_2(x),...,\pi_N(x))$, where $\pi_i(x)=P_x(I_i=1)$. By taking expectation in (2.11) and recalling (2.8) we get;

$$E_x[L(X;w)]=\sum_{i\in U} x_i\cdot w_i\cdot\pi_i(x)\,,\quad x\in\Omega\,. \tag{2.12}$$

In the rest of this section we presume that the unit-sample distribution is canonical, denoted **P**, and, hence, that the inference model is $(\Omega,\{x^{(n)};\ x\in\Omega\},\ \{P\circ x;\ x\in\Omega\})$. Then the inclusion probabilities no longer depend on **x**, and we return to the notation of (2.8) and (2.9). Moreover we assume that $\pi_i>0$ for $i=1,2,...,N$. Then the following well-known result follows readily from (2.12). The statistic

$$\hat{\tau}(x)_{HT}:=L(X;\frac{1}{\pi})=\sum_{i\in U}\frac{x_i}{\pi_i}\cdot I_i\,. \tag{2.13}$$

yields unbiased estimation of the population **x**-total $\tau(x)$ whatever Ω is. In (2.13), and henceforth, we employ the customary notation that the "hat" ^ signifies that the statistic gives (approximately) unbiased estimation of the characteristic underneath the hat. (The previous statement would be true also for a general family $\{P_x\circ x;\ x\in\Omega\}$ if π_i is replaced by $\pi_i(x)$ in (2.13). However, $\pi_i(x)$ would in the general case be unknown, and the statistic would be "uncomputable" in the practical situation.) The statistic in (2.13) is referred to as the **Horvitz-Thompson estimator**, shorter **HT-estimator, for the population total** $\tau(x)$. Moreover, the HT-estimator is under very general conditions (to the effect that Ω is not "extremely meagre") unique as unbiased linear estimator of $\tau(x)$. As a follow-up to (2.13), the **HT-estimator for a ratio of totals** characteristic (see (2.3)) is defined as stated in (2.14). This estimator is, however, only approximately unbiased in the general case, but consistent under quite general conditions.

$$\hat{\mu}(x;y)_{HT}:=\hat{\tau}(x)_{HT}/\hat{\tau}(y)_{HT}\,. \tag{2.14}$$

Our subsequent estimation efforts will relate to the problem of finding statistics Q so that $\hat{\tau}(x)_{HT}$ and Q are pivotal for $\tau(x)$, and more generally so that $\hat{\mu}(x,y)_{HT}$ and Q are pivotal for $\mu(x,y)$. Thereby the interest will focus on variances of linear statistics, notably HT-estimators. However, for the sake of "simplicity of formulas" we shall let the basic investigations relate to the "standardized" version of a linear statistic which is constituted by the **sample sum**. We reserve the letter T for this statistic, and we allow (according to convenience) for both of the notations T(X) and T(x);

$$T(X)=T(x)=\sum_{i\in U} x_i\cdot I_i=\sum_{v=1}^{n} X_v\,. \tag{2.15}$$

Restriction to sample sums means no loss of generality, though, because any linear statistic can be viewed as the sample sum for a suitably transformed variable, as stated below;

$$L(\mathbf{X};\mathbf{w}) = T(\mathbf{x}\cdot \mathbf{w}). \tag{2.16}$$

By employing (2.16), a result for sample sums is readily transformed to a corresponding result for linear statistics. In the following Sections 3, 4 and 5 we mainly confine the studies to the behaviour of sample sums, but in Section 6 we specialize to HT-estimators. We conclude this sub-section with a brief review of some well-known results on variances and variance estimation for sample sums. The inference model is still assumed to be $(\Omega,\{\mathbf{x}^{(n)}; \mathbf{x}\in\Omega\},\{\mathbf{P}\circ\mathbf{x}; \mathbf{x}\in\Omega\})$. The inferential variance for T(X) is given by the following formula;

$$V_x[T(X)] = \sum_{i\in U}\sum_{j\in U} x_i\cdot x_j\cdot(\pi_{ij}-\pi_i\cdot\pi_j), \quad x\in\Omega. \tag{2.17}$$

Provided that $\pi_{ij} > 0$ for all i,j=1,2,..,n, the statistic in (2.18) below yields unbiased estimation of the inferential variance V_x[T(X)], whatever Ω is;

$$\hat{V}_x[T(X)] = \sum_{i\in U}\sum_{j\in U} x_i\cdot x_j\cdot(1-\pi_i\cdot\pi_j/\pi_{ij})\cdot I_i\cdot I_j. \tag{2.18}$$

Under the premise of prescribed sample size, there are some well-known alternative (but algebraically equivalent) forms of the right hand sides in (2.17) and (2.18).

The more general HT-estimator in (2.14) is, however, not a linear statistic in the general case, it is a ratio of linear statistics. Therefore, (2.17) and (2.18) do not apply directly to yield results pertaining to µ(x,y). In fact, there exists no closed form expression for the variance of µ(x,y) (not even for its expected value), and in line with that no closed form unbiased estimator of the inferential variance. This obstacle is usually circumvented by resorting to so called Taylor approximation, which leads to the following well-known approximate variance formula (in which the middle form requires that µ(y)>0, and the right hand form also that µ(x)>0));

$$\begin{aligned} V_{x,y}[\hat{\mu}(x;y)_{HT}] &\approx \frac{1}{\mu(y)^2}\cdot V_{x,y}[L(X-\mu(x;y)\cdot Y;\mathbf{1/\pi})] = \\ &= \mu(x,y)^2\cdot V_{x,y}[L(x/\mu(x)-y/\mu(y);\mathbf{1/\pi})], \quad (x,y)\in\Omega. \end{aligned} \tag{2.19}$$

As can be seen, the formula (2.19) brings problems about the variance of the general HT-estimator in (2.14) back on problems about the variance of a linear statistic, which in turn (by (2.16)) have been brought back on problems on the variance of a sample sum.

Example 2.1: In the special case when **P** is size n **simple random sampling,** (2.17) and (2.18) specialize to the following well-known formulas;

$$V_x[T(X)] = n\cdot(1-n/N)\cdot\sigma^2(x), \quad \hat{V}_x[T(X)] = n\cdot(1-n/N)\cdot S^2(X). \tag{2.20}$$

where σ^2(x) and S^2(X) denote population variance and sample variance respectively i.e., with $\bar{X}$ denoting the sample mean;

$$\sigma^2(x) = \frac{1}{N-1}\cdot\sum_{i=1}^{N}(x_i-\mu(x))^2, \quad S^2(X) = \frac{1}{n-1}\cdot\sum_{v=1}^{n}(X_v-\bar{X})^2. \tag{2.21}$$

■

3. Systematic sampling and inference from systematic samples

3.1 Systematic pps-sampling

There are various (minor) variations on the theme of systematic pps-sampling, and we start by stating the definition which will be used throughout this paper. As usual, let $U=(1,2,\dots,N)$ denote a population. We presume that a list of the population units is available for use as sampling frame. When drawing a systematic unit-sample, the order of the units in the sampling frame plays a crucial role. We shall soon elaborate on that matter, but until further notice we assume that an order has been decided upon, and for simplicity we assume that the labels $1,2,\dots,N$ also specify the frame order. Moreover, the sampling frame is assumed to contain a distinguished variable $s=(s_1,s_2,\dots,s_N)$, called the **size (measure) variable**, all values of which are positive. Define the **cumulative sizes** $c=(c_1,c_2,\dots,c_N)$ by $c_0=0$, $c_i=s_1+s_2+\dots+s_i$, $i=1,2,\dots,N$, and call $\mathfrak{I}_i = (c_{i-1},c_i]$ the **size interval** associated with unit i, $i=1,2,\dots,N$.

A **size** n **systematic pps-sample** from the (ordered) population U, drawn with **size measure s** is selected by executing the following steps (i)-(iii).

(i) Determine the **sampling interval** d by $d = c_N/n = (s_1+s_2+\dots+s_N)/n$.

(ii) Select the random **starting point** B_1 with uniform probability distribution on the interval (0,d]. Then compute the (random) points $B_\nu = B_1 +(\nu-1)\cdot d$, $\nu=2,3,\dots,n$.

(iii) The **unit-sample** consists of the units which are associated with the size intervals $\mathfrak{I}$ into which the points $B_1,B_2,\dots,B_n$ fall. The sample is regarded to be **ordered** in the same way as the B:s.

If all sizes are equal, i.e. if $s_1 =s_2 =\dots= s_N$, the sample is called a **simple systematic sample**.

In order that the steps (i)-(iii) with certainty shall lead to a sample with n different units (and not fewer), it is necessary and sufficient that no size interval is longer than the sampling interval, i.e. that the following condition (3.1) is met. Unless stated otherwise, (3.1) is assumed to be in force in the sequel.

$$s_i \le c_N / n, \quad i=1,2,\dots,N. \tag{3.1}$$

Remark 3.1: If (3.1) is not satisfied in a practical situation, one usually proceeds as follows. Units which violate (3.1) are (with certainty) taken out for observation. Then condition (3.1) is checked for the "remaining population" and the "remaining sample size", and if (3.1) is still violated, further units are brought to the "certainty part" of the sample, and (3.1) is checked again, etc. Ultimately (3.1) becomes satisfied, and then a systematic pps-sample of the remaining sample size is selected from the remaining population.

In some applications, though, one executes the steps (i)-(iii) as they stand even if (3.1) is not satisfied. Then the sample may contain replicated units, and the number of different units in the sample is a random quantity. ■

The above steps (i)-(iii) specify a unique unit-sample distribution, which depends on the frame order, though, and next we give a more formal specification of this dependence. Then we have to distinguish between two kinds of labels, which we, for simplicity, let coincide above. Redenote the **identifying labels** by setting them in parenthesis; $U=((1),(2),\dots,(N))$, and let

1,2,...,N be the **frame (order) labels.** The relation between the two types of labels is specified by a permutation $\mathbf{q}=(q_1,q_2,\dots,q_N)$ of (1,2,...,N), referred to as a **frame order**, with the following interpretation. The first unit in the frame is the one with (identifying) label (q_1), the second is the one with label (q_2), etc. The values of the variable $\mathbf{x}=(x_{(1)},x_{(2)},\dots,x_{(N)})$ on U are given frame orders as follows;

Under $\boldsymbol{q}$ the ***frame order for the x-values*** is $(x_{(q_1)},x_{(q_2)},\dots,x_{(q_N)})$.

Given the frame order **q**, the sample selection steps (i)-(iii) in Section 3.1 are applied, with 1,2,...,N in **s**, **c** and $\mathfrak{S}$ interpreted as frame order labels. Thereby, for each permutation **q** a unit-sample distribution is specified, which we denote **P**(· ;n;N;**s**;**q**). However, unless otherwise stated, we regard N as "canonical" and suppress it, leading to the shorter notation **P**(· ;n;**s**;**q**). We also use **SYS** as a mode for referring to systematic pps-sampling, and combinations like SYS(n;**s**;**q**). The notation SYS(n;**s**) refers to systematic pps-sampling with some **q**-order, which may be more or less unspecified, though. Analogously, **SSYS** refers to simple systematic sampling. The **collection of all permutations** of (1,2,...,N) is denoted by (N!).

It is readily realized that **P**(· ;**s**;**q**) and **P**(· ;**t**;**q**) coincide if the size measures **s** and **t** are proportional, i.e. if for some constant k we have; $s_i= k\cdot t_i$, i=1,2,...,N. This many-to-one correspondence between size measures and unit-sample distributions is eliminated if one confines to size measures which are standardized to the effect that the sizes sum to 1. The **standardized version of the size measure s**, denoted $\mathbf{z}(\mathbf{s})=(z_1(\mathbf{s}),z_2(\mathbf{s}),\dots,z_N(\mathbf{s}))$, is;

$$z_i(\boldsymbol{s}) = s_i \Big/ \sum_{r=1}^{N} s_r\,, \quad i=1,2,\dots,N. \tag{3.2}$$

On the other hand, for different frame orders $\mathbf{q}_1$ and $\mathbf{q}_2$, **P**(· ;**s**;$\mathbf{q}_1$) and **P**(· ;**s**;$\mathbf{q}_2$) are in general different. All **P**(· ;**s**;**q**) with the same **s** have, however, the same "marginal distributions" in the sense of inclusion probabilities, as is stated in the following well-known result.

For every frame order **q**∈(N!), **P**(· ;n;**s**;**q**) has the inclusion probabilities
$\pi_i = n\cdot z_i(\mathbf{s})$, i=1,2,...,N. (3.3)

Remark 3.2: From (3.3) and (3.2) it is seen that simple systematic sampling SSYS(n;**q**) is self-weighting, whatever be the frame order **q** and, hence, has inclusion probabilities n/N. ■

3.2 Inference models

We now turn to inference from **x**-values in a SYS-sample, and we adhere to the general inference frame-work which was introduced in Section 2. We start by stating a special structural assumption on inference models for SYS-samples The concrete interpretations of the assumption are discussed afterwards.

DEFINITION 3.1: A univariate inference model for **x**-values in a SYS(n;**s**)-sample is assumed to be of the form $(\Omega,\{\mathbf{x}^{(n)};\ \mathbf{x}\in\Omega\},\{\mathbf{P}_x\circ\mathbf{x};\ \mathbf{x}\in\Omega\})$, where the family $\{\mathbf{P}_x;\ \mathbf{x}\in\Omega\})$ of **x**-sample distributions has the following structure;

$$\boldsymbol{P}_x = \sum_{\boldsymbol{q}\in(N!)} h_x(\boldsymbol{q})\cdot \boldsymbol{P}(\cdot\,;n;\boldsymbol{s};\boldsymbol{q}), \quad \boldsymbol{x}\in\Omega, \tag{3.4}$$

where $\{h_x(\cdot\ );\ \mathbf{x}\in\Omega\}$ is a family of probability distributions on the set (N!) of all permutations **q** of (1,2,...,N), i.e.;

$$h_x(\boldsymbol{q})\geq 0\,,\ \boldsymbol{q}\in(N!) \quad\text{and}\quad \sum_{\boldsymbol{q}\in(N!)} h_x(\boldsymbol{q}) = 1\,,\ \boldsymbol{x}\in\Omega\,. \tag{3.5}$$

Now to the interpretation of the above definition. In general an inference model shall codify the "uncertainty" which depends on lack of a unique relation between sampled **x**-values and those in the population. In the SYS case, an inference model for the variable **x** shall formalize two main sources of uncertainty; the randomness in the selection of the starting point in step (ii) above, and the lack of precise knowledge of the frame order for the **x**-values. To that end the above model comprises two main ingredients; a family $\{\mathbf{P}(\cdot\ ;n;s;\mathbf{q});\ \mathbf{q}\in(N!)\}$ of unit-sample distributions and a family $\{h_{\mathbf{x}}(\cdot\ );\ \mathbf{x}\in\Omega\}$ of **frame order distributions** for **x**. The first family formalizes the uncertainty which enters via the random selection of the starting point, and this family is specified in Section 3.1.

The interpretation of the family $\{h_{\mathbf{x}}(\cdot\ );\ \mathbf{x}\in\Omega\}$ is usually more intricate. Its scope is to codify the uncertainty which emanates from the fact that one does not have full knowledge of the frame order for the **x**-values. In most practical situations the sampling frame order is not generated by an "objective" random mechanism, though (which would specify $h_{\mathbf{x}}(\cdot\ )$ in an "objective" way). On the contrary, the surveyor usually strives for a sampling frame arrangement which renders as much "regular trend" as possible to one or more of the study variables. Therefore, in general $\{h_{\mathbf{x}}(\cdot\ );\ \mathbf{x}\in\Omega\}$ should reflect the "remaining uncertainty" about the frame order for the **x**-values, given the surveyor's knowledge about the frame ordering process. Normally $\{h_{\mathbf{x}}(\cdot\ );\ \mathbf{x}\in\Omega\}$ has to be viewed as a family of subjective/personalistic probability distributions, which reflect the surveyor's opinion about the frame order for the **x**-values, and this will be our main view on $\{h_{\mathbf{x}}(\cdot\ );\ \mathbf{x}\in\Omega\}$ in the following.

Among the possible opinions on the frame order for **x**, there is a distinguished one which we refer to as **total ignorance of the frame order for x, or totally random frame order for x.** In everyday language this state of knowledge would be described in words like: "I have no idea of the structure in the frame order for the **x**-values. Any order is as likely as any other." This type of opinion is specified formally as follows;

$$h_{\mathbf{x}}(\boldsymbol{q}) = \frac{1}{N!} \quad \text{for } \boldsymbol{q}\in(N!)\,,\quad \boldsymbol{x}\in\Omega\,. \tag{3.6}$$

Under assumption of total ignorance of frame order the following statements hold true, as is well known and readily checked.

(i) The family $\{\mathbf{P}_{\mathbf{x}};\ \mathbf{x}\in\Omega\}$ of unit-sample distributions in the SYS(n;s) inference model is a family with canonical distribution. This canonical unit-sample distribution is called **randomized systematic s-sampling** with sample size n. It will be denoted by $\mathbf{P}_R(\cdot\ ;n;s)$, and is also referred to by **RSYS**. A somewhat more formal definition of $\mathbf{P}_R$ is as follows. The "total" sample selection is carried out by first implementing a totally random order in the sampling frame, and then selecting a unit-sample according to the steps (i)-(iii) in Section 3.1. (3.7)

(ii) The RSYS-procedure corresponding to simple systematic sampling SSYS(n) is nothing but simple random sampling SRS(n). (3.8)

Situations where the assumption (3.6) is regarded as inappropriate are commonly referred to by saying that the **x-values exhibit trend in their frame order**. We illustrate the notion of frame order distribution in the following examples.

Example 3.1: a) The frame is ordered with the aid of an auxiliary variable $\mathbf{a}=(a_1,a_2,...,a_N)$, the values of which are known for all units in the population. The surveyor means that the variable **a** and the study variable $\mathbf{x}=(x_1,x_2,...,x_N)$ are related according to a model $x_i = k\cdot a_i + \varepsilon_i$,

$i=1,2,\ldots,N$, where k is a known constant and $(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_N)$ are independent, identically distributed random variables with known distribution. The ε's should usually be viewed as representing opinions rather than as effects of some "physical" random mechanism. If one prescribes a mode for ordering the units in the sampling frame by their **a**-values (e.g. by increasing **a**-values), the family of frame order distribution $\{h_{\mathbf{x}}(\cdot\,);\ \mathbf{x}\in\Omega\}$ for **x** is determined, even though the determination is considerably implicit, and it will contain **a**, k and the ε-distribution as "parameters".

Comment: The determination of $\{h_{\mathbf{x}}(\cdot\,);\ \mathbf{x}\in\Omega\}$ which is indicated above contains various assumptions (about linearity, independence and distribution form), assumptions which usually are subjective in nature. However, this fact does not introduce any new element into the inference model. For SYS-sampling the family $\{h_{\mathbf{x}}(\cdot\,);\ \mathbf{x}\in\Omega\}$, and thereby also the family $\{P_{\mathbf{x}};\mathbf{x}\in\Omega\}$, are practically always to be interpreted as families of subjective probability distributions.

b) For k≠0, the above determination of the family $\{h_{\mathbf{x}}(\cdot\,);\ \mathbf{x}\in\Omega\}$ leads, in the general case, to a family which does not satisfy (3.6). The special case k=0 is of great practical interest. Its interpretation is: "The values of the variables **a** and **x** are unrelated." Then the above determination of $\{h_{\mathbf{x}}(\cdot\,);\ \mathbf{x}\in\Omega\}$ leads to (3.6), whatever be the (common) distribution of the ε's. This fact is often formulated as follows: "Frame ordering with the aid of an auxiliary variable **a** which is unrelated to the study variable **x** leads to total ignorance of the frame order for the **x**-values.

c) A consequence of what is said in a) and b) above runs as follows. Let **x** and **y** be two study variables. If the frame is ordered with the aid of an auxiliary variable **a**, it may very well happen (if **x** and **y** are related to **a** in different manners), that $h_{\mathbf{x}}$ and $h_{\mathbf{y}}$ differ, leading to different $\mathbf{P}_{\mathbf{x}}$ and $\mathbf{P}_{\mathbf{y}}$. In particular, it may ver well happen that the **x**-values exhibit trend in their frame order, while there is total ignorance of the frame order for the **y**-values. ■

Example 3.2: Let, as usual, $s=(s_1,s_2,\ldots,s_N)$ be the (known) size measure and $\mathbf{x}=(x_1,x_2,\ldots,x_N)$ a study variable. Assume that the surveyor thinks that **x** and **s** are related according to the model $x_i = k\cdot s_i\cdot(1+\varepsilon_i)$, $i=1,2,\ldots,N$, where k is a constant which may be unknown, and the ε's are independent, identically distributed random variables with interpretation as in the previous example. Introduce the following transformed **x**-variable: $\mathbf{y}=\mathbf{x}/\mathbf{s}=(x_1/s_1,x_2/s_2,\ldots,x_N/s_N)$. Then the previous relation transforms to $y_i=\varepsilon'_i$ $(= k\cdot(1+\varepsilon_i))$, $i=1,2,\ldots,N$, where $(\varepsilon'_1,\varepsilon'_2,\ldots,\varepsilon'_N)$ are independent, identically distributed random variables. As stated in the previous example, this leads to a family $\{h_{\mathbf{y}}(\cdot\,);\ \mathbf{y}\in\Omega\}$ of frame order distributions for **y**, which satisfies (3.6). Hence, transformation of a study variable may cause substantial change of the family of frame order distributions. ■

In more pronounced probability language, an inference model with the structure (3.4) + (3.5) can be described as follows. Given **x**, the distribution $\mathbf{P}_{\mathbf{x}}$ is the **unconditional** (or mixed) unit-sample distribution which is specified by the conditional distributions given the frame order, $\{\mathbf{P}(\cdot\,;n;s;q);\ q\in(N!)\}$, and the mixing distribution $h_{\mathbf{x}}(\cdot)$. Let expectation and variance relative to the "conditional" distributions $\mathbf{P}(\cdot\,;n;s;q)$ be denoted by $\mathbf{E}(\cdot\,|q)$ and $\mathbf{V}(\cdot\,|q)$ respectively and referred to as the **conditional mean** respectively the **conditional variance given the frame order q**. Then, (3.4) leads to the following expectations formula;

$$\boldsymbol{E}_{\mathbf{x}}(\cdot) = \sum_{q\in(N!)} h_{\mathbf{x}}(q)\cdot\boldsymbol{E}(\cdot\,|\,q),\quad \boldsymbol{x}\in\Omega, \tag{3.9}$$

We conclude this sub-section by commenting on multivariate situations. The following is a continuation of the discussion at the end of Section 2.3. We assume that a SYS(n;s) sample is (or is to be) selected, and that the values of several study variables, say $\mathbf{x}(1),\mathbf{x}(2),\ldots,\mathbf{x}(m)$, have been (or are to be) observed for the unit-sample. We describe the inference situation by a collection of separate univariate SYS(n;s) inference models, and we employ the following notation; $P(\cdot;n;s;h_x)$ denotes the $\mathbf{x}$-sample distribution which is given by (3.4) and (3.5) when the frame order distribution is h_x. The collection of univariate inference models is; $(\Omega_1, \{x(1)^{(n)};\ x(1)\in\Omega_1\},\{P(\cdot;n;s;h_{x(1)})\circ x(1);\ x(1)\in\Omega_1\})$, $(\Omega_2,\{x(2)^{(n)};\ x(2)\in\Omega_2\},\{P(\cdot;n;s;h_{x(2)})\circ x(2);\ x(2)\in\Omega_2\})$, ..., $(\Omega_m,\{x(m)^{(n)};\ x(m)\in\Omega_m\},\{P(\cdot;n;s;h_{x(m)})\circ x(m);\ x(m)\in\Omega_m\})$.

As was discussed in Example 3.1, the families $\{h_{x(k)}(\cdot);\ x\in\Omega_{(k)}\}$ of frame order distributions, $k=1,2,\ldots,m$, may differ between the study variables. Hence, for SYS-sampling it can be quite reasonable to employ different unit-sample distributions $\mathbf{P}_x$ for different study variables, even though the same unit-sample is used to collect observations on all study variables. The reason is that the surveyor's opinion on frame order may differ among the variables. For instance, it may be quite reasonable to have the canonical unit-sample distribution $\mathbf{P}_R$ for some subset of the variables (reflecting total ignorance of frame order for these variables), but have other unit-sample distributions for the other variables. In such a situation, $\mathbf{P}_K$ is uniformly canonical for a particular subset of variables, but not for the entire collection of study variables.

3.3 Point estimation

Here we shall consider estimation based on $\mathbf{x}$-observations on a SYS(n;s) sample, under the general inference model $(\Omega,\{x^{(n)};x\in\Omega\},\{\mathbf{P}_x\circ x;x\in\Omega\})$, which is assumed to comply with Definition 3.1. Let $I_1,I_2,\ldots,I_N$ be the (frame label) inclusion indicators and let $\pi(x)=(\pi_1(x),\pi_2(x),\ldots,\pi_N(x))$ be the inclusion probabilities. Then, as a straightforward consequence of (3.9), (2.8) and (3.3) we have;

$$\pi_i(x) = \sum_{q\in(N!)} h_x(q)\cdot E(I_i\,|\,q) = n\cdot z_i(s),\quad i=1,2,\ldots,N,\quad x\in\Omega. \tag{3.10}$$

Similar reasoning using (2.12) and (3.10) yields for a linear statistic L (see (2.11));

$$E_x[L(X;w)] = n\cdot\sum_{i\in U} x_i\cdot z_i(s),\quad x\in\Omega. \tag{3.11}$$

(3.10) and (3.11) state that inclusion probabilities and expected values for linear statistics are the same under all SYS(n;s) inference models, whatever be the frame order distribution h_x. Hence, although a SYS(n;s) inference model in the general case has a non-canonical unit-sample distribution, the notion of HT-estimator (see (2.13) and (2.14)) applies, and moreover HT-estimators have the usual unbiasedness properties. In view of (3.3), the HT-estimator for a population total under a SYS(n;s) model has the following appearance;

$$\hat{\tau}(x)_{HT} = \frac{1}{n}\cdot\sum_{i\in U}\frac{x_i}{z_i(s)}\cdot I_i\,. \tag{3.12}$$

Remark 3.3: In view of the claim in Remark 3.2; For SSYS the HT-estimator of the population x-mean is simply the sample mean X, i.e. $\hat{\mu}_{HT}(x)=\bar{X}$. ■

The following inferential variance analogue to (3.9) is a straight-forward consequence of the definition of variance, (3.9) and the above-mentioned fact that a linear statistic L has "frame order independent expectation";

$$V_x(L) = \sum_{q\in(N!)} h_x(q)\cdot V(L\,|\,q),\quad x\in\Omega. \tag{3.13}$$

For the "standard" linear statistic, i.e. sample sum, an explicit formula for the inferential variance is obtained by combining (3.13) with the result below on the conditional variance given the frame order, the justification of which is left to the reader.

$$V_x[\mathrm{T}(X)\,|\,q] = \frac{1}{\mathrm{d}}\cdot\int_0^{\mathrm{d}} \sum_{i=1}^{N}\Big[\mathrm{x}_{q_i}\cdot\delta_{q_i}(\mathrm{u}) - \frac{\mathrm{n}}{\mathrm{C_N}}\cdot\sum_{r=1}^{N}\mathrm{x}_r\cdot\mathrm{s}_r\Big]^2\mathrm{du}, \quad x\in\Omega, \tag{3.14}$$

where $\delta_i(u)$ indicates whether unit i is included or not in the sample when the basic random point falls in u, i.e. $\delta_i(u)=1$ if $\mathfrak{I}_i\cap\{u, u+d, u+2\cdot d, ..., u+(n-1)\cdot d\}$ is non-empty, and $\delta_i(u)=0$ otherwise, $0\leq u\leq d$.

The theoretical inferential variance V_x plays (at least) two important roles. One is that it gives information on the precision of an estimator, and another that it is a main instrument for finding pivotal quantities (see Section 2.2). The chief interest in this paper relates to the latter aspect, but before we enter that topic in more detail, we shall employ the above formulas to make some general comments on estimation in SYS sampling situations, and to provide background for the comments we start with an example.

Example 3.3: Let $\mathbf{x}=(x_{(1)},x_{(2)},...,x_{(N)})$ be a variable on the population U, having mean $\mu(\mathbf{x})$. We assume that $\mu(\mathbf{x})=0$, which in fact means no loss of generality in the following, just a technical simplification. Consider SSYS(n) sampling from U and, as usual, let T(X) denote the sample sum for **x**. Also as usual, let **q** denotes a frame order, which implies a frame order for the **x**-values (see Section 3.1).

Case α: Here we assume that $N=2\cdot n$, which means that the sampling fraction is ½. Then, for a fixed frame order **q** there are only two possible unit-samples, namely (in frame order labelling); (1,3,5,...) and (2,4,6,...). Hence, the conditional distribution of T(X) given **q** has the following two-point distribution;

$$P_x(\mathrm{T}(X)=\mathrm{x}_{(q_1)}+\mathrm{x}_{(q_3)}+\mathrm{x}_{(q_5)}+...\,|q) = P_x(\mathrm{T}(X)=\mathrm{x}_{(q_2)}+\mathrm{x}_{(q_4)}+\mathrm{x}_{(q_6)}+...\,|q) = ½\,. \tag{3.15}$$

The following formulas for conditional variances are straightforward consequences of (3.15), the assumption that $\mu(\mathbf{x})=0$ and Remark 3.3 which states that $\hat{\mu}_{HT}(\mathbf{x})=T(X)/n$;

$$\mathrm{V}_x[\mathrm{T}(X)\,|\,q] = [\sum_{i=1}^{n}\mathrm{x}_{(q_{2i})}]^2, \quad \mathrm{V}_x[\hat{\mu}(x)_{\mathrm{HT}}|\,q] = [\frac{1}{n}\cdot\sum_{i=1}^{n}\mathrm{x}_{(q_{2i})}]^2, \quad x\in\Omega. \tag{3.16}$$

Let $\mathbf{x}_0$ denote the particular **x** which is specified by $x_{(1)}= x_{(3)}= ...= x_{(2n+1)}= +1$ and $x_{(2)}= x_{(4)}=... = x_{(2n)}= -1$, and let **1** denote the identity permutation. (A situation of the type $(\mathbf{x}_0,\mathbf{1})$ is commonly referred to as a **periodic case.**) From (3.16) we get;

$$\mathrm{V}_{x_0}[\hat{\mu}(x)_{\mathrm{HT}}|\,\mathbf{1}] = 1\,. \tag{3.17}$$

Now look at $V_x[T(X)\,|\,q]$ as a function of **q**, while **x** is regarded as fixed. Then, with a probability distribution $h_x(\cdot)$ on (N!), $V_x[T(X)\,|\,q]$ can be viewed as a random variable on (N!). Presume that h_x is given by (3.6), i.e. that we are in the case with total ignorance of frame order. Then, by virtue of (3.8), the x-sums in (3.16) can be viewed as sample sums for **x** under SRS(n). Moreover, assume that n is large enough and that **x** is "regular" enough for the SRS central limit theorem (CLT) (see Hájek (1960)) to apply with good approximation. According to the CLT, the x-sums in (3.16) are approximately normally distributed with mean 0 (since $\mu(\mathbf{x})=0$) and variance $n\cdot\sigma^2(\mathbf{x})/2$ (see (2.20)), which leads to the following;

Under (3.6) $V_x[T(X)\,|\,q]$ behaves (with good approximation) as $H^2\cdot n\cdot\sigma^2(\mathbf{x})/2$, where H is a standard normal random variable. (3.18)

Case β: We modify the situation in Case α to the effect that instead of N=2· n, we assume that N=k· n, where k ranges over the natural numbers. Hence, the sampling fraction is 1/k. The frame order distribution h_x is still assumed to be according to (3.6).

The following formula is a consequence of (3.8) and (2.21);

$$V_x[\hat{\mu}(x)_{HT}|q] = \frac{1}{n}\cdot(1-n/N)\cdot\sigma^2(x). \qquad (3.19)$$

The proof of the following assertion is not entirely straight-forward, but it is left to the reader.

> Under (3.6) $V_x[T(X)|q]/n$ converges in P_x-probability, as k→∞, to the unconditional variance $V_x[T(X)]/n$ (which in turn converges to $\sigma^2(x)$) (3.20)

■

With the above examples as background we now make some general remarks.

Remark 3.4: Under any SYS(n;s) inference model, the HT-estimator in (3.12) is "decent" in the sense that it is unbiased (relative to the inferential model). However, it does not necessarily be "decent" in the sense that it yields consistent estimation. If it does or not, depends essentially on the family $\{h_x(\cdot); x\in\Omega\}$ of frame order distributions.

The above general statement is illustrated in Example 3.2 (which concerns SSYS sampling). It is seen from (3.17), that if $h_x(\cdot)$ is specified by $h_x(1)=1$, then the HT-estimator of $\mu(x)$ is not consistent. On the other hand, if $h_x(\cdot)$ is given by (3.6) then the HT-estimator is consistent, as is seen from (3.19).

The first $h_x(\cdot)$-family is quite extreme, though, in having a periodic trend. In general trends in x-values usually pull in the opposite direction, and the following holds in the SSYS case. Trend in the frame order normally leads to reduction of the inferential variance, compared with the case with total ignorance of frame order, and hence in particular "improves" on the consistency of HT-estimators. Illustrations of the claim can be found in most text-books on sampling theory, and also in e.g the review papers by Bellhouse (1988) and Murthy & Rao (1988). ■

Remark 3.5: When the sampling fraction is non-negligible, conditional and unconditional inferential variances are substantially different notions.

The statement is illustrated in Case α in the example, where (3.18) tells that the conditional variance $V_x[T(X)|q]$ may vary considerably, and in a highly erratic way, if the frame is reordered (while $V_x[T(X)]$ is a constant, given x). A practical consequence of this is as follows. If one makes simulation experiments concerning variance estimation in cases with large sampling fraction, it makes a great difference whether or not the frame order is selected anew (according to a prescribed h_x) for each new simulation case.

However, when the sampling fraction is small, the distinction between conditional and unconditional inferential variance is much less important, and the difference more or less disappears as the sampling fraction tends to 0. This is illustrated for SSYS by the statement in (3.20). (In the author's view, most text-books are quite unclear on the distinction between inferential variance and conditional variance given the frame order. However, an excuse is given by the previous statement together with the fact that most sample surveys have small sampling fractions.) ■

3.4 Brief review of estimation of inferential variances (for sample sums)

We now turn to estimation of inferential variances on the basis of **x**-samples $\mathbf{X}=(X_1,X_2,...,X_n)$ generated by SYS(n;s) sampling. We presume an inference model $(\Omega,\{\mathbf{x}^{(n)};\ \mathbf{x}\in\Omega\},\{\mathbf{P}_{\mathbf{x}}\circ\mathbf{x};\ \mathbf{x}\in\Omega\})$ in accordance with Definition 3.1. When restricting to sample sums (as we do for notational simplicity), the characteristic to be estimated is $V_{\mathbf{x}}(X_1+X_2+ ...+X_n)$, $\mathbf{x}\in\Omega$. As a help to keep different variance estimator candidates apart, we introduce the following notational system. Generally, an "inferential variance estimator for sample sums" is referred to by VESS, and when applied to the sample **X**, by VESS[**X**]. Distinction between different estimators is made by notation of the type VESS[**X**;"label"], where the "labels" will be introduced successively. We treat the cases with "total ignorance of frame order" respectively with "trend in frame order" separately.

3.4.1 Variance estimation under total ignorance of frame order

Throughout this sub-section we presume that (3.6) is in force, i.e. that the surveyor "knows next to nothing" about the frame order for the **x**-values. By (3.7) the inference model then is $(\Omega,\{\mathbf{x}^{(n)};\ \mathbf{x}\in\Omega\},\{\mathbf{P}_R\circ\mathbf{x};\ \mathbf{x}\in\Omega\})$, where $\mathbf{P}_R$ is randomized systematic sampling. Thus, the inference model has canonical unit-sample distribution and, hence, the results in Section 2.4 apply. For $\mathbf{P}_R$ holds, as is readily checked, $\pi_{ij} > 0$ for $i,j=1,2,...,N$, and accordingly the variance estimator formula (2.18) could be applied. Unfortunately, though, so far nobody has managed to compute the corresponding π_{ij} exactly. In lack of an exact solutions, one resorts to various approximation approaches, which are briefly discussed below. Also the novel approach to be discussed in the subsequent sections, belongs to the category "approximation solutions".

First we comment on a special case where an exactly unbiased variance estimator in fact can be exhibited under the prevailing presumption about total ignorance of frame order, namely the SSYS(n) case. By virtue of (3.8) and Example 2.1, the estimator in (3.21) below yields unbiased estimation of the inferential variance;

$$\mathrm{VESS}[\mathbf{X};a] = n\cdot(1-n/N)\cdot S^2(\mathbf{X}). \tag{3.21}$$

Next we turn to approximation solutions of the variance estimation problem in the general SYS(n;s) case. The simplest is based on the following heuristic argument.

> At least if the sampling fraction is small, a SYS(n;s) sample behaves very much like a sample which is generated as follows. Instead of selecting the points $B_1,B_2,..\ .,B_n$ as stated in step (ii) in Section 3.1, they are selected independently of each other, each one with uniform distribution on the whole size interval $[0,c_N]$. (3.22)

Under the sampling procedure in (3.22), the sampled **x**-values are independent, identically distributed (iid) random variables. For iid random variables the following well-known estimator yields unbiased variance estimation;

$$\mathrm{VESS}[\mathbf{X};b] = n\cdot S^2(\mathbf{X}). \tag{3.23}$$

The estimator (3.23) is frequently employed in SYS(n;s) practice. When doing so, one relies on an approximation of SYS(n;s), which is a without-replacement sampling procedure, by a with-replacement sampling procedure. This kind of approximation "feels", and is, quite reasonable when the sampling fraction is small, but it "feels", and is, the more questionable, the larger the sampling fraction becomes. Formulas (3.21) and (3.22) show that, in the SSYS case, the "defect" in using (3.23) amounts to omission of the finite population correction factor (fpc) (1-n/N). On intuitive grounds one "feels" that also in the general SYS(n;s) case, (3.23) should be adjusted by some sort of fpc. A simple and near at hand idea is to use the "traditional" fpc, i.e. (1-n/N). This choice transforms VESS[X;b] into VESS[X;a]. Other fpc's have been suggested in the literature, we refer to Wolter (1985), pp 286-290 for a review. However, a common feature for all suggestions is that they are considerably ad hoc.

To the best of our understanding, VESS[· ;a] and VESS[· ;b] are the variance estimators which are most widely used in SYS-practice, the latter when the sampling fraction is negligible and the former when it is non-negligible, but "moderate". VESS[· ;a] can also be viewed as a "combination" of the two. This "strategy" works quite satisfactorily as long as one does not face situations with "strongly non-negligible" sampling fractions (say roughly 25% or more). However, as mentioned in Section 1, such cases do occur in practice, and they constitute the main concern of this paper, which is to find a variance estimator which performs satisfactorily over a wide range of sampling fractions (preferably the entire range).

An approximation approach of a different kind is due to Hartley & Rao (1962). The crucial ingredient in their approach is an approximation of the second order inclusion probabilities π_{ij} via an asymptotic (as N and n become large) expansion. A variance estimator is then obtained by applying (2.18) with the approximate version of π_{ij}. The resulting estimator has cased controversy, though, and counterexamples to the Hartley-Rao approximation claims can be found in e.g Håjek (1981), Chapter 10. Anyway, either it is good or not, the estimator is presented below, the z's being the normalized size measures in (3.2);

$$\mathrm{VESS}[X;c] = \frac{1}{2\cdot(n-1)}\cdot\sum_{i=1}^{N}\sum_{j=1}^{N}[1-n\cdot(z_i+z_j-\sum_{k=1}^{N}z_k^2)]\cdot(x_i-x_j)^2\cdot I_i\cdot I_j\,. \qquad (3.24)$$

3.4.2 Variance estimation when the study variable has frame order trend

There are two main approaches to estimation of inferential variances in situations where one knows, or believes, that the study variable exhibits trend in the frame order. The approaches will be referred to as **modelling of trend** respectively **frame segmentation**, and they are discussed separately in the following.

Modelling of trend

From a technical point of view, trend modelling amounts to specification of a particular, and explicit version of the family $\{h_x(\cdot); x\in\Omega\}$ of frame order distributions. To qualify as "trend modelling", the specification should differ from (3.6).

As discussed in Section 3.3, the choice of $\{h_x(\cdot); x\in\Omega\}$ will not affect the unbiasedness of HT-estimators, but it will affect inferential variances as $V_x[T(X)]$. Under a particular family $\{h_x(\cdot); x\in\Omega\}$ a first question will therefore be to find an, exact or approximate, expression for $V_x[T(X)]$, as a basis for the search for a pivotal variance estimator. Ideally one should also justify the pivotality, by proving some result of the type (2.6). Under general SYS(n;s) sampling this is a huge program to accomplish, and we shall not enter any details in this paper, only make some comments.

One can claim that trend modelling is less relevant for general SYS(n;s) sampling (which is our main concern) than it is for SSYS sampling, using the following type of argumentation. When estimating a total $\tau(x)$ on the basis of x-values in a SYS(n;s) sample, one can distinguish between two alternatives (which are somewhat extreme, though) for the relationship between the study variable **x** and the auxiliary size variable **s**. (i) The variable **s** is strongly related to **x** and the SYS(n;s)-procedure was chosen to achieve good estimation precision due to appropriately varied inclusion probabilities. Then the interesting statistics are the corresponding HT-estimators, the essential ingredients of which are sample sums for the transformed variable **x**/π. Then one is typically in the type of situation that was discussed in Example 3.2. If so, the frame order for **x**/π is described by total ignorance of frame order, whatever procedure is used to order the frame. (ii) Often frames are ordered by **s**, and we assume that this is the case. Assume, that **x** is (fairly) unrelated to **s**, being a study variable of "minor importance". Then, the discussion in Example 3.1 applies to **x** as well as to **x/s**, and their frame orders are both modelled by total ignorance of frame order. In both cases (i) and (ii) one comes to the conclusion that total ignorance of frame order is a good description. However, even if the arguments were meant to support the claim that total ignorance of frame order is the relevant model in quite many SYS(n;**s**) situations, we certainly do not claim that this is always true. However, we shall not pursue the trend modelling approach any further in this paper, but for the frame segmentation approach which is discussed below, and which offers a method for handling situations with frame trend for a study variable.

Next we consider the SSYS case and view **s** as an auxiliary variable which is strongly related to the study variable **x**. Then, the "**x**-information potential" of **s** is not "used up" for the purpose of creating appropriately varying inclusion probabilities. Hence, frame ordering with the aid of **s** can have considerable variance reducing effect, and we presume that the frame is ordered in that way. A consequence of this is that variance estimation based on total ignorance of frame order (or equivalently based on SRS) will be misleading, usually to the effect that the true estimator variance is over-estimated. There is a wealth of suggestions for how to cope with this type of situation by trend modelling approaches, and an extensive review is given in Wolter (1985), Chapter 7. It should be mentioned, though, that the modelling approaches in the literature usually differ somewhat from the one in this paper, to the effect that most writers work in super-population frame-works, and not in our frame-work with families of frame order distributions $\{h_{\mathbf{x}}(\cdot); \mathbf{x}\in\Omega\}$.

Variance estimation by frame segmentation

The basic idea in the frame segmentation approach is as follows. For a specific study variable **x** it may be difficult to specify a trend structure over the entire frame, but this task may be easier to accomplish "part-wise", if the sampling frame is partitioned into appropriately chosen segments. The segmentation idea lies close to stratification thinking, and from a more formal point of view segmentation can be regarded as a special case of stratification. From a practical point of view, though, there are (at least) the following differences. A stratification, should be carried out before the sample is selected, while segmentation typically is carried out at the estimation stage. Moreover, one may perfectly well use different segmentations for different study variables (while a stratification holds uniformly for the whole survey, with all its study variables). Finally, as is meant to be indicated by the very term "segmentation", a segment should consist of consecutive units (in the frame order), while a stratum could be any subset of the frame. Segmentation and post-stratification are pretty close notions, though. Below we introduce some formal notions, terminology and notation.

> A **segmentation** $\mathcal{H}=(\mathcal{H}_1,\mathcal{H}_2,...,\mathcal{H}_G)$ of the sampling frame $(1,2,...,N)$ is specified by integers $N_1,N_2,...,N_G$ such that $N_g\geq1$, $g=1,2,...,G$ and $N_1+N_2+...+N_G=N$, and the **frame segments** are $\mathcal{H}_1=(1,2,...,N_1)$, $\mathcal{H}_2=(N_1+1,N_1+2,...,N_1+N_2)$, ..., $\mathcal{H}_G=(N_1+...+N_{G-1}+1,...,N)$.

The corresponding **segmentations of the size measure s, ($s_1,s_2,...,s_G$), of the sample size** n, ($n_1,n_2,...,n_G$) and **of the x-sample X, ($X_1,X_2,...,X_G$),** are obtained by the "natural" allocations of the quantities in question to the respective frame segments. In particular, X_g is the collection of x-values which are observed for sampled units from $\mathcal{H}_g$.

Segmented variance estimator are defined below, and their rationale is discussed afterwards.

Let X be an x-sample selected by SYS(n;s) from the frame (1,2,...,N), which is given the segmentation $\mathcal{H}=(\mathcal{H}_1,\mathcal{H}_2,...,\mathcal{H}_G)$. Then a **segmented variance estimator** for the x sample sum, generally referred to by the notation VESS[X;$\mathcal{H}$], has the following structure, where VESS_g denotes a variance estimator for sample sums selected by SYS(n_g;s_g) from $\mathcal{H}_g$;

$$\mathrm{VESS}[X;\mathcal{H}]=\sum_{g=1}^{G}\mathrm{VESS}_g[X_g]. \tag{3.25}$$

To give justification of the segmented variance estimator, we assume in the first round that the frame segmentation in fact is made before the sample selection and that independent samples are selected from the respective segments, a SYS(n_g;s_g) from $\mathcal{H}_g$. With T(X;$\mathcal{H}_g$) denoting the x sample sum in the sample from $\mathcal{H}_g$ we have; T(X)=T(X;$\mathcal{H}_1$)+ T(X;$\mathcal{H}_2$)+ ... +T(X;$\mathcal{H}_G$), and the independence assumption yields; V_x[T(X)]=V_x[T(X;$\mathcal{H}_1$)]+V_x[T(X;$\mathcal{H}_2$)]+... +V_x[T(X;$\mathcal{H}_G$)]. By estimating term-wise in this last relation (3.25) is obtained.

However, in a practical SYS-sampling situation the assumptions that the above reasoning was based on are virtually never exactly fulfilled. In particular, the assumption about independence between the samples from different segments is not fulfilled (at least not in the strict probabilistic meaning of independence), due to the common starting point for all segment samples. Moreover, if the segmentation is carried out when the SYS(n;s) sample is already selected, as is the usual case in practice, the sample sizes ($n_1,n_2,...,n_G$) should be viewed as random quantities. However, they are "ancillary" relative to most estimation objectives. Hence, if one follows the general recommendation for statistical inference: "Make inference conditional on all available ancillary statistics", and makes the inference conditional on ($n_1,n_2,...,n_G$), then one is back in the situation with fixed sample sizes, which was considered in the above "justification". (Rao (1985) gives a fuller discussion of conditional inference in survey sampling.) We let this conclude the "justification" of the segmented variance estimator. Admittedly, a good deal of "hand-waving" was used and, and as always for SYS-sampling, in the end it becomes a matter of judgement whether a procedure should be regarded as sufficiently accurate or not.

The formula (3.25) only provides a general structure, and in the specific situation one also has to decide on which VESS_g-estimators to use in the right hand side of (3.25). In this decision one is "back to scratch" in the sense that the task now is to specify frame order distributions for the different segments. This could be done by a trend model or in terms of total ignorance of frame order, and there is no general requirement to use the same VESS-estimator in all segments. A special case, which is highly relevant for practice, is that the surveyor judges that total ignorance of frame order is a good enough model for each separate segment, although he/she means that a total ignorance description would be inappropriate for the whole sampling frame. In such situations one can say that the problem of trend in the study variable is resolved by a segmentation which leads to **removal of the trend**. (This judgement is typically made if one knows, or believes, that the x-values are fairly homogeneous within segments, but have different mean levels between segments.)

If VESS[· ;a] is applied in each segment, one gets the following estimator;

$$\mathrm{VESS}[X;\mathcal{H};\mathrm{a}]=\sum_{g=1}^{G}\mathrm{VESS}[X_g;a]=\sum_{g=1}^{G}\mathrm{n}_g\cdot(1-\mathrm{n}_g/\mathrm{N}_g)\cdot\mathrm{S}^2(X)\,. \tag{3.26}$$

A further specialization of the estimator in (3.26) is obtained as follows. For simplicity, we assume that the sample size n is an even number. (The estimator is readily modified if this is not the case.) Chose a segmentation $\mathcal{H}^*$ such that exactly two sampled units come from each segment (which does not lead to a unique determination of $\mathcal{H}^*$, though). Then (3.26) takes the following form;

$$\mathrm{VESS}[X;\mathcal{H}^*;\mathrm{a}]=\tfrac{1}{2}\cdot(1-\mathrm{n}/\mathrm{N})\cdot\sum_{r=1}^{n/2}(\mathrm{X}_{2r-1}-\mathrm{X}_{2r})^2\,. \tag{3.27}$$

In the SSYS case, a natural choice of $\mathcal{H}^*$ is to let each segments consist of N/2 units (for simplicity assume that also N is even).

4. On successive pps-sampling.

In this section we shall review and study some aspects of the sampling procedure which usually goes under the name of successive pps-sampling, in particular we shall study variance estimation.

4.1 Some general notions and results

As before we presume a sampling frame which lists the units in the population U=(1,2,...,N). Here the frame order will be irrelevant, though, in contrast to the systematic sampling case. The sampling frame is assumed to carry a distinguished variable $\mathbf{p}=(p_1,p_2,...,p_N)$, referred to as the **draw probability proportionates**, which are assumed to satisfy;

$$p_i > 0,\ i=1,2,...,N, \quad \text{and} \quad \sum_{i=1}^{N} p_i = 1. \tag{4.1}$$

A size n **successive pps unit-sample** from U, selected with draw probability proportionates **p**, referred to by the notation SUC(n;**p**), is drawn as follows.

(i) n units are drawn successively and without replacement from U, and they constitute the unit-sample.

(ii) Given the "remaining" population", each new draw is made independently of previous ones, and so that the probability of selecting a unit which is still in the population is proportional to its **p**-value.

Remark 4.1: a) The assumption (ii) is expressed more formally as follows. Let $J_1(v),J_2(v),...,J_N(v)$ denote the "sample exclusion indicators" after v draws (hopefully a self-explaining term). Then;

$$P\begin{pmatrix}\text{Unit i is drawn} \\ \text{in the v:th draw}\end{pmatrix} = p_i \cdot J_i(v-1) / \sum_{r=1}^{N} p_r \cdot J_r(v-1),\ i=1,2,...,N,\ v=1,2,...,n. \tag{4.2}$$

b) A probabilistically equivalent way to select a SUC(n;**p**) sample runs as follows. (i) Units are drawn from U successively and with replacement until n different units have been selected, and these different units constitute the unit-sample. (ii) At each draw, unit i is selected with probability p_i, i=1,2,...,N, and all draws are made independently of each other. ■

Remark 4.2: When the draw probability proportionates are equal, $p_1=p_2=...=p_N=1/N$, successive pps-sampling is nothing but simple random sampling, i.e. SUC(n;1/N) = SRS(n). ■

Next we review some results on successive sampling, which will play an instrumental role in the sequel. The results are taken from Rosén (1972a,b) where they are formulated as limit theorems, while we here prefer to express them as approximation results. For the validity of the limit theorems, and accordingly for good approximation in the corresponding approximation formulas, certain conditions must be fulfilled. In (4.3)-(4.5) below we formulate the relevant conditions in "pedestrian's versions" which admittedly are somewhat vague, and a reader who wants more details is referred to Rosén (1972). Another possible source, and probably more comprehensible, is Chapter 9 in Hájek (1981).

The sample size n is "fairly large" and the size N of the population is "larger". (The sampling fraction n/N may range over the entire interval (0,1), though.) (4.3)

None of the draw probability proportionates p_i , i=1,2,...,N, is an "outlier" compared with the majority of p-values. (4.4)

For the variable **x**, none of the deviations from its mean, i.e none of the values x_i-$\mu(\mathbf{x})$, i=1,2,...,N, is an "outlier" compared with the majority of deviations. (4.5)

To formulate the approximation results we need some notation. The variables $\delta=(\delta_1,\delta_2,\ldots,\delta_N)$ and $\phi=(\phi_1,\phi_2,\ldots,\phi_N)$ are determined by n and **p** as stated in (4.6) and Definition 4.1 below. When we want to stress the dependence on n and **p** we use notation as $\delta_i(n;\mathbf{p})$ and $\phi_i(n;\mathbf{p})$.

$$\delta_i(n;\mathbf{p}) = n\cdot p_i\ ,\ i=1,2,\ldots,N. \tag{4.6}$$

The variable ϕ, which will be used to approximate inclusion probabilities for SUC(n;**p**), has a somewhat more complicated definition.

DEFINITION 4.1: First determine the number u=u(n;**p**), 0<u≤1, by the relation;

$$\sum_{i=1}^{N} u^{\delta_i(n;\boldsymbol{p})} = N - n\ , \quad n=1,2,\ldots,N\,. \tag{4.7}$$

Then set,

$$\phi_i(n;\boldsymbol{p}) = 1 - u^{\delta_i(n;\boldsymbol{p})}, \quad i=1,2,\ldots,N\,. \tag{4.8}$$

That the above definition always leads to a unique variable $\phi=(\phi_1,\phi_2,\ldots,\phi_N)$ can be realized by the following arguments. View the left hand side in (4.7) as a function f(u) of u, 0≤u≤1. Then, as is readily seen, f(u) is continuous, it increases strictly with u from f(0)=0 to f(1)=N. Hence (4.7) is satisfied for a unique u-value. ■

LEMMA 4.1: Approximate inclusion probabilities. The inclusion probabilities $\pi=(\pi_1,\pi_2,\ldots,\pi_N)$ for SUC(n;**p**) are under (4.3) and (4.4) well approximated as follows;

$$\pi_i \approx \phi_i(n;\mathbf{p})\ ,\ i=1,2,\ldots,N. \tag{4.9}$$

Justification of the claim in Lemma 4.1 can be found in Rosén (1972) or in Hájek (1981), Chapter 9. Next we consider inferential variances of linear statistics. As regards inference, throughout this section we work under the inference model $(\Omega,\{\mathbf{x}^{(n)};\ \mathbf{x}\in\Omega\},\{\mathbf{P}\circ\mathbf{x};\ \mathbf{x}\in\Omega\})$, where **P** is the canonical unit-sample distribution given by SUC(n;**p**). Also her we confine the discussion to sample sums.

LEMMA 4.2: Consider SUC(n;**p**) and let π be the inclusion probabilities and δ and ϕ be according to (4.6) and Definition 4.1. T(**X**) denotes, as usual, the sample sum for the variable x. Assume that (4.3)-(4.5) are satisfied. (i) Then the following variance formula holds with good approximation;

$$V_{\boldsymbol{x}}[T(\boldsymbol{X})] \approx \sum_{i=1}^{N} [x_i - \mu(\boldsymbol{x};\pi;\delta)]^2\cdot\pi_i\cdot(1-\pi_i)\ ,\quad \boldsymbol{x}\in\boldsymbol{\Omega}\,, \tag{4.10}$$

where

$$\mu(x;\pi;\delta) = \sum_{i=1}^{N} x_i \cdot \delta_i \cdot (1-\pi_i) / \sum_{i=1}^{N} \delta_i \cdot (1-\pi_i) . \tag{4.11}$$

(ii) Also the following variance formula holds with good approximation, where $\mu(x;\phi;\delta)$ in (4.12) is specified in (4.13);

$$V_x[T(X)] \approx \sum_{i=1}^{N} [x_i - \mu(x;\phi;\delta)]^2 \cdot \phi_i \cdot (1-\phi_i) , \quad x \in \Omega , \tag{4.12}$$

$$\mu(x;\phi;\delta) = \sum_{i=1}^{N} x_i \cdot \delta_i \cdot (1-\phi_i) / \sum_{i=1}^{N} \delta_i \cdot (1-\phi_i) . \tag{4.13}$$

Justification of (4.10) can be found in Rosén (1972) or in Hájek (1981), Chapter 9. The difference between (4.10) and (4.12) is that the exact inclusion probabilities in (4.10) are changed in (4.12) to the approximate versions given by (4.9). Hence, when (4.10) and (4.9) hold with good approximation so does (4.12) (but probably with a bit less good approximation.)

4.2 Variance estimation

First we introduce a notation, and with regard to its future use we make it a bit elaborate.

DEFINITION 4.2: Q*-statistics. Consider a size n unit-sample from the population U, on which there are specified variables **x**, **a** and **b**. The values of **x**, **a** and **b** for sampled units are denoted by $X=(X_1,X_2,...,X_n)$, $A=(A_1,A_2,...,A_n)$ and $B=(B_1,B_2,...,B_n)$ respectively. First define the statistics D as follows;

$$C_1 = \sum_{v=1}^{n} X_v \cdot \frac{B_v}{A_v} \cdot (1-A_v) , \quad C_2 = \sum_{v=1}^{n} \frac{B_v}{A_v} \cdot (1-A_v) , \quad D = C_1 / C_2 . \tag{4.14}$$

The statistic Q*, fuller denoted by Q*(X;**a**,**b**) or Q*(X;**A**,**B**) according to convenience, is then defined as;

$$Q^*(X;a,b) = Q^*(X;A;B) = \sum_{v=1}^{n} (X_v - D)^2 \cdot (1-A_v) . \tag{4.15}$$

We are now prepared to formulate a preliminary variance estimation result for SUC-sampling.

LEMMA 4.3: Let $X=(X_1,X_2,...,X_n)$ be the sampled **x**-values for a SUC(n;**p**) unit-sample. Assume that (4.3)-(4.5) are satisfied. Then approximately unbiased estimation of $V_x[T(x)]$ is given by Q*(X;**a**;**b**) with **a**=ϕ(n;**p**) and **b**=δ(n;**p**) as specified in Definition 4.1 and (4.6).

Justification: Let $I_1,I_2,...,I_N$ denote the inclusion indicators for the sample. In the first round regard the quantity $\mu(x;\phi;\delta)$ in (4.13) as known, although this will not be the case in a practical situation since $\mu(x;\phi;\delta)$ is a characteristic which depends on the x-values over the entire population (cf. (4.13)). Consider the following sample statistic;

$$R(X) = \sum_{i=1}^{N} [x_i - \mu(x;\phi;\delta)]^2 \cdot (1-\phi_i) \cdot I_i . \tag{4.16}$$

By taking expectation in (4.16) and recalling the relation $E[I_i]=\pi_i$ (see (2.8)) we get;

$$E_x[R(X)] = \sum_{i=1}^{N} [x_i - \mu(x;\phi;\delta)]^2 \cdot (1-\phi_i) \cdot \pi_i, \quad x \in \Omega. \tag{4.17}$$

From (4.17), (4.9) and (4.12) it is seen that $E_x[R(X)] \approx$ the right hand side in (4.12). Hence, in view of (ii) in Lemma 4.2, R(X) yields approximately unbiased estimation of $V_x[T(X)]$. However, as already stated, $\mu(x;\phi;\delta)$ is not known in the practical situation and, hence, R(X) cannot be computed. We take the usual way around this obstacle, i.e. to exchange the unknown population characteristic $\mu(x;\phi;\delta)$ by an estimated version of it. Introduce the following statistics;

$$F(X) = \sum_{i=1}^{N} x_i \cdot \frac{\delta_i}{\phi_i} \cdot (1-\phi_i) \cdot I_i, \quad W = \sum_{i=1}^{N} \frac{\delta_i}{\phi_i} \cdot (1-\phi_i) \cdot I_i. \tag{4.18}$$

Then, again using that $E(I_i)=\pi_i$, we get;

$$E_x[F(X)] = \sum_{i=1}^{N} x_i \cdot \frac{\delta_i}{\phi_i} \cdot (1-\phi_i) \cdot \pi_i, \quad E(W) = \sum_{i=1}^{N} \frac{\delta_i}{\phi_i} \cdot (1-\phi_i) \cdot \pi_i, \quad x \in \Omega. \tag{4.19}$$

From (4.19), (4.9) and (4.13) it is seen that (with good approximation) $E_x[F(X)] \approx$ the nominator in $\mu(x;\phi;\delta)$ and $E(W) \approx$ the denominator in $\mu(x;\phi;\delta)$. From this we conclude that F/W yields approximately unbiased estimation of $\mu(x;\phi;\delta)$. It is readily checked, and we leave the details to the reader, that exchange of $\mu(x;\phi;\delta)$ for its estimator F/W in the statistic R leads to the estimator $Q^*(X;\phi(n;\mathbf{p});\delta(n;\mathbf{p}))$, which thereby is justified as an approximately unbiased estimator of $V_x[T(X)]$. ■

As stated in Remark 4.2, SUC(n;1/N) equals SRS(n). Hence, Lemma 4.3 provides a possible variance estimator also for SRS(n). It is readily checked, and we leave the details to the reader, that for SUC(n;1/N)=SRS(n) we have $\phi(n;\mathbf{p})=\delta(n;\mathbf{p})=n/N$ and, hence, that the variance estimator becomes;

$$Q^*(X;(n/N)\cdot \mathbf{1};(n/N)\cdot \mathbf{1}) = (n-1)\cdot (1- n/N)\cdot S^2(X). \tag{4.20}$$

From (4.20) and (2.20) it is seen that the above complicated way of deriving a variance estimation procedure for SRS in fact leads to an estimator which lies close to the traditional one in (2.20). There is a slight discrepancy, though, amounting to the factor n/(n-1). With this observation as background we introduce an adjusted (admittedly a bit ad hoc) version of the estimator. First we introduce a modification of the statistic Q*. **The statistic** Q is defined as follows, where $Q^*(X;\mathbf{a};\mathbf{b})$ is according to Definition 4.2;

$$Q(X;\mathbf{a};\mathbf{b}) = Q^*(X;\mathbf{a};\mathbf{b})\cdot n/(n-1). \tag{4.21}$$

The following estimator VESS[X;SUC] will be regarded as the "basic" one for SUC-sampling.

THEOREM 4.1: Consider SUC(n;**p**) and let $\mathbf{a}=\phi(n;\mathbf{p})$ and $\mathbf{b}=\delta(n;\mathbf{p})$ be as specified in Definition 4.1 respectively (4.6). Let $X=(X_1,X_2,\ldots,X_n)$ be the sampled values for the variable x. Assume that (4.3)-(4.5) are satisfied. Then approximately unbiased estimation of $V_x[T(X)]$ is provided by

$$VESS[X;SUC] = Q(X;\phi(n;\mathbf{p});\delta(n;\mathbf{p})). \tag{4.22}$$

This concludes our study of SUC-sampling.

5. Basic results on variance estimation for systematic pps-sampling by approximation with successive pps-sampling

In this section we carry out the basic part of the program which was outlined in Section 1, namely to employ (1.2) as a basis for derivation of variance estimators in conjunction with systematic pps-sampling. The approximation idea in (1.2) is restated below in a somewhat different wording.

> Inference from a SYS(n;**s**) sample under total ignorance of frame order is approximately equivalent with inference from a SUC(n;**p**) sample with an appropriately chosen **p**. (5.1)

Also in this section we confine the discussion to the standardized linear statistic, i.e. the sample sum. In (5.2) below we state a more special, and more technical version of (5.1).

> An estimator of $V_x[T(X)]$ under SYS(n;**s**) with total ignorance of frame order is obtained by a (good) estimator for $V_x[T(X)]$ under a SUC(n;**p**) inference model, with an appropriately chosen **p**. (5.2)

When implementing (5.2), we shall use the SUC variance estimator in Theorem 4.1, i.e an estimator of the type Q(X;**a**;**b**) as defined by Definition 4.2 and (4.21). The problem is to chose **a** and **b**, or more generally formulated, to answer the following question which is raised by (5.1) and (5.2). For which **p**'s can SUC(n;**p**) be expected to give "appropriate" approximation of SYS(n;**s**) with total ignorance of frame order? The following idea for an answer is near at hand.

> Approximate SYS(n;**s**) with total ignorance of frame order by the SUC(n;**p**) which has draw probability proportionates **p** equal to the standardized version **z**(**s**) of the size measure **s** (see (3.2)), i.e. approximate by SUC(n;**z**(**s**)). (5.3)

Under the approximation mode (5.3), **a** and **b** in Q(X;**a**;**b**) shall be chosen as;

$$a_i = \phi_i(n;z(s)), \quad b_i = \delta_i(n;z(s)), \quad i=1,2,\ldots,N, \tag{5.4}$$

where ϕ and δ are according to (4.6) and Definition 4.1. This leads to the following variance estimator (and we continue on the labelling a,b,c which was used in Section 3.4.1);

$$\mathrm{VESS}[X;\mathrm{d}] = Q(X;\phi(n;z(s));\delta(n;z(s))). \tag{5.5}$$

As will be seen in Section 7, VESS[X;d] works quite satisfactorily in many cases, but in some types of situations it does not perform well. An explanation for this sometimes bad performance is given by the fact that SUC(n;z(s)) in fact may approximate SYS(n;s) quite badly in the sense that the two procedures may have considerably different inclusion probabilities. By virtue of (3.3) and Lemma 4.1 the following relation holds, and "non-inequality" occurs only exceptionally in it;

$$\pi_i[\mathrm{SYS}(n;s)] = n\cdot z_i(s) \neq 1-u(n;z(s))^{n\cdot z_i(s)} \approx \pi_i[\mathrm{SUC}(n;z(s)]\,, \quad i=1,2,\ldots,N\,. \tag{5.6}$$

The relative difference between the inclusion probabilities in (5.6) is often small but it can be considerable, as is illustrated in the following example.

Example 5.1: Let U=(1,2,...,N) be a population, **z** a standardized size measure and n an integer. Assume that for some unit in the population, say unit 1, we have $z_1 = 1/n$. Then, for SYS(n;**z**), $\pi_1 = n\cdot z_1 = 1$, which means that unit 1 will be included in a SYS(n;**z**) sample with probability 1. However, for SUC(n;**z**) the inclusion probability for unit 1 is less than 1, since the unit may be "missed" in each of the n successive draws (without replacement). If N is large and the sampling fraction is small we have $\pi_1 \approx 1-1/e = 0.37$ (Lemma 4.1 provides more details.) Hence, inclusion probabilities for SYS(n;s) and SUC(n;**z**(s)) may differ considerably.

The above situation is perhaps a bit extreme to the effect that it involves inclusion with probability 1. However, it is readily realized that the phenomenon, that one (or more) units have much higher inclusion probabilities for SYS(n;**z**) (but less than 1) than for SUC(n;**z**), occurs as soon as some z_i lies close to 1/n, even if it lies a bit below. (Since the sum of the inclusion probabilities equals the sample size n in both cases, other inclusion probabilities must be ordered in the opposite way.) ■

With the above discussion as background, the following alternative approximation mode sounds more appealing than the one in (5.3).

> Approximate SYS(n;s) with total ignorance of frame order by the SUC(n;**p**) which has the same inclusion probabilities as SYS(n;s). (5.7)

Then the problem is to find a **p** which makes (5.7) satisfied. Strict fulfilment of (5.7), i.e. to find a **p** such that π_i[SUC(n;**p**)]=π_i[SYS(n;s)], i=1,2,...,N, is impossible for general n and s, at least from a practical point of view. The reason for this is that no numerically manageable formula for SUC inclusion probabilities is available. Therefore, we shall consider a "next best" solution, by seeking a **p** which yields equality between the inclusion probabilities for a given SYS(n;s) and the approximate SUC inclusion probabilities given by Lemma 4.1. By (3.3) this means that we want to solve for **p**, **p**=**p**(n;s), in the following equation (system); ϕ(n;**p**)= n$\cdot$ **z**(s), which by virtue of (4.8) has the following explicit form;

$$n \cdot z_i(s) = 1 - u^{n \cdot p_i(n;s)}, \quad i=1,2,\ldots,N. \tag{5.8}$$

By taking logarithms (with an arbitrary base) in (5.8) we get;

$$p_i(n;s) = \frac{\log(1 - n \cdot z_i(s))}{n \cdot \log u}, \quad i=1,2,\ldots,N. \tag{5.9}$$

Now recall (4.1), which states that probability proportionates should sum to 1. By applying this condition to (5.9) we arrive at the following p_i's, which have the pleasant feature that u has disappeared;

$$p_i(n;s) = \log(1 - n \cdot z_i(s)) / \sum_{r=1}^{N} \log(1 - n \cdot z_r(s)), \quad i=1,2,\ldots,N. \tag{5.10}$$

Hence, the strategy (5.7) leads to the variance estimator Q(X;**a**;**b**) which is obtained by setting **a**=n$\cdot$ z(s) and **b**=n$\cdot$ **p**(n;s), with **p** according to (5.10), and the resulting estimator is denoted

$$\text{VESS}[X;e] = Q(X;\ n\cdot \mathbf{z}(s);\ n\cdot \mathbf{p}(n;s)), \tag{5.11}$$

where Q is given by (4.21) and Definition 4.2. This is the estimator which we want to recommend for practical use, and therefore its computation is written out in full in Algorithm 5.1 below. Justification for the algorithm is given after its formulation. The main reasons for our preference for this estimator, which will be discussed more fully in Section 7, are in brief the following. We strongly believe that VESS[X;e] is good as regards estimation accuracy and in addition to that it is computationally simple.

ALGORITHM 5.1. Computation of VESS[· ;e]: Let $X=(X_1,X_2,\ldots,X_N)$ be the observed x-values for a SYS(n;s) sample. Then VESS[X;e] in (5.11) is computed by the following steps (i)-(iii).

(i) Compute the standardized size measure **z**(s) by (5.12). Henceforth let $Z_1,Z_2,\ldots,Z_n$ denote the **z**-values for the sampled units.

$$z_i(s) = s_i / \sum_{r=1}^{N} s_r, \quad i=1,2,\ldots,N. \tag{5.12}$$

(ii) Compute

$$C_1 = \sum_{v=1}^{n} X_v \cdot \frac{\log(1-n\cdot Z_v)}{Z_v} \cdot (1-n\cdot Z_v), \quad C_2 = \sum_{v=1}^{n} \frac{\log(1-n\cdot Z_v)}{Z_v} \cdot (1-n\cdot Z_v), \tag{5.13}$$

$$D = C_1/C_2. \tag{5.14}$$

(iii) Finally set;

$$\text{VESS}[X;e] = \frac{n}{n-1} \cdot \sum_{v=1}^{n} (X_v - D)^2 \cdot (1-n\cdot Z_v). \tag{5.15}$$

Justification of Algorithm 5.1: The task is to give an explicit computation formula for the right hand side in (5.11), using Definition 4.2 and (4.21). The following observation admits simplifications. The Q*-statistic in Definition 4.2 is "homogeneous in **b**" in the sense that Q*(X;**a**;k· **b**) = Q*(X;**a**;**b**) for any k≠0. This follows readily from (4.14) and the fact that the **b**-values come in only in $D=C_1/C_2$. From (4.21) follows that Q has the same homogeneity property. In our case **b** is given as n· **p**(n;s), where **p**(n;s) is defined in (5.10). By the homogeneity we can change **b** to b_i=n· log(1-z(s)$_i$), i=1,2,...,N, which leads to the algorithm. ■

Remark 5.1: As the variance estimator VESS[X;d] in (5.5) also is based on the Q-statistic in (4.21) it is computed by an algorithm which is quite analogous to Algorithm 5.1. The computation of VESS[X;d] requires some non-trivial preliminary computation efforts, though. To start with one has to compute the ϕ(n;z)-values by Definition 4.1, the crucial part being to solve the equation (system) (4.7) for u. In our experience this is quite feasible by using iterative methods, but it does require some computational efforts. We leave the details to the reader. ■

Remark 5.2: It is readily checked that VESS[X;d] and VESS[X;e] agree for SSYS (which under total ignorance of frame order agrees with SRS), and then both variance estimators lead to the "correct" estimator $n\cdot(1-n/N)\cdot S^2(X)$. Hence, to experience difference between VESS[X;d] and VESS[X;e] one has to consider situations with "genuine" systematic pps-sampling (i.e. "non-SSYS" cases). In Section 7 we consider such cases. ■

Thereby we have advised on a variance estimator for SYS(n;s) under total ignorance of frame order. As regards situations where the study variable has frame order trend, we confine in this paper to "trend removal" by segmentation (see Section 3.4.2). In line with the previous general recommendation of VESS[X;e] under total ignorance of frame order, the following segmented estimator is recommended in cases where one knows of a segmentation which leads to (or is believed to lead to) segment-wise total ignorance of frame order;

$$\text{VESS}[X;\mathcal{H};e] = \sum_{g=1}^{G} \text{VESS}[X_g;e]. \tag{5.16}$$

6. On practical application of the SYS variance estimation method

In this section we shall consider various issues of relevance for practical applications of the SYS(n;s) variance estimation approach, which was introduced in the previous section. In Sections 4 and 5 we were "unpractical" to the effect that we considered the standardized linear statistic, the sample sum, while the most interesting linear statistic in practice is the HT-estimator. Variance estimation for HT-estimators is treated in Section 6.1. The ever-present problem in practice, non-response, is commented upon in Section 6.2. Variance estimation under measurement errors is discussed in Section 6.3. The results there have independent interest, and they also prepare for the considerations in Section 6.4, which concern stage-wise sampling with systematic pps-sampling in the first stage.

6.1 Variance estimation for Horvitz-Thompson estimators

We start with the HT-estimator of a population total $\tau(x)$. Its appearance in the SYS(n;s) case is stated in (3.12). As is pointed at by (2.16) and (2.13), the HT-estimator for a population total can be viewed as the sample sum for the transformed variable $y=x/\pi$, which in the SYS(n;s) case is $y=x/[n\cdot z(s)]$. Hence, Algorithm 5.1 can be applied to $\hat{\tau}(x)_{HT}=T(x/[n\cdot z(s)])$, and no principally new problems turn up. However, because of its practical interest we make the effort to write down the special case of Algorithm 5.1 which corresponds to the HT-estimator for a population total. The checking is left to the reader.

ALGORITHM 6.1: Variance estimation for the HT-estimator of a population total: Let $X=(X_1,X_2,...,X_n)$ be the observed x-values for a SYS(n;s)-sample. Let z(s) and $(Z_1,Z_2,..,Z_n)$ be as stated in Algorithm 5.1. Then the HT-estimator for the population total $\tau(x)$ is;

$$\hat{\tau}(x)_{HT}=\frac{1}{n}\cdot\sum_{v=1}^{n}\frac{X_v}{Z_v}. \qquad (6.1)$$

Under total ignorance of frame order for the variable x/s, the inferential variance of the HT-estimator in (6.1) is estimated by the steps (ii) and (iii) below.

(ii) Compute

$$C_1=\frac{1}{n}\cdot\sum_{v=1}^{n}\frac{X_v\cdot\log(1-n\cdot Z_v)}{Z_v^2}\cdot(1-n\cdot Z_v),\quad C_2=\sum_{v=1}^{n}\frac{\log(1-n\cdot Z_v)}{Z_v}\cdot(1-n\cdot Z_v), \qquad (6.2)$$

$$D=C_1/C_2. \qquad (6.3)$$

(iii) Finally set

$$\hat{V}_x[\hat{\tau}(x);e]=\frac{n}{n-1}\cdot\sum_{v=1}^{n}[\frac{X_v}{n\cdot Z_v}-D]^2\cdot(1-n\cdot Z_v). \qquad (6.4)$$

Variance estimation for the more general HT-estimator in (2.14), of a ratio of totals, can be handled very much along the same lines. Take as point of departure is the approximate variance formula (2.19). As is seen, the middle and right hand parts of (2.19) concern variances of linear statistics and Algorithm 5.1 comes into mind. A well-known obstacle turns up, though, namely that the characteristics $\mu(x)$ and $\mu(y)$ in (2.19) are not known. There is also a well-known (and "usual") way around this obstacle, namely the following. Replace $\mu(x)$ and $\mu(y)$ by their estimates $\hat{\mu}_{HT}(x)$ and $\hat{\mu}_{HT}(y)$, and then proceed as if the estimated values of $\mu(x)$ and $\mu(y)$ were the true ones. With this modification, Algorithm 5.1 can be set to work. We leave the details to the reader. We adopt the following **notation convention** as regards estimators of the variance for HT-estimators. A variance estimator for the HT-estimator $\hat{\chi}(\mathbf{x})_{HT}$ which is based on VESS[· ;γ] is referred to by the notation $\hat{V}[\hat{\chi}(\mathbf{x})_{HT};\gamma]$.

6.2 Variance estimation when non-response occurs

Here we make a detour from the general assumption in Section 2 that all sampled units respond in the observation phase. It is a well-known (and unpleasant) fact that some non-response occurs in virtually every practical sample survey. Then one cannot exhibit unbiased estimators, neither point nor variance estimators, unless assumptions are made about the "mechanism" for generation of the non-response. It is also a well-known fact that, whatever assumptions are made, they are pretty difficult to verify. The following assumption is a standard one in this context (and it is often as good as any other).

> **MODEL 6.1 for non-response:** Chance determines whether a unit, which is taken out for observation, responds or not. As regards responding, units act independently of each other, and all units in the population have the same response propensity.

The following result belongs to the "folk-lore".

> **Estimation under Model 6.1:** Consider inference on **x**-characteristics based on SYS(n;s) with total ignorance of frame order for **x**. Non-responses may occur, and Model 6.1 is assumed to hold, at least for the variable **x**. Let n' denote the **number of responding units**. Then the inference from the **x**-values for the responding units should be carried out as for SYS(n';s).

As regards variance estimation, the practical consequence of the above statement is that VESS[· ;e] can be applied also when non-response occurs, provided Model 6.1 can be assumed to be realistic. The modification needed is that the sample size n should be substituted by the number n' of responding units.

6.3 Variance estimation when the observations are afflicted by measurement errors

Here we make another detour from the general assumption in Section 2, and allow for the possibility that measurement errors occur. Then we have to modify the previous inference model, which is done below. Note, though, that the following model has quite special features, and a variety of other models are possible in the measurement error context.

> **MODEL 6.2 for sampling and measurements:** The inference interest concerns **x**-characteristics where, as usual, $\mathbf{x}=(x_1,x_2,...,x_N)$ is a variable on the population $U=(1,2,...,N)$. The **apriori possible x-values** are, as before, denoted by Ω. An ordered, size n unit-sample is selected without replacement from U. Observations are made on the sampled units and the measurement for unit i, provided it is sampled, is generated as $y_i=x_i+\varepsilon_i$, $i=1,2,...,N$, where the **measurement errors** $\varepsilon_1,\varepsilon_2,...,\varepsilon_N$ are viewed as random variables. The observed values in their sample order are denoted by $\mathbf{Y}=(Y_1,Y_2,...,Y_n)$. To set up a model for the distribution of **Y** we make the following assumptions.
>
> The sampling and measurement processes are independent of each other, i.e. the inclusion indicators $(I_1,I_2,...,I_N)$ and the measurement errors $(\varepsilon_1,\varepsilon_2,...,\varepsilon_N)$ are independent random vectors. The distribution of $(I_1,I_2,...,I_N)$ is, as before, modeled by a family $\{\mathbf{P}_\mathbf{x};\ \mathbf{x}\in\Omega\}$ of unit-sample distributions. The measurement errors $(\varepsilon_1,\varepsilon_2,...,\varepsilon_N)$ are assumed to be mutually

independent, and for each one of them there is specified a distribution family in which **x** enters as parameter. The following relation (6.5) is assumed to be satisfied for a set of parameters $(\sigma_1^2,\sigma_2^2,...,\sigma_N^2)$, referred to as the **measurement variances.** A consequence of (6.5) is that the distribution of ε_i in fact is assumed not to depend on **x** as regards it mean and variance;

$$E_x(\varepsilon_i) = 0 \text{ and } V_x(\varepsilon_i) = \sigma_i^2 \text{ , } i=1,2,...,N. \tag{6.5}$$

It should be clear that the above assumptions in fact specify the joint distribution of $(I_1,I_2,...,I_N,\varepsilon_1,\varepsilon_2,...,\varepsilon_N)$. In particular, $\mathbf{x}\in\Omega$ is the relevant parameter for the family of possible joint distributions.

Remark 6.1: Note that Model 6.2 is invariant under transformation of the variable **x** to the following effect. If the model holds for **x**, and **a** is a variable which can be observed without measurement errors, then the model holds also for **x/a.** ■

In Model 6.2 we do not make any specific assumptions about the nature of the unit-sample distribution $\mathbf{P_x}$, more than that it should be an ordered, size n sample drawn without replacement. Before entering our special concern, SYS-sampling, we shall, for future use, consider a result which holds without additional assumptions to Model 6.2. In the following we let $X=(X_1,X_2,...,X_n)$ denote the x-values for the sampled units (which are not observed, though). Moreover, T(**Y**) and T(**X**) as usual denote the following sums;

$$T(Y) = \sum_{i\in U} y_i\cdot I_i = \sum_{v=1}^{n} Y_v \text{ , } \quad T(X) = \sum_{i\in U} x_i\cdot I_i = \sum_{v=1}^{n} X_v \text{ .} \tag{6.6}$$

The contents of the following lemma is certainly well-known but we prove it, partly for the sake of completeness and partly as a preparation for future considerations.

LEMMA 6.1: Under Model 6.2 the following relation holds;

$$V_x[T(Y)] = V_x[T(X)] + \sum_{i=1}^{N} \sigma_i^2\cdot\pi_i \text{ , } \quad x\in\Omega \text{ .} \tag{6.7}$$

Proof: Recall the relation $y_i=x_i+\varepsilon_i$, i=1,2,...,N. Let E_x^I denote conditional expectation given the outcome of the unit-sample, i.e. given $I=(I_1,I_2,...,I_N)$. For notational simplicity we will mostly suppress the subscript **x**, though. By taking iterated expectation $E_x=E_xE^I$ in (6.9) and its squared version and employing well-known rules for conditional expectation we get;

$$E_x[T(Y)] = E_x[\sum_{i=1}^{N} x_i\cdot I_i] + E_x[\sum_{i=1}^{N} I_i\cdot E^I(\varepsilon_i)] \text{ , } \quad x\in\Omega \text{ ,} \tag{6.8}$$

$$\begin{aligned} E_x[T(Y)^2] &= E_x[(\sum_{i=1}^{N} (x_i+\varepsilon_i)\cdot I_i)^2] = E_x[(\sum_{i=1}^{N} x_i\cdot I_i)^2] + \\ &+ E_x[\sum_{i=1}^{N}\sum_{j=1}^{N} I_i\cdot I_j\cdot E^I(\varepsilon_i\cdot\varepsilon_j)] + 2\cdot E_x[\sum_{i=1}^{N}\sum_{j=1}^{N} x_i\cdot I_i\cdot I_j\cdot E^I(\varepsilon_j)] \text{ , } \quad x\in\Omega \text{ .} \end{aligned} \tag{6.9}$$

By the independence of $(I_1,I_2,...,I_N)$ and $(\varepsilon_1,\varepsilon_2,...,\varepsilon_N)$ and (6.5) the following relations hold; $E^I(\varepsilon_i)=E(\varepsilon_i)=0$ and $E^I(\varepsilon_i^2)=E(\varepsilon_i^2)=\sigma^2_i$, i=1,2,...,N, and under the assumption about mutual independence between the ε's we have; $E^I(\varepsilon_i\cdot \varepsilon_j)=E(\varepsilon_i\cdot \varepsilon_j)=0$ for $i\neq j$. By employing these relations, (6.8) and (6.9) can be simplified. Next, combine the simplified formulas with the general formula $V(T)=E(T^2) - [E(T)]^2$, and (6.7) results. ■

We now specialize to SYS(n;s). Moreover, we return to considering only sample sums, again with "simplicity of formulas" as reason. We leave to the reader to carry out the necessary modifications to obtain corresponding results for general linear statistics.

LEMMA 6.2: Consider inference for SYS(n;s) under total ignorance of the frame order for x, and assume that Model 6.2 is in force. The Z's are as in Algorithm 5.1. Moreover, assume that there are statistics $(H_1,H_2,...,H_N)$ which yield (approximately) unbiased estimation of the measurement variances, i.e. that $E(H_i) = \sigma_i^2$, i=1,2,...,N, and that these statistics are available for sampled units. Then VESS[Y;f] in (6.10) below yields approximately unbiased estimation of $V_x[T(Y)]$;

$$\mathrm{VESS}[Y;f] := \mathrm{VESS}[Y;e] + \sum_{v=1}^{n} n \cdot Z_v \cdot H_v . \qquad (6.10)$$

Remark 6.2: The formula (6.10) exhibits an unfortunate notational awkwardness, which may cause confusion. A VESS[·] is always a **statistic**, which can be applied to a set of observed values. If certain conditions are met, the statistic VESS[X] yields "decent" estimation of $V_x[T(X)]$. However, under other conditions VESS[X] may in fact be a bad estimator for $V_x[T(X)]$. What is claimed in Lemma 6.2 is, that under the conditions stated in the lemma, VESS[Y;f] yields decent estimation of $V_x[T(Y)]$, while VESS[Y;e] does not in general. ■

Remark 6.3: The claim in Lemma 6.2 of course still holds if VESS[X;e] is substituted by a statistic which is approximately equal to it. ■

Justification of Lemma 6.2: The following is a heuristic justification, rather than a strict proof. According to Algorithm 5.1 we have;

$$\mathrm{VESS}[Y;e] = \frac{n}{n-1} \cdot \sum_{i=1}^{N} (y_i - D(Y))^2 \cdot (1-\pi_i) \cdot I_i , \qquad (6.11)$$

where we use the notation D(Y) to stress that D is a statistic which depends on the observations $Y=(Y_1,Y_2,...,Y_n)$. Make the following operations in (6.11). (i) Substitute y_i by $x_i+\varepsilon_i$, in accordance with Model 6.2. (ii) Substitute D(Y) by D(X), and regard that as an "acceptable" approximation, with justification provided by the formulas (5.13) and (5.14) and the fact that the ε's have zero means. (iii) Expand the square. (iv) Approximate n/(n-1) by 1. The listed operations (i)-(iv) together with (5.15) lead to the following approximate relation;

$$\mathrm{VESS}[Y;e] \approx \mathrm{VESS}[X;e] + \sum_{i=1}^{N} \varepsilon_i^2 \cdot (1-\pi_i) \cdot I_i + 2 \cdot \sum_{i=1}^{N} (x_i - D(X)) \cdot \varepsilon_i \cdot (1-\pi_i) \cdot I_i . \qquad (6.12)$$

Now take expectation in (6.12), for the summations as term-wise iterated expectation $E_x E^I$ (cf the proof of Lemma 6.1), and allow for the approximation to regard D(X) as a constant. By using the previously listed formulas $E^I(\varepsilon_i)=0$ and $E^I(\varepsilon_i^2)=\sigma_i^2$, we get;

$$E_x(\mathrm{VESS}[(Y;e]) \approx E_x(\mathrm{VESS}[X;e]) + \sum_{i=1}^{N} \sigma_i^2 \cdot (1-\pi_i) \cdot \pi_i . \qquad (6.13)$$

Under the conditions of Lemma 6.2, VESS[X;e] yields approximately unbiased estimation of $V_x[T(X)]$. By inserting this into (6.13) and by comparing with (6.7), we obtain;

$$E_x(\mathrm{VESS}[Y;f]) = V_x[T(Y)] - \sum_{i=1}^{N} \sigma_i^2 \cdot \pi_i^2 . \qquad (6.14)$$

Furthermore, by (3.3) and the assumptions on the H's we have;

$$E_x[\sum_{v=1}^{n} n \cdot Z_v \cdot H_v] = E_x[\sum_{i=1}^{N} n \cdot z_i \cdot H_i \cdot I_i] = \sum_{i=1}^{N} \sigma_i^2 \cdot \pi_i^2 . \tag{6.15}$$

Combination of (6.14) and (6.15) now yields the claim in Lemma 6.2. ■

Remark 6.4: Under additional assumptions, the summation term in (6.10) can be neglected, and hence VESS[Y;e] can be used as an estimator for $V_x[T(Y)]$. Note that the estimator VESS[Y;e] has the pleasant feature that it does not require any estimates of the measurement variances. From (6.15) it is seen that the summation term in (6.10) can be omitted if the quantity $\sigma^2_1 \cdot \pi_1^2 + \ldots + \sigma^2_N \cdot \pi_N^2$ is small compared with "surrounding" terms.

One situation where this holds is if all π_i's are small compared with 1, because then $\sigma^2_1 \cdot \pi_1^2 + \ldots + \sigma^2_N \cdot \pi_N^2$ is small compared with $\sigma^2_1 \cdot \pi_1 + \ldots + \sigma^2_N \cdot \pi_N$, which is part of the right hand side in (6.16), and hence part of E_x(VESS[Y;e]). (To a large extent this finding is only a complicated way of saying that when the sampling fraction is small, one can estimate $V(Y_1+Y_2+\ldots+Y_n)$ as if $Y_1, Y_2, \ldots, Y_n$ were independent observations.)

Another situation where the summation term in (6.10) can be omitted, which also covers cases with non-negligible sampling fractions, is when $\sigma^2_1 \cdot \pi_1^2 + \ldots + \sigma^2_N \cdot \pi_N^2$ is small compared with $V(X_1+X_2+\ldots+X_n)$, as is seen from (6.13). This type of situation occurs in particular (and not too surprisingly) if all σ_i^2 are small, i.e. if all measurement errors are small. ■

6.4 On variance estimation at stage-wise sampling with SYS(n;s) in the first stage

Here we shall apply the results of the previous sub-section to situations with stage-wise sampling with SYS(n;s) in the first stage. More specifically we shall consider a version of the well-known variance estimation approach which often is referred to as the **method of ultimate clusters**. We start with some terminology and notation.

The population is assumed to be hierarchically clustered; The study units are (according to some rule) grouped into mutually exclusive clusters, which in turn are grouped into mutually exclusive clusters on a higher level, which in turn ..., etc. The process eventually comes to an end, and then one has reached the "ultimate clusters". A stage-wise sampling procedure relies on this type of hierarchical structure, but goes in the "opposite direction". We adhere to the usual way of naming the clusters, namely by their role in the sampling process. The first stage sample of PSUs (= primary sampling units) is a sample of ultimate clusters. From sampled PSUs subsamples of SSUs (= secondary sampling units) are selected, etc. Finally one comes to samples of USUs (= ultimate sampling units), which are the basic units for observation. Below we specify the vital assumptions about the sampling process.

(i) The sampling process is hierarchical in the sense that only clusters which are sampled on one level are subsampled on the next level.

(ii) Given the clusters which have been sampled on one level, the subsamples on the next level are drawn independently of each other and independently of what happened in previous sampling stages.

Let i index the PSUs/ultimate clusters, i=1,2,...,N. Moreover, let

Y_i = a statistic which can be computed from observations made on sampled units (on any level) which are reached via the i:th PSU, and only from such observations, $i=1,2,..,N$. (6.16)

Assume that the sampling procedure starts with a sample of PSU's of size n. Let, as usual, $T(Y)=Y_1+Y_2+...+Y_n$ be the sample sum of Y's. When stage-wise sampling is employed, an estimator of a population total typically has this sum structure, the Y's being "HT-inflated" versions of estimators of PSU totals. (The "HT-modification" is left to the reader.) The task is, as usual, to derive an estimator for $V_x[T(Y)]$. Consider the following partitioning;

$$Y_i = E[Y_i] + (Y_i - E[Y_i]) =: x_i + \varepsilon_i\, ,\ i=1,2,...,N. \tag{6.17}$$

It is readily checked that the above assumptions (i) and (ii) entail that Model 6.2 applies to (6.17). Hence, if one for sampled PSU's can exhibit unbiased estimators $\hat{V}[Y_i]$ of $\sigma_i^2=V[Y_i]$, Lemma 6.2 can be applied to yield the following estimator, which we call an **ultimate clusters variance estimator**;

$$\mathrm{VESS}[Y;\mathrm{UK};e] = \mathrm{VESS}[Y;e] + \sum_{v=1}^{n} n\cdot Z_v\cdot \hat{V}[Y_v]\,. \tag{6.18}$$

In fact, usually one can exhibit estimators of $V[Y_i]$, which usually also are independent of each other, due to the fact that they are based on observations from stages which are subsequent to the first one.

Remark 6.5: The "classical" ultimate clusters estimator is obtained by using either of the estimators VESS[Y;a] or VESS[Y;b]. ■

7 Numerical illustrations, discussion and conclusions

7.1 Introduction

So far our investigations have been quite theoretical, but we have employed different types of approximation arguments at various stages of the reasoning. Therefore, on the basis of the hitherto considerations one cannot give clear answers to questions about the performance of VESS[· ;e] and VESS[· ;d] in practical applications. To get good insight on that matter one has to carry out well designed simulation experiments. A first step on the road of simulation evaluations is reported on in Section 7.2. However, these simulations certainly do not cover all types of situations that may turn up in practice, since there is such a great number of "parameters" which may be varied; the population size, the sampling fraction, the distribution of study variables, the distribution of the size measure variable and the structure of its covariation with the study variables, to mention the most pertinent ones.

In Section 7.2 we report on simulation findings which enable comparison of the performances of the new variance estimators VESS[· ;e] and VESS[· ;d] with "old" ones; VESS[· ;a], VESS[· ;b] and VESS[· ;c], and which also exhibit results for (estimated) true variances. The comparisons concern expected values of the estimators as well as the confidence level for confidence intervals based on them. The comparisons mainly relate to situations with total ignorance of frame order, but we also consider a situation with frame trend in the study variable.

As was discussed in Section 2, the success of a confidence interval $K\pm\sqrt{Q}$ depends not only on the unbiasedness of the employed variance estimator, approximate normality is also required. We comment on that matter in Section 7.3.

In Section 7.4 finally, we formulate our conclusions concerning the practical applicability of the methods which are presented in this paper, the main one being that we mean there are good grounds to employ VESS[· ;e] in practical use, notably in cases with high sampling rates.

7.2 Simulation results

In this sub-section we report on simulation findings which relate to two different types of situations. The first one, in 7.2.1, concerns the behaviour of the VESS estimators, notably that of VESS[· ;e] and VESS[· ;d], in situations with total ignorance of frame order. The second one, in 7.2.3, is meant to illustrate the need for alternative estimators in situations with trend in the study variable.

7.2.1 Variance estimation under total ignorance of frame order

The main interest of this paper concerns situations with large sampling fractions. To achieve such situations under good "simulation economy" we chose a fairly small population. The "unordered" version of the population (cf Section 3.1) and the variables on it were chosen in the following way. The population size was set to N=100, and the **size measure variable s** was selected as follows; $s_{(i)}=1$ for $1 \le (i) \le 25$, $s_{(i)}=2$ for $26 \le (i) \le 50$, $s_{(i)}=3$ for $51 \le (i) \le 75$ and $s_{(i)}=4$ for $76 \le (i) \le 100$. To generate the "main" **study variables x** and **y** we started by introducing two "auxiliary" variables **x*** and **y*** as follows;

Values for the variable x* were generated as outcomes of independent normal random variables with mean 10 and standard deviation 2, (7.1)

Values for the variable y* were generated as outcomes of independent random variables with uniform distributions; for units with s=1 on the interval [0,10], for s=2 on the interval [5,20], for s=3 on the interval [10,30] and for s=4 on the interval [20,45]. (7.2)

The values of x and y were then set to;

$$x_{(i)} = s_{(i)} \cdot x^*_{(i)}, \quad y_{(i)} = s_{(i)} \cdot y^*_{(i)}\ , \quad 1 \leq (i) \leq 100. \tag{7.3}$$

Hence, in all there are four variables on the population; x, y, x* and y*. The reason for the "duplication" of variables is that HT-estimators relative to x and y correspond to sample sums relative to x* and y*, as is readily checked. We find it instructive to be able to alternate between the two views; "sample sum" respectively "HT-estimator". The distributions of the variables over the whole population are presented graphically in the Figures 7.1 - 7.4.

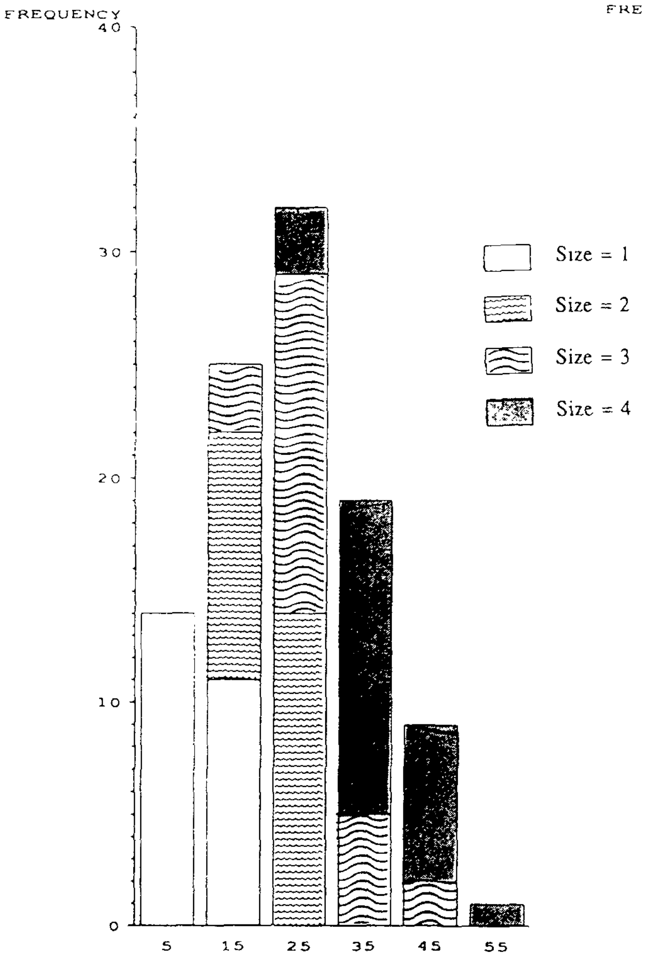

Figure 7.1. Historgam for the variable x.

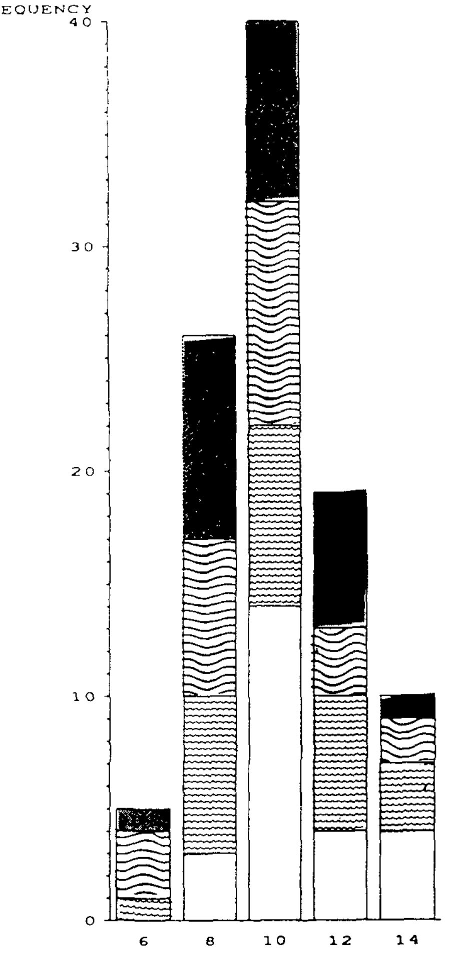

Figure 7.2. Historgam for the variable x*.

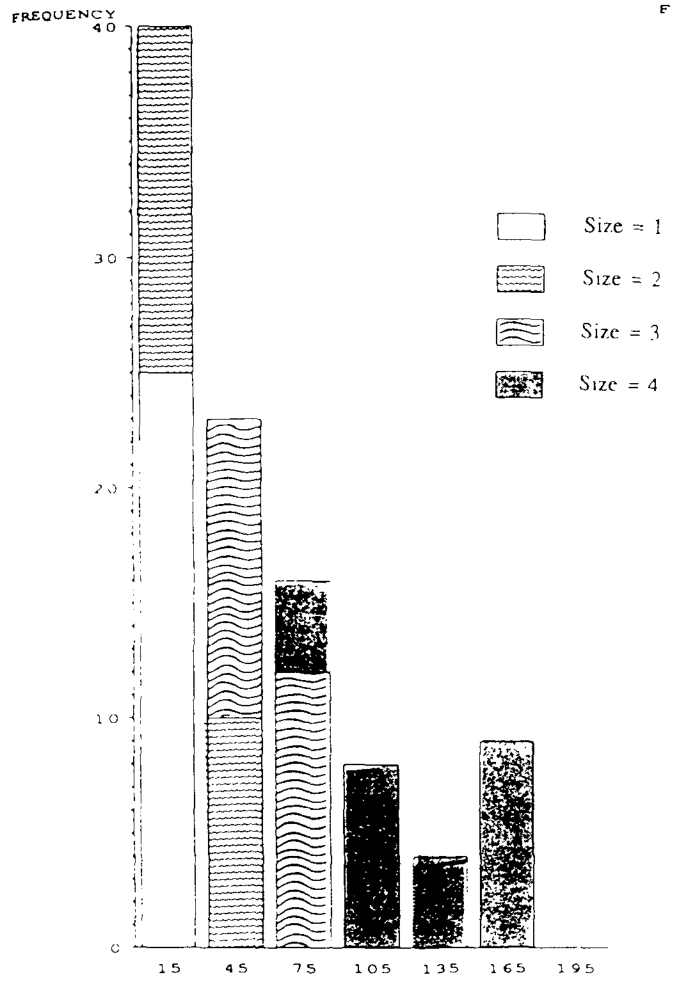

Figure 7.3. Historgam for the vanable y.

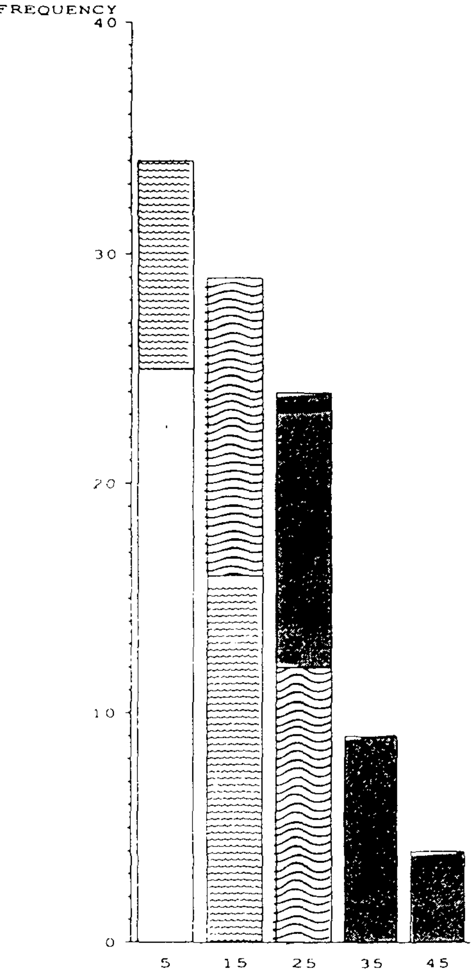

Figure 7.4. Historgam for the variable y*.

The design of the simulation experiment was as follows. For the sample sizes n=10, 20, 30, 40, 50 and 60, yielding sampling rates 10%, 20%, ..., 60%, repeated independent SYS(n;s)-samples were drawn from the population. Condition (3.1) sets the upper limit n≤62. In all 5000 independent samples were selected. For each sample, the values of s, x, y, x*, and y* were recorded for the sampled units, the latter as usual referred to by X, Y, X* and Y*. Since the simulation findings are meant to illustrate inference under total ignorance of frame order, the sampling frame was reordered by a totally random permutation before each new sample selection (cf. Remark 3.3). Statistics were computed from the sample values, and compiled as stated below.

(i) The sample sums T(X*) and T(Y*) (see (2.15)) were computed for each sample. From (7.3) and what is said in Section 2.2 it is readily checked that these sample sums relate to the HT-estimators for the population means of x, y as stated in (7.4). Note that N=100 and z=s/250.

$$\hat{\mu}(x)_{HT} = \frac{2.5}{n}\cdot T(X^*), \quad \hat{\mu}(y)_{HT} = \frac{2.5}{n}\cdot T(Y^*). \tag{7.4}$$

(ii) The true values of the variances V[T(X*)] and V[T(Y*)] were estimated by the empirical variances for the simulated sample sums T(X*) and T(Y*). The findings concerning "true" variances are reported in the bottom lines of Tables 7.1 and 7.3. The computation of the associated margins of error is described in Section 7.2.2.

(iii) For each sample the variance estimators VESS[X*;·] and VESS[Y*;·] were computed for · = a,b,c,d,e, see (3.21), (3.23), (3.24), (5.5) and (5.11). The mean values of the variance estimators over the simulated samples were derived, and these findings are reported in Tables 7.1 and 7.3. The computation of the associated margins of errors is described in Section 7.2.2.

(7.4) yields that if one multiplies a VESS[X*;·] or VESS[X*;·] by $6.25/n^2$ one obtains the corresponding estimate of the variance of the HT-estimator for the population mean.

(iv) Confidence intervals for μ(x) and μ(y) were computed by the formula

$$\hat{\mu}(x)_{HT} \pm 2\cdot\sqrt{\hat{V}[\hat{\mu}(x)_{HT};\cdot]} \tag{7.5}$$

for · = a,b,c,d,e and also for the "true" variance discussed in (i). (For notational system see the end of Section 6.1. Each confidence interval was checked with respect to coverage or not of the (true) population mean, and empirical coverage rates are reported in Tables 7.2 and 7.4.

Variance estimator	Estimated expected values with margins of error					
	n=10	n=20	n=30	n=40	n=50	n=60
VESS[X*;a]	36.4 ± 0.4	64.8 ± 0.5	85.1 ± 0.4	96.7 ± 0.4	101.0 ± 0.3	96.9 ± 0.2
VESS[X*;b]	40.4 ± 0.5	81.0 ± 0.6	121.6 ± 0.6	161.2 ± 0.6	202.0 ± 0.6	242.3 ± 0.6
VESS[X*;c]	36.4 ± 0.4	61.4 ± 0.4	77.5 ± 0.4	83.3 ± 0.4	79.9 ± 0.4	66.5 ± 0.3
VESS[X*;d]	35.5 ± 0.4	61.7 ± 0.4	78.5 ± 0.4	85.8 ± 0.4	85.2 ± 0.3	76.9 ± 0.3
VESS[X*;e]	35.6 ± 0.4	61.4 ± 0.4	77.4 ± 0.4	83.1 ± 0.4	79.4 ± 0.3	65.6 ± 0.3
V[T(X*)] estimated	35.3 ± 1.3	60.2 ± 2.3	76.8 ± 3.1	79.8 ± 3.1	81.6 ± 3.3	67.4 ± 2.8

Table 7.1: Results for the variable x*. The contents of the table are illustrated graphically in Figure 7.5.

Variance estimator in the confidence interval	Non-coverage percentages for confidence intervals for the population mean μ(x*) (see (7.5).					
	n=10	n=20	n=30	n=40	n=50	n=60
$\hat{V}[\hat{\mu}(x);a]$	7.9%	5.4%	4.2%	3.1%	3.2%	2.0%
$\hat{V}[\hat{\mu}(x);b]$	6.6%	3.3%	1.9%	0.6%	0.4%	0.0%
$\hat{V}[\hat{\mu}(x);c]$	8.0%	5.9%	5.5%	4.4%	5.5%	5.5%
$\hat{V}[\hat{\mu}(x);d]$	8.0%	5.9%	5.3%	4.2%	4.8%	3.7%
$\hat{V}[\hat{\mu}(x);e]$	8.0%	5.9%	5.5%	4.5%	5.5%	5.7%
"True" value from Table 7.1	4.7%	4.5%	4.5%	4.5%	4.6%	4.7%

Table 7.2. Non-coverage percentages for confidence intervals for μ(x). The margin of error for 5% non-coverage risk is 0.6%. The contents of the table are illustrated graphically in Figure 7.6.

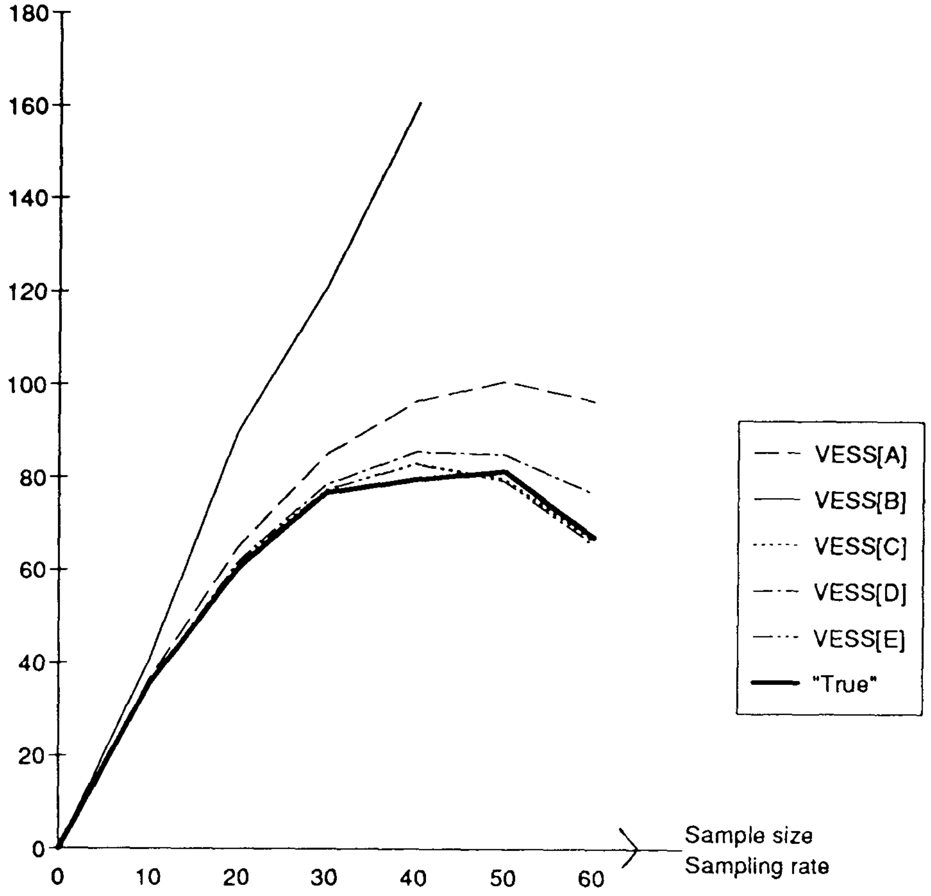

Figure 7.5. Graphical representation of the VESS values in Table 7.1. "True" denotes the estimated value of V[T(X*)]. Note that the curves for VESS[X*;c] and VESS[X*;e] hardly can be distinguished.

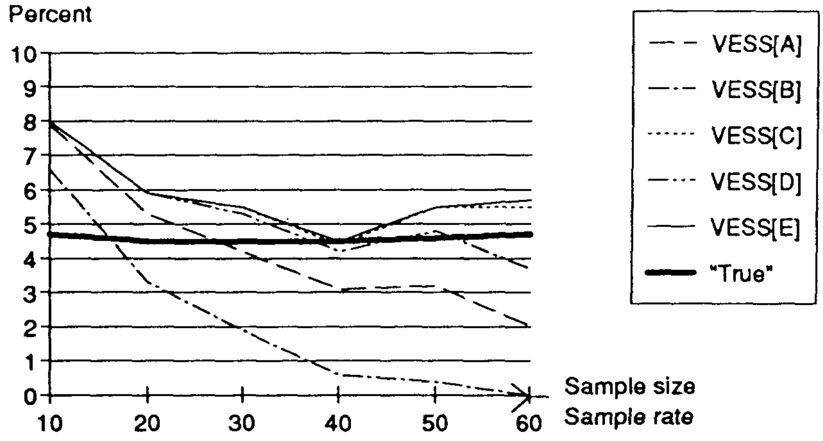

Figure 7.6. Graphical representation of the non-coverage percentages in Table 7.2. "True" corresponds to the estimated value of V[T(X*)].

Variance estimator	Estimated expected values with margins of error					
	n=10	n=20	n=30	n=40	n=50	n=60
VESS[Y*;a]	1099 ± 12	1932 ± 13	2533 ± 12	2906 ± 11	3031 ± 9	2907 ± 6
VESS[Y*;b]	1221 ± 13	2415 ± 16	3618 ± 18	4843 ± 18	6061 ± 17	7263 ± 15
VESS[Y*;c]	1079 ± 11	1851 ± 13	2353 ± 13	2583 ± 12	2521 ± 11	2170 ± 11
VESS[Y*;d]	1079 ± 11	1854 ± 13	2361 ± 12	2608 ± 10	2583 ± 8	2304 ± 7
VESS[Y*;e]	1077 ± 11	1840 ± 13	2308 ± 12	2454 ± 10	2187 ± 7	1268 ± 4
V[T(Y*)] estimated	1071 ± 42	1836 ± 75	2255 ± 94	2716 ±106	2560 ±102	1371 ± 55

Table 7.3: Results for the variable y*. The contents of the table are illustrated graphically in Figure 7.7.

Variance estimator in the confidence interval	Non–coverage percentages for confidence intervals for the population mean μ(y*) (see (7.5).					
	n=10	n=20	n=30	n=40	n=50	n=60
$\hat{V}[\hat{\mu}(y);a]$	7.9%	5.4%	4.5%	4.5%	3.5%	0.5%
$\hat{V}[\hat{\mu}(y);b]$	6.5%	3.4%	1.9%	1.0%	0.3%	0.0%
$\hat{V}[\hat{\mu}(y);c]$	8.2%	5.9%	5.4%	6.0%	5.6%	1.7%
$\hat{V}[\hat{\mu}(y);d]$	8.2%	5.8%	5.2%	5.8%	5.3%	1.2%
$\hat{V}[\hat{\mu}(y);e]$	8.2%	6.0%	5.6%	6.7%	7.1%	5.6%
"True" value from Table 7.3	4.3%	4.5%	4.4%	4.5%	4.6%	4.4%

Table 7.4. Non-coverage percentages for confidence intervals for μ(y). The margin of error for 5% non-coverage risk is 0.6%. The contents of the table are illustrated graphically in Figure 7.8.

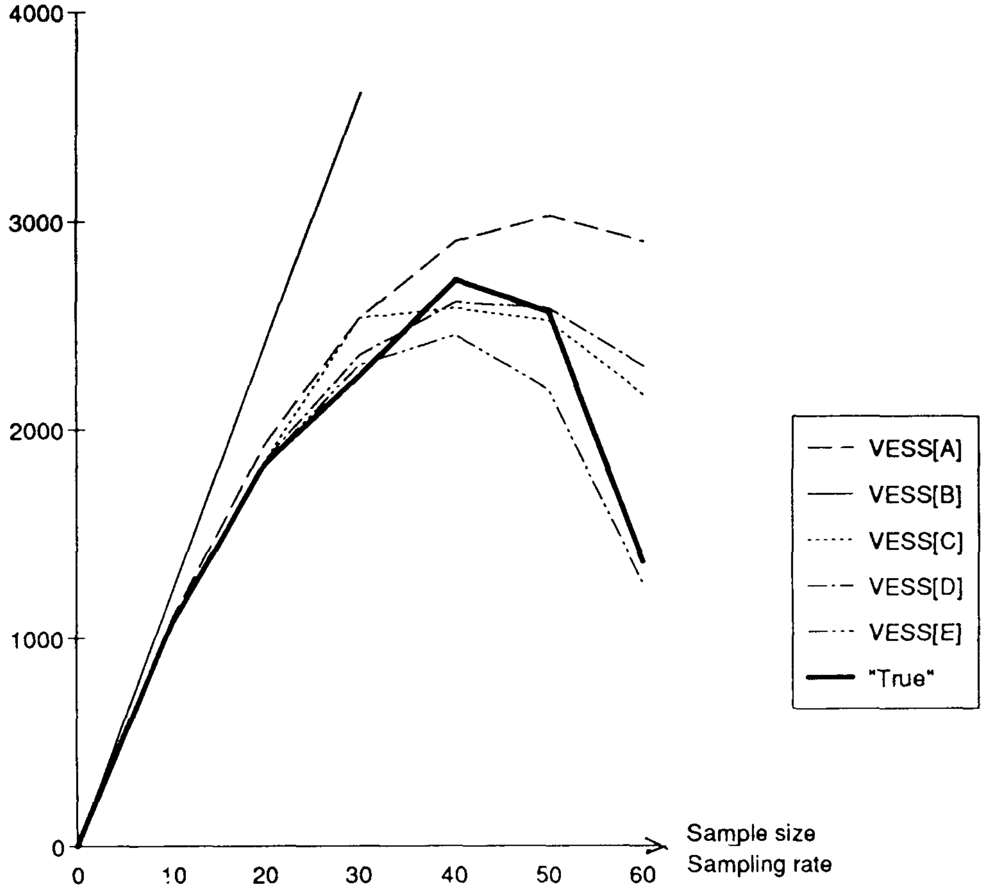

Figure 7.7. Graphical representation of the VESS values in Table 7.2. "True" denotes the estimated value of V[T(Y*)]. Note that the curves for VESS[Y*;c] and VESS[Y*;e] hardly can be distinguished.

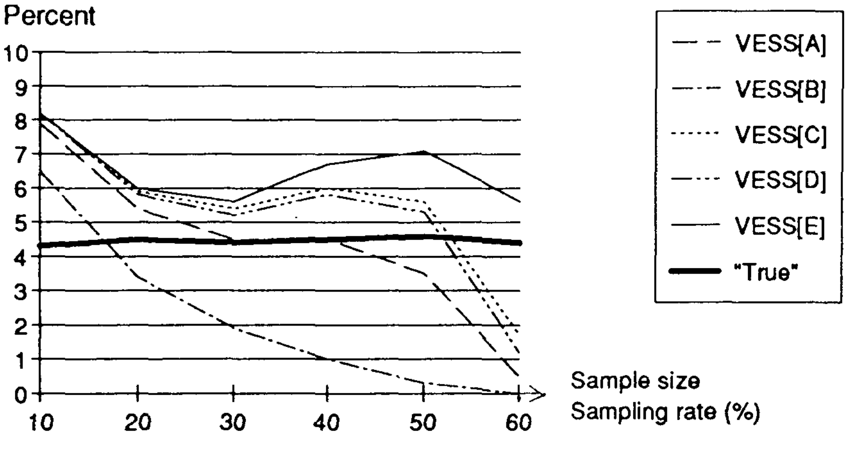

Figure 7.8. Graphical representation of the non-coverage percentages in Table 7.2. "True" corresponds to the estimated value of V[T(Y*)].

7.2.2 On the computation of margins of error

The term **margin of error** is used for twice the standard deviation of an estimated quantity. The margins of error for the estimated expected values of the VESS's were computed in the "straightforward" way; as twice the empirical standard deviations for the simulated VESS values divided by the square root of the number of iterations. Similarly, margins of error for the non-coverage percentages were computed by the usual "pq-formula".

The margin of error for estimates of "true" variances, i.e. the margins of error in the bottom lines of Tables 7.1 and 7.2, were derived by applying well-known large sample theory (see e.g. Chapter 27 in Cramér (1946)) which leads to the following (large sample) estimator for the variance of an empirical variance S^2 (see (2.21)) based on independent, equally distributed random variables $X_1,X_2,...,X_k$;

$$\hat{V}(S^2) = [\frac{1}{k}\cdot\sum_{v=1}^{k}(X_v-\bar{X})^4 - S^4]/k. \tag{7.6}$$

7.2.3 Comments on situations with trend in the study variable

So far we have almost exclusively considered the variance estimation problem under the premise of total ignorance of frame order, but in Section 3.4.2 we discussed briefly variance estimation in cases with frame trend in the study variable. However, we shall not consider the "frame trend aspect" in any depth in this paper. The aim here is just to illustrate that trend in the study variable affects matters considerably.

To that end we have made some simulations which are closely related to those in Section 7.2.1. A population with size variable **s** and study variables **x** and **y** was constructed as follows. The population in 7.2.1 was "duplicated" into two strata, which were concatenated as regards frame labelling, each stratum containing 100 units. The size values **s** were left intact in the duplication process. In one stratum also the values of **x** and **y** were left intact, while in the other stratum 10 was added to each "old" value.

The following inference model was considered. The surveyor knows that the mean levels differ between the two strata, but has total ignorance of frame order within strata. (More technically formulated; Before a new sample was drawn, the frame was randomly re-ordered within each stratum.) 500 independent samples of sizes 20, 40, 60, 80, 100 and 120 were selected, i.e. the sampling fractions were as before 10%, 20%,...,60%. According to the discussion in Section 3.4.2, segmented variance estimatimation is appropriate, with the two strata as segments. The variance estimators VESS[$\cdot$;a] to VESS[$\cdot$;e] were also computed for the entire samples. Results on the means of the latter are shown in Figure 7.9 for **x** and in Figure 7.10 for **y**. As can be seen in these figures, frame trend in the study variable can lead to over-estimation as well as to under-estimation of the true variance.

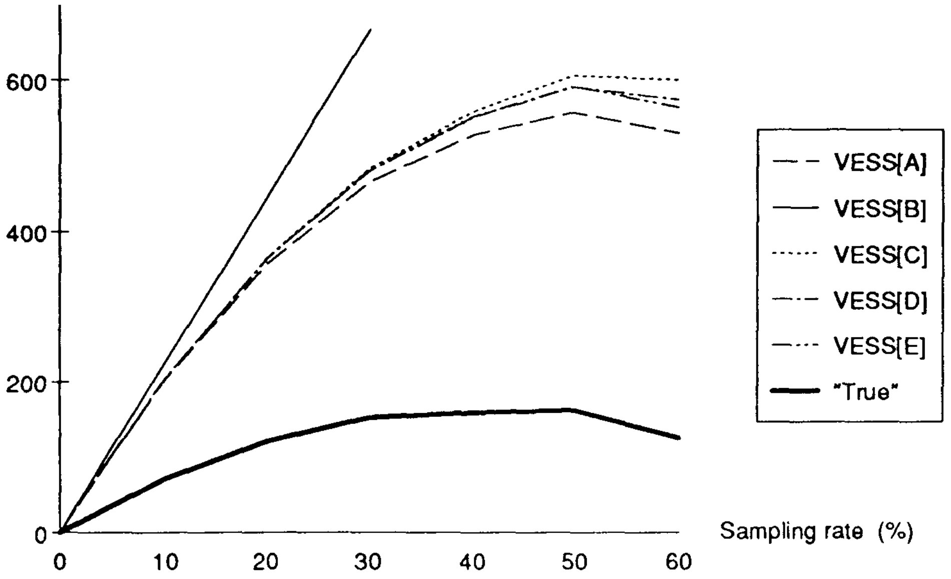

Figure 7.9. Graphical representation of the means of the VESS estimators applied to the entire x-samples. Means for the corresponding segmented variance estimators can be "copied" from Figure 7.5.

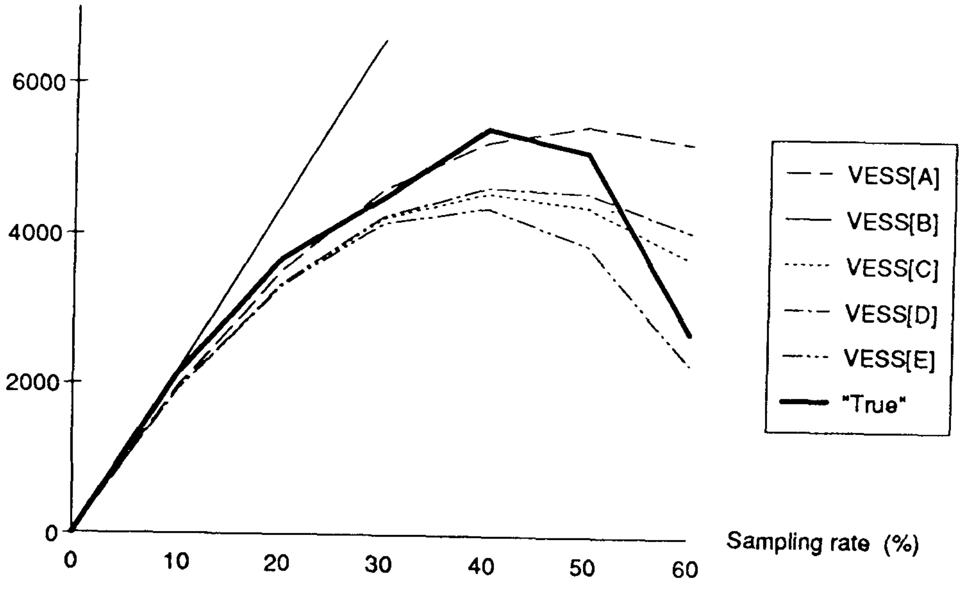

Figure 7.10. Graphical representation of the means of the VESS estimators applied to the entire y-samples. Means for the corresponding segmented variance estimators can be "copied" from Figure 7.7.

7.3 Comments on asymptotic normality

As was discussed in Section 2.2, a variance estimator Q (relative the point estimator K) is really interesting only if it can be used as a pivotal quantity in the sense that $K \pm 2\cdot\sqrt{Q}$ yields a confidence interval with approximately 95% confidence level. One requirement for this is that the variance estimator in fact yields consistent estimation of the true estimator variance. Another requirement is that the distribution of the point estimator K in fact is well approximated by a normal distribution. In our situation, this leads to the question if sample sums selected by SYS(n;s) are approximately normally distributed or not. In Section 2.2 we alluded to the "meta-theorem" which says that every "reasonable" statistic, as e.g. a sample sum, is approximately normally distributed provided the sample size is fairly large and other circumstances are not too "wild". If one thinks this meta-theorem is justification enough, everything is of course ok. However, below we give some supplementary arguments of, hopefully, "harder" nature.

(i) In Rosén (1972) approximate/asymptotic normality of the sample sum at SUC sampling is proved. This does of course not say anything directly about the behaviour of sample sums at SYS(n;s) sampling. However, provided that one believes that the approximation argument (1.2) in fact is valid with good approximation, and this is the basis for the present variance estimation method, then a consequence is also that the sample sum under total ignorance of frame order is approximately normally distributed.

(ii) Here we adhere to what is said in Section 7.2.2 on estimation of $V(S^2)$. If the X-variables are normally distributed, thereby having 4:th central moment equal to $3\cdot\sigma^4$, the following simpler (than (7.6)) estimator also applies;

$$\hat{V}(S^2) = 2\cdot S^4/k. \tag{7.7}$$

The reported margins of errors in Tables 7.1 and 7.3 are based on (7.6). However, we also computed the variance estimators in (7.7) for all the different sample sizes, and found that they gave very similar results. This provides some justification for the fact that SYS(n;s) sample sums in fact have approximately normal distributions under general conditions, at least under random frame order.

(iii) Another (partial) check of the approximate normality is obtained by the actual confidence level of the confidence interval $K \pm 2\cdot\sqrt{V[K]}$. Under perfect normality this confidence level i 4.5%. As can be seen from the bottom lines in the Tables 7.2 and 7.4, this fits well with the simulation findings.

7.4 Conclusions

One should of course be careful, not to draw too far-reaching conclusions from the limited numerical findings which are available yet. However, we mean that they indicate the following.

- All the estimators VESS[· ;a], VESS[· ;b], VESS[· ;c], VESS[· ;d] and VESS[· ;e] work satisfactorily as long as sampling rates are "negligible". VESS[· ;b] is the first to run into trouble as the sampling fraction increases, and already for "moderate" sampling fractions it becomes unreliable.

- As long as the sampling rate is "moderate" VESS[· ;a], VESS[· ;c], VESS[· ;d] and VESS[· ;e] all work fairly satisfactorily, but VESS[· ;a] is no longer reliable when the sampling rate becomes "large".
- For "large" sampling fractions VESS[· ;c], VESS[· ;d] and VESS[· ;e] all work fairly satisfactorily, but when the sampling fraction becomes "very large" VESS[· ;a] seems to be superior to the others.

One gets the impression that the two best candidates are VESS[· ;c], i.e. the Hartley-Rao estimator and the new estimator VESS[· ;e], but it would be hasty to make any firm claim about their "ranking". We want to add the following circumstances in favour of VESS[· ;e].

(i) VESS[· ;e] has the advantage that it requires knowledge of inclusion probabilities only for sampled units, while VESS[· ;c] requires knowledge of the inclusion probabilities for all units in the population (cf. (3.24)).

(ii) VESS[· ;e] has a simple rationale, namely the approximation argument (1.2), which makes it possible to judge in a particular situation whether it can be regarded as "reasonably accurate" or not.

In conclusion we want to recommend VESS[· ;e] for practical use, notably in situations with large sampling fractions. It can be employed "straightforwardly" in situations with total ignorance of frame order, as well as in segmented variance estimation procedures.

Acknowledgements. The author wants to thank Lennart Nordberg for constructive criticism on earlier versions of the paper, and Dan Hedlin for patient assistance and teaching on the SAS programme package, which was used to carry out the simulations.

References

Bellhouse, D. R. (1988). *Systematic Sampling with Illustrative Examples.* In *Handbook of Statistics, Vol 6* (eds. P. R. Krishnaiah & C. R. Rao), 125-145. North-Holland, Amsterdam.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

Hájek, J. (1960). Limiting Distributions in Simple Random Sampling from a Finite Population. *Publ Math Inst Hungar Acad Sci, Ser B* **5**, 361-374.

Hájek, J. (1981). *Sampling from a Finite Population.* Marcel Decker, New York.

Hartley, H. O. & Rao, J. N. K. (1962). Sampling with Unequal Probabilities and Without Replacement. *Ann Math statist* **33**, 350-374.

Murthy, M. N. & Rao, T. J. (1988). *Systematic Sampling with Illustrative Examples.* In *Handbook of Statistics, Vol 6* (eds. P. R. Krishnaiah & C. R. Rao), 147-185. North-Holland, Amsterdam.

Rao, J. N. K. (1985). Conditional Inference in Survey Sampling. *Survey Methodology* **11**, 15-31.

Rosén, B. (1972). Asymptotic Theory for Successive Sampling with Varying Probabilities Without Replacement, I and II. *Ann Math Statist* **43**, 373-397 and 748-776.

Savage, L. J. (1954). *The Foundations of Statistics.* Wiley, New York.

Wolter, K. M. (1985). *Introduction to Variance Estimation.* Springer-Verlag, New York.

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 during 1991:

1991:1 (grön) Computing elementary aggregates in the Swedish consumer price index **(Jörgen Dalén)**

1991:2 (grön) Översikt av estimatorer för stora och små redovisningsgrupper **(Sixten Lundström)**

1991:3 (grön) Interacting nonresponse and response errors **(Håkan L Lindström)**

1991:4 (grön) Effects of nonresponse on survey estimates in the analysis of competing exponential risks **(Ingrid Lyberg)**

1991:5 (grön) Study of the estimation precision in the Chinese MIS surveys **(Bengt Rosén)**

1991:6 (beige) Abstracts I - Sammanfattning av metodrapporter från SCB

1991:7 (grön) Kvalitetsfonden 1990: Projekt - handläggning - utvärdering **(Roland Friberg och Håkan L Lindström)**

1991:8 (grön) Tre återintervjustudier: Inkomstfördelningsundersökningen **(Jan Eriksson)**, Högskoleenkätuppföljningen **(Anders Karlsson)**, Årsarbetskraften **(Majvor Karlsson och Solveig Thudin)**

1991:9 (gul) What metainformation should accompany statistical macrodata **(Bo Sundgren)**

1991:10 (grön) The Family Expenditure Survey: An Experiment with Incentives **(Håkan L. Lindström)**, Seasonal Variation and Response Behaviour in Swedish Households Expenditure **(Peter Lundquist)**, Reducing Nonresponse Rates in Family Expenditure Surveys by Forming Ad Hoc Task Forces **(Lars Lyberg)**, A Study of Errors in Swedish Consumption Data **(Martin G. Ribe)**

1991:11 (gul) Statistical Metainformation and Metainformation Systems **(Bo Sundgren)**

1991:12 (beige)	Abstracts II - Sammanfattning av metodrapporter från SCB
1991:13 (grön)	Bortfallsbarometern nr 6 **(Mats Bergdahl, Sonia Ekman, Anders Lindberg, Peter Lundquist, Monica Rennermalm)**
1991:14 (grön)	Kvalitetsrapport 1991 - Utveckling av kvaliteten för SCBs statistikproduktion **(Jan Eklöf, Per Nilsson)**

Kvarvarande **beige** och **gröna** exemplar av ovanstående promemorior kan rekvireras från Inga-Lill Pettersson, U/LEDN, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 49 56.

Kvarvarande **gula** exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 47.