R&D Report 1998:6
Research – Methods – Development

Statistiska centralbyrån Statistics Sweden

On variance estimation for measures of change when samples are coordinated by a permanent random numbers technique

by

Lennart Nordberg

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 1998:6. On variance estimation for measures of change when samples are coordinated by a permanent random numbers technique / Lennart Nordberg.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1998-X101OP9806

R&D Report **1998:6**
Research – Methods – Development

On variance estimation for measures of change when samples are coordinated by a permanent random numbers technique

by

Lennart Nordberg

R&D Report 1998:6

Research - Methods - Development

On variance estimation for measures of change when samples are coordinated by a permanent random numbers technique

---

Från trycket	Januari 1999
Producent	Statistiska centralbyrån, *Statistics Sweden,* metodenheten Box 24300, *SE*-104 51 STOCKHOLM
Utgivare	Lars Lyberg

Förfrågningar	Lennart Nordberg lennart.nordberg@scb.se telefon 019-17 60 12

ISSN 0283-8680

Printed in Sweden
SCB-Tryck, Örebro 1999

Abstract

A common objective in business surveys is to compare two estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ of the same characteristic taken on two occasions, e.g. the level of production the same month in two subsequent years, and to judge whether the observed change is statistically significant or merely subject to random variation.

Business surveys often use samples at subsequent occasions that are positively coordinated, i.e. overlapping, in order to increase precision in estimates of change over time. Such sample coordination will make $\hat{\theta}_0$ and $\hat{\theta}_1$ become correlated.

Under common rotating panel designs this correlation can often be estimated in a straightforward way. However, some systems used for sample coordination are designed not only to generate samples that are positively coordinated between subsequent occasions but also to obtain negative coordination between different surveys in order to spread the response burden. One way of simultanously creating positive and negative coordination is to use permanent random number techniques. In some of these systems the rotation pattern becomes random which makes the estimation of the correlation more difficult than under common rotating panels.

The so called SAMU system for sample coordination of business surveys at Statistics Sweden is such a system. The purpose of the present paper is to show how to estimate the variance for measures of change such as $\hat{\psi} = \hat{\theta}_1 - \hat{\theta}_0$ or $\hat{\psi} = \hat{\theta}_1 / \hat{\theta}_0$ when $\hat{\theta}_0$ and $\hat{\theta}_1$ are estimated from two separate SAMU samples.

Note: This manuscript is an extended version, translated into English, of Nordberg (1994), see reference list.

Acknowledgements: I am indepted to Dr. Tiina Orusild for letting me use a result from her forthcoming paper Orusild (1999), see reference list. I also want to thank Dr. Sixten Lundström for helpful comments on an earlier version of this manuscript.

Key words: Survey Sampling, Variance Estimation, Estimates of Change, Panel Designs, Permanent Random Numbers.

Contents

1. Introduction

Often in business surveys one wants to compare two estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ of the same characteristic taken on two occasions 0 and 1, e.g. the level of production the same month in two subsequent years, and to judge whether the observed change is statistically significant or merely subject to random variation.

Business surveys often use samples at subsequent occasions that are positively coordinated, i.e. overlapping, in order to increase precision in estimates of change over time. Such sample coordination will make $\hat{\theta}_0$ and $\hat{\theta}_1$ become correlated.

Under common rotating panel designs this correlation can often be estimated in a straightforward way. However, some systems used for sample coordination are designed not only to generate samples that are positively coordinated between subsequent occasions but also to obtain negative coordination between different surveys in order to spread the response burden. One way of simultanously creating positive and negative coordination is to use permanent random number techniques. In some of these systems the rotation pattern becomes random which makes the estimation of the correlation more difficult than under common rotating panels.

The so called SAMU system for sample coordination of business surveys at Statistics Sweden is such a system. The word SAMU ('SAMordnade Urval') is an abbreviation in Swedish for co-ordinated samples.

The purpose of the present paper is to show how to estimate the variances for measures of change such as $\hat{\psi} = \hat{\theta}_1 - \hat{\theta}_0$ or $\hat{\psi} = \hat{\theta}_1 / \hat{\theta}_0$ when θ_0 and θ_1 are estimated from two separate SAMU samples. Although SAMU applies also to other types of designs we will confine the discussion here to the simple and common case of stratified sampling of elements (businesses) with simple random sampling without replacement within strata. Next we give a brief presentation of SAMU. For more complete descriptions, see Ohlsson (1992), (1995).

Every sample in the SAMU system is drawn from an up-to-date version of the Business Register. The co-ordination of samples is obtained by the so called JALES technique, the basic idea being to associate a random number to every element (enterprise or local unit) as soon as it enters the Business Register and to keep it as long as the element remains in the register. All generated random numbers are to be independent. Suppose that we want to sample 10 elements in a particular stratum. All the frame elements in the current stratum are ordered by the size of their random numbers. An arbitrary starting point is chosen and the first 10 elements 'to the right' (say) of this starting point are included in the sample. It can be shown, see Ohlsson (1992), that this sampling mechanism is equivalent to simple random sampling.

We will in the following consider sampling in the SAMU system on two occasions, time 0 and time 1, and hence apply the JALES technique to two different versions of the Business Register. The JALES technique will then obviously introduce some additional randomness compared to the case of common rotating panels. Whether a certain element (business) which was included in the sample at time 0 will remain in the sample at the next sampling occasion at time 1 depends not only on the element itself but also on the behaviour of other elements, notably the random numbers associated with the births and deaths in the frame.

If $\hat{\psi} = \hat{\theta}_1 - \hat{\theta}_0$ we can write the variance of $\hat{\psi}$ as follows.

$$V(\hat{\psi}) = V(\hat{\theta}_0) + V(\hat{\theta}_1) - 2 \cdot C(\hat{\theta}_0, \hat{\theta}_1) . \tag{1.1}$$

If $\hat{\psi} = \hat{\theta}_1 / \hat{\theta}_0$ we have by Taylor linearisation,

$$\frac{V(\hat{\psi})}{\psi^2} \approx \frac{V(\hat{\theta}_0)}{\theta_0^{\,2}} + \frac{V(\hat{\theta}_1)}{\theta_1^{\,2}} - 2 \cdot \frac{C(\hat{\theta}_0, \hat{\theta}_1)}{\theta_0 \theta_1} \tag{1.2}$$

Although θ_0 and θ_1 may be more complex parameters than population totals (see relation (1.5) ahead) the main problem concerns the covariance term, not the variance components $V(\hat{\theta}_0)$ and $V(\hat{\theta}_1)$.

Estimation at time 0: Consider a set of variables $y_1, \ldots, y_j, \ldots, y_J$ and let y_{jk} be the value of the variable y_j for element k in the finite population U at time 0. We associate a population total $t_j = \sum_{k \in U} y_{jk}$ with every variable y_j. The population U is stratified into H strata, $U_1, \ldots, U_h, \ldots, U_H$, and a simple random sample is drawn from each stratum. Let s denote the chosen sample and let N_h and n_h be the number of population- and sample elements respectively in stratum h, $h = 1, 2, \ldots, H$.

As estimator for the total t_j we consider the Horvitz-Thompson (H-T) or the Generalised Regression (GREG) estimator. The GREG estimator is treated in Section 4 ahead so until then we will assume that the H-T estimator is used.

Hence

$$\hat{t}_j = \sum_h \frac{N_h}{n_h} \sum_{k \in U_h} y_{jk} \delta_k \ , \tag{1.3}$$

where

$$\delta_k = \begin{cases} 1 & \text{if element } k \in s \\ 0 & \text{otherwise} \end{cases} \qquad (1.4)$$

We will assume that the estimator $\hat{\theta}_0$ can be expressed on the following form:

$$\hat{\theta}_0 = f(\hat{t}_1, \hat{t}_2, .., \hat{t}_J), \qquad (1.5)$$

where f is an arbitrary rational function. A common estimator for a population mean or a difference between two domain totals, a ratio estimator or a poststratified ratio estimator for a population total are examples of such functions.

Estimation at time 1: The population U' at time 1 consisting of N' elements is stratified into L strata, $U'_1,...,U'_l,...,U'_L$. The stratification at time 1 does not have to be the same as the one at time 0. Let s' be the sample and let N'_l and n'_l be the population- and sample size in stratum l, $l = 1, 2,...,L$.

The population totals at time 1 are estimated in analogy with (1.3).

$$\hat{t}'_j = \sum_l \frac{N'_l}{n'_l} \sum_{r \in U'_l} y'_{jr} \delta'_r, \qquad (1.6)$$

where

$$\delta'_r = \begin{cases} 1 & \text{if element } r \in s' \\ 0 & \text{otherwise} \end{cases} \qquad (1.7)$$

We assume that the estimator $\hat{\theta}_1$ can be written on the form

$$\hat{\theta}_1 = f(\hat{t}'_1, \hat{t}'_2, .., \hat{t}'_J). \qquad (1.8)$$

Notice that the same function f is assumed in (1.5) and (1.8). This should be the most common case in practice. Generalisation to the case of different functions f_0 and f_1 is straight forward.

The covariance: By standard Taylor linearisation we can now write the covariance in (1.1) and (1.2) as follows.

$$C(\hat{\theta}_0,\hat{\theta}_1) \approx \sum_i \sum_j f_i'(t_1,t_2,..,t_J) f_j'(t_1',t_2',..,t_J') C(\hat{t}_i,\hat{t}_j'), \qquad (1.9)$$

f_i' being the partial derivative $\frac{\partial f}{\partial t_i}$ and $C(\hat{t}_i,\hat{t}_j')$ the covariance for the pair $(\hat{t}_i,\hat{t}_j')$.

Next we will study the term $C(\hat{t}_i,\hat{t}_j')$ with its estimator $\hat{C}(\hat{t}_i,\hat{t}_j')$ and $C(\hat{\theta}_0,\hat{\theta}_1)$ with its estimator $\hat{C}(\hat{\theta}_0,\hat{\theta}_1)$ in detail.

2. Estimator for the covariance

By combination of (1.3) and (1.6) we can write the covariances as follows.

$$C(\hat{t}_i,\hat{t}'_j)=\sum_h\sum_l\sum_{k\in U_h}\sum_{r\in U'_l}\frac{N_h\cdot N'_l}{n_h\cdot n'_l}\cdot y_{ik}\cdot y'_{jr}\cdot C(\delta_k,\delta'_r)\ . \qquad (2.1)$$

The expression (2.1) can be simplified in the following way. The union of the sampling frames for time 0 and 1 respectively can be divided into three non-overlapping parts. The first part consists of the elements that were included in the frame at time 0 but not at time 1, i.e. those elements that have disappeared between time 0 and 1. We call this group D (D for 'deaths').

The second part consists of the elements that were included in both frames (time 0 and 1). We call this group P (P for 'persistors').

The third part consists of the elements that are included in the frame at time 1 but not at time 0. We call this group B (B for 'births').

The division into the three groups D, P and B is illustrated by the following figure.

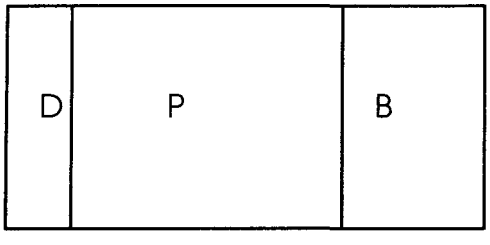

Figure 2.1

The set D can be split into the non-overlapping subsets $\{D_h, h=1,\ 2,..H\}$ where D_h is the set of frame elements that belonged to stratum h at time 0 and had left the frame before time 1.

Correspondingly, the set B can be split into the non-overlapping subsets $\{B_l, l=1,\ 2,..,L\}$ where B_l is the set of frame elements in stratum l , time 1, l =1, 2,..,L which were not found anywhere in the frame at time 0.

The set P can be further divided into the non-overlapping sets $\{P_{hl}, h=1\ ,2,..,H,\ l=1\ ,2,..,L\}$ where P_{hl} is the group of frame elements that belonged to stratum h at time 0 and stratum l at time 1.

Among the n_h elements sampled from stratum h at time 0 we assume that d_h belong to D_h, $h=1, 2,...,H$ and that a_{hl} belong to the stratum combination P_{hl}. Hence $n_h = d_h + \sum_l a_{hl}$.

Among the n_l' elements sampled from stratum l at time 1 we assume that b_l' belong to B_l, $l=1, 2,...,L$ and that a_{hl}' belong to the stratum combination P_{hl}. Hence $n_l' = b_l' + \sum_h a_{hl}'$.

Let G_{hl} be the number of frame elements in P_{hl}. Furthermore, let g_{hl} be the number of elements that belong to P_{hl} ***and*** are included in ***both*** samples (time 0 and 1). We illustrate this by the following figure.

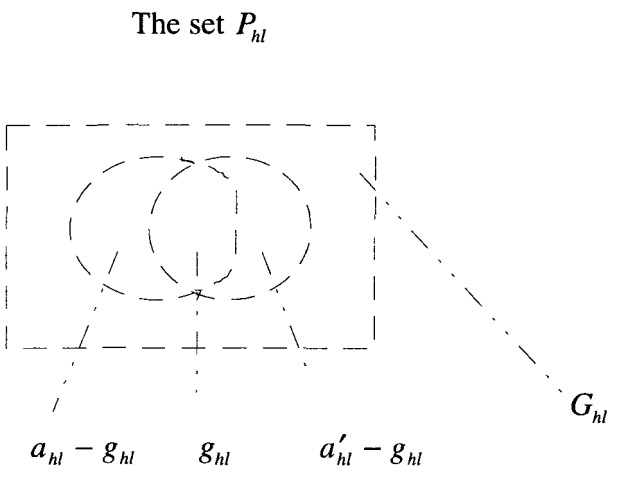

Figure 2.2 (Number of elements)

The quantity $\Omega = \{a_{hl}, a_{hl}', g_{hl}, d_h, b_l', h = 1, 2,..H, l = 1, 2,.., L\}$ could be considered as random (Sjölinder (1971), Bäcklund (1972) and Garås (1989). However, we will in the following make the analysis conditional on Ω since Ω can be considered as an ancillary quantity for the present analysis.

Hence in the general expression

$$V(\hat{\psi}) = E_{\Omega}(V(\hat{\psi}|\Omega)) + V_{\Omega}(E(\hat{\psi}|\Omega)). \tag{2.2}$$

we will estimate the *conditional* variance $V(\hat{\psi}|\Omega)$, and ignore the second component of (2.2). We can then rewrite (2.1) as a conditional covariance:

$$C\left(\hat{t}_i, \hat{t}'_j|\Omega\right) = \sum_h \sum_l \sum_{k \in U_h} \sum_{r \in U'_l} \frac{N_h \cdot N'_l}{n_h \cdot n'_l} \cdot y_{ik} \cdot y'_{jr} \cdot C(\delta_k, \delta'_r|\Omega) \ . \tag{2.3}$$

It follows from the use of independent random numbers under the JALES - technique that we only have to consider the covariance of (2.3) for those pairs of frame elements k and r that both come from the same stratum (h) at time 0 and the same stratum (l) at time 1, i.e only those pairs of elements k and r where both elements belong to P_{hl}. (Sjölinder (1971) and Bäcklund (1972).

Hence the relation (2.3) can be written on the following form.

$$C(\hat{t}_i, \hat{t}'_j|\Omega) = \sum_h \sum_l \sum_{k \in P_{hl}} \sum_{r \in P_{hl}} \frac{N_h \cdot N'_l}{n_h \cdot n'_l} \cdot y_{ik} \cdot y'_{jr} \cdot C(\delta_k, \delta'_r|\Omega), \tag{2.4}$$

or, equivalently

$$C(\hat{t}_i, \hat{t}'_j|\Omega) =$$
$$\sum_h \sum_l \sum_{k \in P_{hl}} \sum_{r \in P_{hl}} \frac{N_h \cdot N'_l}{n_h \cdot n'_l} \cdot y_{ik} \cdot y'_{jr} \cdot (E(\delta_k \cdot \delta'_r|\Omega) - E(\delta_k|\Omega) \cdot E(\delta'_r|\Omega)). \tag{2.5}$$

The following quantity is unbiased for $C(\hat{t}_i, \hat{t}'_j|\Omega)$:

$$\hat{C}(\hat{t}_i, \hat{t}'_j|\Omega) =$$
$$\sum_h \sum_l \sum_{k \in P_{hl}} \sum_{r \in P_{hl}} \frac{N_h \cdot N'_l}{n_h \cdot n'_l} \cdot y_{ik} \cdot y'_{jr} \cdot \left(1 - \frac{E(\delta_k|\Omega) \cdot E(\delta'_r|\Omega)}{E(\delta_k \cdot \delta'_r|\Omega)}\right) \cdot \delta_k \cdot \delta'_r. \tag{2.6}$$

Orusild (1999) computes the three expectations included in (2.6) and shows that the estimator (2.6) can be put on the following form (see Appendix A for an alternative proof).

$$\hat{C}(\hat{t}_i,\hat{t}_j'|\Omega)=\sum_h\sum_l A_{hl}\cdot\left\{\sum_{k\in P_{hl}} y_{ik}\cdot y_{jk}'\delta_k\cdot\delta_k'-\frac{1}{\tilde{a}_{hl}}(\sum_{k\in P_{hl}} y_{ik}\delta_k)(\sum_{r\in P_{hl}} y_{jr}'\delta_r')\right\},\quad 2.7)$$

where $\tilde{a}_{hl}=\dfrac{a_{hl}\cdot a_{hl}'}{g_{hl}}$

and $A_{hl}=\dfrac{N_h N_l'\tilde{a}_{hl}}{G_{hl}^{\,2} n_h n_l'}\left(\dfrac{G_{hl}(G_{hl}-\tilde{a}_{hl})}{(\tilde{a}_{hl}-1)}\right).$

Comment: The quantities a_{hl}, a_{hl}', g_{hl} och G_{hl} were defined in connection with figure 2.2 above.

The complete covariance, conditional on Ω,

$$\hat{C}(\hat{\theta}_0,\hat{\theta}_1|\Omega)=\sum_i\sum_j f_i'(\hat{t}_1,\hat{t}_2,..,\hat{t}_J)f_j'(\hat{t}_1',\hat{t}_2',..,\hat{t}_J')\hat{C}(\hat{t}_i,\hat{t}_j'|\Omega),\qquad(2.8)$$

is obtained by inserting the expressions (1.3), (1.6) and (2.7) into (1.9).

3. A procedure for estimation of the covariance and the measures of change

We will first show how the covariance formula (2.7) can be transformed so that it can be written on the 'usual' form under stratified SRS.

Set $z_{jk} = y_{jk} \frac{N_h \tilde{a}_{hl}}{n_h G_{hl}}$ and $z'_{jk} = y'_{jk} \frac{N'_l \tilde{a}_{hl}}{n'_l G_{hl}}$, $j = 1,2,..,J$, for $k \in P_{hl}$, $h = 1,\ 2,..,H,\ l = 1,\ 2,..,L$.

Then (2.7) can be written on the following form, i.e. the 'usual' formula for the estimator of the covariance between two π-weighted totals under stratified SRS where strata comprise every combination of (h,l) in P_{hl}, the population size "Capital N" equals G_{hl} and the sample size "small n" equals $\tilde{a}_{hl}$.

$$\hat{C}(\hat{t}_i, \hat{t}'_j | \Omega) = \sum_h \sum_l \frac{G_{hl}(G_{hl} - \tilde{a}_{hl})}{\tilde{a}_{hl}(\tilde{a}_{hl} - 1)} * \left\{ \sum_{k \in P_{hl}} z_{ik} \cdot z'_{jk} \delta_k \cdot \delta'_k - \frac{1}{\tilde{a}_{hl}} \left(\sum_{k \in P_{hl}} z_{ik} \delta_k \right) \left(\sum_{k \in P_{hl}} z'_{jk} \delta'_k \right) \right\} \quad (3.1)$$

In order to get the whole covariance term (2.8) right we must also compute the point estimates $\hat{t}_i$ and $\hat{t}'_j$ according to (1.3) and (1.6) respectively. This means that also the observations that are not included in P_{hl} must be accounted for. These observations contribute to the partial derivatives in (2.8) but not to $\hat{C}(\hat{t}_i, \hat{t}'_j | \Omega)$. We will now introduce a data transformation which will enable us to write (2.8) on a "standard form" which can be handled by a 'normal' variance algorithm.

Procedure 3.1

1) Distribute every sample element (which appeared in ***either*** *one of the time 0 and time 1 samples) to the following groups :*

$\boldsymbol{q} = (q_1, q_{2,..,} q_Q) = (D_h, P_{hl}, B_l, h = 1,\ 2,..,H,\ l = 1,\ 2,..,L)$

where

- D_h *includes the elements in stratum h at time 0 which disappeared from the frame before time 1, i.e. the deaths in stratum h.*

- P_{hl} *includes the elements which belonged to stratum h at time 0 and stratum l at time 1. Notice that elements that are included in just one of the two samples must be included, not only the ones included in both samples.*

- B_l *includes the elements that are not found in the frame at time 0 but that belong to stratum l at time 1, i.e the births in stratum l .*

2) *For every group q, compute the following quantities* $\tilde{N}$ och $\tilde{n}$*:*

$$\tilde{N}_q = \begin{cases} N_h & \text{if } q = D_h,\ h = 1,2,..,H \\ G_{hl} & \text{if } q = P_{hl}, h = 1,\ 2,..,H, l = 1,\ 2,..,L. \\ N_l' & \text{if } q = B_l,\ l = 1,\ 2,..,L. \end{cases} \tag{3.2}$$

$$\tilde{n}_q = \begin{cases} n_h & \text{if } q = D_h,\ h = 1,\ 2,..,H \\ \tilde{a}_{hl} & \text{if } q = P_{hl}, h = 1,2,..,H,\ l = 1,\ 2,..,L. \\ n_l' & \text{if } q = B_l,\ l = 1,2,..,L. \end{cases} \tag{3.3}$$

3) *Transform* y_{jk} and y'_{jk} *to* z_{jk} and z'_{jk} *for* $j = 1,\ 2,..,J$ *as follows.*

a) if $k \in D_h,\ h = 1,\ 2,..,H.$ *:*

$$\begin{aligned} z_{jk} &= y_{jk} \\ z'_{jk} &= 0 \end{aligned} \tag{3.4}$$

b) if $k \in B_l,\ l = 1,\ 2,..,L.$*:*

$$\begin{aligned} z_{jk} &= 0 \\ z'_{jk} &= y'_{jk} \end{aligned} \tag{3.5}$$

c) if $k \in P_{hl},\ h = 1,\ 2,..,H,\ l = 1,\ 2,..,L.$*:*

$$z_{jk} = \begin{cases} y_{jk} \cdot \dfrac{N_h \cdot \tilde{a}_{hl}}{n_h \cdot G_{hl}} & \text{if } k \in s \\ 0 & \text{if } k \notin s \end{cases} \tag{3.6}$$

$$z'_{jk} = \begin{cases} y'_{jk} \cdot \dfrac{N_l' \cdot \tilde{a}_{hl}}{n_l' \cdot G_{hl}} & \text{if } k \in s'. \\ 0 & \text{if } k \notin s' \end{cases} \tag{3.7}$$

End of Procedure 3.1

It only takes some elementary algebra to see that (3.8) och (3.9) below are equivalent to (1.3) and (1.6) respectively. Furthermore, the expression (3.10) is equivalent to (2.7). Notice that the contributions to (3.10) from D_h and B_l are zero as they should be.

$$\hat{t}_j = \sum_q \frac{\tilde{N}_q}{\tilde{n}_q} \sum_{k \in q} z_{jk} \quad , \tag{3.8}$$

$$\hat{t}'_j = \sum_q \frac{\tilde{N}_q}{\tilde{n}_q} \sum_{k \in q} z'_{jk} \quad , \tag{3.9}$$

$$\hat{C}(\hat{t}_i, \hat{t}'_j \mid \Omega) =$$
$$= \sum_q \frac{\tilde{N}_q(\tilde{N}_q - \tilde{n}_q)}{\tilde{n}_q(\tilde{n}_q - 1)} * \left\{ \sum_{k \in q} z_{ik} \cdot z'_{jk} \cdot \delta_k \cdot \delta'_k - \frac{1}{\tilde{n}_q} \left(\sum_{k \in q} z_{ik} \delta_k \right) \left(\sum_{k \in q} z'_{jk} \delta'_k \right) \right\} \tag{3.10}$$

Hence by generating the pseudo data z and z' as in Procedure 3.1 and then applying standard formulas for point-, variance- and covariance estimators under ordinary stratified simple random sampling (STSI) with q serving as strata, $\tilde{N}_q$ serving as population size parameter ('Capital N') and $\tilde{n}_q$ as sample size parameter ('small n') we can compute the covariance estimator (2.7) by (3.10).

Furthermore, suppose that a software is available which - under this STSI design - can compute proper variance estimates for $\hat{\theta}_0$, $\hat{\theta}_1$ and $\hat{\theta}_0 + \hat{\theta}_1$ where $\hat{\theta}_0 = f(\hat{t}_1, \hat{t}_2, .., \hat{t}_J)$, $\hat{\theta}_1 = f(\hat{t}'_1, \hat{t}'_2, .., \hat{t}'_J)$ and $\hat{t}_j$ and $\hat{t}'_j$ are defined by (3.8) and (3.9). Then the covariance estimate of (2.8) can be extracted from the following general relation:

$$\hat{V}(\hat{\theta}_0 + \hat{\theta}_1) = \hat{V}(\hat{\theta}_0) + \hat{V}(\hat{\theta}_1) + 2 \cdot \hat{C}(\hat{\theta}_0, \hat{\theta}_1) \tag{3.11}$$

Notice that this software must also meet the following requirements:

- Since $\tilde{a}_{hl}$ and hence $\tilde{n}_q$ may not be an integer (see (2.7) and (3.3)) the software must be able to accept arbitrary non-negative values for the sample size parameter ('small n'). Softwares that compute n by counting elements in the input data set will not be appropriate for this task.

- Even though $\tilde{N}_q$ is normally larger than $\tilde{n}_q$ there is no absolute guarantee for this. As a consequence the software must be able to accept negative variance contributions from some (extreme) strata. Notice that the quantities

$\tilde{V}$ that appear in step c of Procedure 3.2 below are based on the pseudo data and may include negative contributions from certain strata whenever $\tilde{a}_{hl} > G_{hl}$ (see (2.7)).

The software CLAN developed at Statistics Sweden, see Andersson and Nordberg (1994) and (1998) meet the requirements mentioned above, and can readily be used to implement the following Procedure 3.2.

Procedure 3.2 *(Estimation of variances for the measures of change)*

a) Estimation for time 0: Compute $\hat{\theta}_0$ *by (1.5) and an estimate* $\hat{\mathrm{V}}(\hat{\theta}_0)$ *for* $\mathrm{V}(\hat{\theta}_0)$.

b) Estimation for time 1: Compute $\hat{\theta}_1$ *by (1.8) and an estimate* $\hat{\mathrm{V}}(\hat{\theta}_1)$ *for* $\mathrm{V}(\hat{\theta}_1)$.

c). *Estimation of the covariance:*

(i) Perform procedure 3.1.

(ii) Use the same software as in the previous steps a) and b) to compute the estimates $\tilde{V}(\tilde{\theta}_0), \tilde{V}(\tilde{\theta}_1)$ and $\tilde{\mathrm{V}}(\tilde{\theta}_0 + \tilde{\theta}_1)$ *for* $V(\tilde{\theta}_0), V(\tilde{\theta}_1)$ and $\mathrm{V}(\tilde{\theta}_0 + \tilde{\theta}_1)$ *based on the pseudo data generated in (i). The tilde sign symbolizes the fact that computations are based on the pseudo data.*

(iii) Estimate the covariance (2.8) by the following relation .

$$\tilde{C} = 0.5 \cdot (\tilde{V}(\tilde{\theta}_0 + \tilde{\theta}_1) - \tilde{V}(\tilde{\theta}_0) - \tilde{V}(\tilde{\theta}_1)), \qquad (3.12)$$

where $\tilde{V}$ *symbolizes variance estimate based on pseudo data.*

d) Estimation of the measures of change:

The estimated variances for the measures of change as in (1.1) and (1.2) can now be easily computed by inserting the estimates $\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\mathrm{V}}(\hat{\theta}_0)$, $\hat{\mathrm{V}}(\hat{\theta}_1)$ *and* $\tilde{C}$ *into (1.1) and (1.2) respectively.*

End of Procedure 3.2

4. Extension to GREG estimation

We will now show how the procedure presented above can be modified to incorporate the case when the H-T estimators of (1.3) and (1.6) respectively replaced by GREG estimators. Before proceeding we need some notation and results concerning GREG -estimation. See Särndal, Swensson and Wretman (1992) for a comprehensive presentation of the theory.

Let y be a variable of interest and let y_k be the y-value for element k in a finite population. Also available is a value $\mathbf{x}_k = (x_{1k},...,x_{mk},...,x_{Mk})^T$ of the vector $\mathbf{x}$ of length M.

The GREG-estimator can be motivated in terms of regression theory along the following lines. Suppose that the scatter of the N points $(y_k, x_{1k},...,x_{mk},...,x_{Mk}) : k = 1, ..., N$ looks as if it had been generated by a linear regression model with y as the response and the x:s as covariates. The values, $y_1,...,y_k,...,y_N$ are assumed to be realised values of independent random variables, $Y_1, ..., Y_k, ..., Y_N$. Moreover it is assumed that,

$$\begin{cases} E_\xi(Y_k) = \mathbf{x}_k'\boldsymbol{\beta} \\ V_\xi(Y_k) = \sigma_k^2 \left(= \sigma^2/\tau_k\right) \end{cases} \tag{4.1}$$

where E_ξ and V_ξ denote expected value and variance with respect to the model ξ while $\boldsymbol{\beta}$ and σ_k^2 are usually unknown model parameters and τ_k (>0) is a known scale parameter. Suppose that the values of y are only known for the elements in the sample while $\mathbf{x}$ is known for every element in the population.

The GREG-estimator of the population total t_y for y can now be expressed on the form

$$\hat{t}_{y,GREG} = \mathbf{t}_x^T \cdot \hat{\mathbf{B}} + \hat{t}_{e,HT}, \tag{4.2}$$

where $\mathbf{t}_x^T$ is the (transposed) vector of totals for $\mathbf{x}$, $\hat{t}_{e,HT}$ is the H-T estimator for the residual:

$$e_{ks} = y_k - \mathbf{x}_k^T \cdot \hat{\mathbf{B}} \ , \tag{4.3}$$

and

$$\hat{\mathbf{B}} = \left(\sum_r \frac{\mathbf{x}_k \mathbf{x}_k' \tau_k}{\pi_k} \right)^{-1} \sum_r \frac{\mathbf{x}_k y_k \tau_k}{\pi_k} \ , \tag{4.4}$$

π_k being the inclusion probability for element k. The choice of $\hat{\mathbf{B}}$ can be justified as the weighted least squares estimator of the model parameter β.

Since the GREG- estimator is non-linear some approximation is needed to estimate its variance. In Särndal, Swensson and Wretman (1992), section 6.6, the following large sample approximation is obtained by Taylor linearisation:

$$\hat{t}_{y,GREG} \approx \mathbf{t}_x^T \cdot \mathbf{B} + \hat{t}_{E,HT} \,, \tag{4.5}$$

where $\hat{t}_{E,HT}$ is the H-T estimator for $E_k = y_k - \mathbf{x}_k^T \cdot \mathbf{B}$, where $\mathbf{B}$ is the hypothetical least-squares estimator for the model parameter β if we could base the fit on the whole population.

If the residual E_k were known for the sample - which would require the full knowledge of $\mathbf{B}$ - then the usual formulas for variances and covariances, including the ones appearing in the section 3 above, would apply directly, at least as a large sample approximation, simply by replacing y with E. In particular, the procedure suggested in section 3 would only have to be modified by applying the data transformation (3.2)- -(3.7) to all the covariates x as well as to y.

However, $\mathbf{B}$ and, as a consequence, the residual E are unknown and must be estimated from the sample. One approximation would be to replace the unknown E_k with the estimate e_{ks} defined in (4.3) above. However, Särndal et al (1992) suggest a modified variance/covariance estimator where e_{ks} is replaced by the product $(g_{ks} \cdot e_{ks})$, the so called g-weight being defined as follows:

$$g_{gs} = \left\{ 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_x)^T \cdot \left(\sum_s \frac{\mathbf{x_k} \mathbf{x_k}^T \cdot \tau_k}{\pi_\mathbf{k}} \right)^{-1} \cdot \mathbf{x}_k \cdot \tau_k \right\} \tag{4.6}$$

If we, as suggested above, apply the data transformation (3.2) - - (3.7) to y and to every covariate x then it follows from (4.4) and (4.6) that $\hat{\mathbf{B}}$ and g_{ks} will be affected (which they should not be). However, we can adjust for this by modifying the scale factor τ_k in such a way that τ_k is replaced by $\tau_k \cdot \frac{n_h \cdot G_{hl}}{N_h \cdot \tilde{a}_{hl}}$ if $k \in s$ and by $\tau_k \cdot \frac{n'_l \cdot G_{hl}}{N'_l \cdot \tilde{a}_{hl}}$ if $k \in s'$ (c.f (3.6) and (3.7)).

After these modifications the procedure for the estimation of the covariance, as suggested in section 3 above, applies also when the GREG estimator is used as estimator for the population totals involved.

5. References

Andersson, C. and Nordberg, L. (1994): A Method for Variance Estimation of Non-Linear Functions of Totals in Surveys - Theory and a Software Implementation. Journal of Official Statistics, 10, 395-405.

Andersson, C. and Nordberg, L. (1998): A User's Guide to CLAN 97 - a SAS Program for Computation of Point- and Standard Error Estimates in Sample Surveys. Statistics Sweden.

Bäcklund, S. (1972): Tillämpning av JALES metod vid stratifierade urval. Skattning av en förändringsparameter. Bestämning av variansen i denna skattning när enheter byter stratum mellan två tidpunkter. Memo, Statistics Sweden. (In Swedish).

Garås, T. (1989): Förändringsestimatorer vid dynamiska populationer. Memo, Statistics Sweden. (In Swedish).

Nordberg, L. (1994): En Procedur för att Skatta Varianser för Förändringstal Baserade på Två Separate SAMU-Urval. Technical Report, Dept. for Economic Statistics, Statistics Sweden. (In Swedish).

Ohlsson, E. (1992): SAMU - The System for Co-Ordination of Samples from the Business Register at Statistics Sweden. R&D Report 1992:18, Statistics Sweden.

Ohlsson, E. (1995): Coordination of Samples Using Permanent Random Numbers. In Business Survey Methods. Ed. Cox, B. et al Wiley 1995.

Orusild, T. (1999): Confidence Intervals for Functions of Quantiles under Finite Population Sampling . Forthcoming Technical Report at Dept. of Mathematical Statistics, University of Stockholm.

Sjölinder, K. (1971): Beräkning av kovarianstermen vid urvalsdragning enligt JALES-metoden. Memo, Statistics Sweden. (In Swedish).

Särndal, C.E., Swensson, B. and Wretman, J. (1992): Model Assisted Survey Sampling. New York: Springer-Verlag.

Appendix A: Proof of the relation (2.7)

The point of departure will be the following expression - (2.6) in Section 2.

$$\hat{C}(\hat{t}_i, \hat{t}'_j|\Omega) =$$

$$\sum_h \sum_l \sum_{k\in P_{hl}} \sum_{r\in P_{hl}} \frac{N_h \cdot N'_l}{n_h \cdot n'_l} \cdot y_{ik} \cdot y'_{jr} \cdot (1 - \frac{E(\delta_k|\Omega) \cdot E(\delta'_r|\Omega)}{E(\delta_k \cdot \delta'_r|\Omega)}) \cdot \delta_k \cdot \delta'_r. \qquad \text{(A1)}$$

The three expectations must be computed in order to make the estimator (A1) operational. The following result - where all the quantities involved have been defined earlier in the main text - is shown, under a different notation, by Orusild (1999). An alternative proof follows here:

Lemma:

a) $E(\delta_k|\Omega) = \frac{a_{hl}}{G_{hl}}$, $k \in P_{hl}$,

b) $E(\delta'_r|\Omega) = \frac{a'_{hl}}{G_{hl}}$, $r \in P_{hl}$,

c) $E(\delta_k \delta'_k|\Omega) = \frac{g_{hl}}{G_{hl}}$, $k \in P_{hl}$,

d) $E(\delta_k \delta'_r|\Omega) = (\frac{a_{hl}a'_{hl} - g_{hl}}{G_{hl}(G_{hl}-1)})$, $k \neq r$, $k, r \in P_{hl}$.

Proof: a) By symmetry $E(\delta_k|\Omega)$ must be a constant for all $k \in P_{hl}$. Set this constant to τ_{hl}. Then $\sum_{k\in P_{hl}} E(\delta_k|\Omega) = \tau_{hl} \cdot G_{hl}$.

But $\sum_{k\in P_{hl}} E(\delta_k|\Omega) = E(\sum_{k\in P_{hl}} \delta_k|\Omega) = a_{hl}$ since $\sum_{k\in P_{hl}} \delta_k = a_{hl}$.

Hence $\tau_{hl} = \frac{a_{hl}}{G_{hl}}$.

b) proved analogously.

c) Let $E(\delta_k \delta'_k|\Omega) = \tau_{hl}$. Hence $\sum_{k\in P_{hl}} E(\delta_k \delta'_k|\Omega) = \tau_{hl} \cdot G_{hl}$.

But $\sum_{k\in P_{hl}} \delta_k \delta'_k \equiv g_{hl}$, i.e $\tau_{hl} = \frac{g_{hl}}{G_{hl}}$.

d) Let $E(\delta_k \delta_r' | \Omega) = \tau_{hl}$, k$\neq$r, since this expectation due to symmetry must be the same for all pairs k,r ; k$\neq$r in P_{hl}.

Now, $E(\delta_k \sum_{r \in P_{hl}} \delta_r' | \Omega) = E(\delta_k \cdot \delta_k' | \Omega) + \tau_{hl}(G_{hl} - 1)$.

But the left hand side is $E(\delta_k \cdot a_{hl}' | \Omega) = a_{hl}' E(\delta_k | \Omega) = \frac{a_{hl} \cdot a_{hl}'}{G_{hl}}$

Hence with assistance of c) above: $\frac{a_{hl} a_{hl}'}{G_{hl}} = \frac{g_{hl}}{G_{hl}} + \tau_{hl}(G_{hl} - 1)$ which proves point d of the lemma.

End of proof of lemma.

By the lemma and some algebra we have,

$$(1 - \frac{E(\delta_k | \Omega) E(\delta_r' | \Omega)}{E(\delta_k \delta_r' | \Omega)}) = 1 - \frac{\tilde{a}_{hl}}{G_{hl}}, \; k = r, \tag{A2}$$

$$(1 - \frac{E(\delta_k | \Omega) E(\delta_r' | \Omega)}{E(\delta_k \delta_r' | \Omega)}) = \frac{\tilde{a}_{hl} - G_{hl}}{G_{hl}(\tilde{a}_{hl} - 1)}, \; k \neq r. \tag{A3}$$

where

$$\tilde{a}_{hl} = \frac{a_{hl} \cdot a_{hl}'}{g_{hl}} \quad . \tag{A4}$$

By inserting (A2)--(A4) into (A1), finally, we arrive at the expression for the covariance estimator conditional on Ω, i.e. formula (2.7) in the main text.

$$\hat{C}(\hat{t}_i, \hat{t}_j' | \Omega) = \sum_h \sum_l A_{hl} \cdot \left\{ \sum_{k \in P_{hl}} y_{ik} \cdot y_{jk}' \delta_k \cdot \delta_k' - \frac{1}{\tilde{a}_{hl}} (\sum_{k \in P_{hl}} y_{ik} \delta_k)(\sum_{r \in P_{hl}} y_{jr}' \delta_r') \right\},$$

where $A_{hl} = \frac{N_h N_l' \tilde{a}_{hl}}{G_{hl}^2 n_h n_l'} (\frac{G_{hl}(G_{hl} - \tilde{a}_{hl})}{(\tilde{a}_{hl} - 1)})$.

Appendix B: On SAS-implementation of Procedure 3.1

To implement Procedure 3.1 it is necessary to merge information that is likely to be found in different computer files. We will here give an outline of a SAS program that may be useful for this implementation.

Suppose that the sample at time 0 is found in the SAS dataset Samp0:

```
Data Samp0;
length strat0 $ 5;
.
resp0=1;
keep id strat0 npop0 nresp0 resp0 y0;
proc sort; by id;
```

Comment: The variables in the keep-list are the identity of current element (e.g. enterprise), the stratum identity for the stratum to which 'id' belongs (we assume here that this stratum identity is a character of (say) five 'positions', hence the 'length' statement above), the number of population elements in current stratum, the number of responding elements in current stratum, a response indicator (to be used below) for current element and finally the observational value y for current element at time 0.

Analogously for sample at time 1:

```
Data Samp1;
length strat1 $ 5;
.
resp1=1;
keep id strat1 npop1 nresp1 resp1 y1;
proc sort; by id;
```

Next consider the frames for time 0 and 1:

```
Data Frame0;
length strat0 $ 5;.
.
keep id strat0;
proc sort; by id;

Data Frame1;
length strat1 $ 5;.
.
keep id strat1;
proc sort; by id;
```

/* Generating data sets D('death'), B ('birth') and P ('persistors') */

```
Data D;
merge Samp0 (in=a) Frame1 (in=b); by id;
if a and not b;

Data B;
merge Samp1 (in=a) Frame0 (in=b); by id;
if a and not b;

Data s0s1;
/* elements included in both samples */
merge Samp0 (in=a) Samp1 (in=b); by id;
if a and b;

Data s0F1_s1;
/* elements included in sample 0 and Frame1 but not in sample 1 */
merge Samp0 (in=a) Frame1 (in=b) Samp1 (in=c); by id;
if a and b and not c;

Data s1F0_s0;
/* elements included in sample1 and Frame0 but not in sample 0 */
merge Samp1 (in=a) Frame0 (in=b) Samp0 (in=c); by id;
if a and b and not c;

Data P;
set s0s1 s0F1_s1 s1F0_s0;
if resp0=1 and resp1=1 then resp01=1; else resp01=0;
if resp0 ne 1 then resp0=0;
if resp1 ne 1 then resp1=0;
proc sort; by strat0 strat1;
```

/* Computation of G_{hl}, here denoted 'glarge' */

```
data tempo; merge Frame0 (in=a) Frame1(in=b); by id;
if a and b;
proc sort; by strat0 strat1;
proc summary data=tempo; by strat0 strat1;
output out=gdata;
data gdata; set gdata;
rename _freq_=glarge;
```

/* Computation of $\tilde{a}_{hl}$ here denoted 'atilde', */

```
proc summary data=P; by strat0 strat1;
var resp0 resp1 resp01;
output out=adata sum=ahl aprhl gsmall;
run;
data adata; set adata;
if gsmall=0 then atilde=1;
if gsmall>0 then atilde=ahl*aprhl/gsmall;
run;
```

/* Transformations (3.2)—(3.7) in Procedure 3.1 */

```
data P;
merge P gdata adata; by strat0 strat1;
strat=strat0 || strat1;
nlarge=glarge;
nsmall=atilde;
if resp0=1 then z0=y0*(npop0*atilde) / (nresp0*glarge); else z0=0;
if resp1=1 then z1=y1*(npop1*atilde) / (nresp1*glarge); else z1=0;
run;

data D; set D;
strat=strat0 ||'00000';
nlarge=npop0;
nsmall=nresp0;
z0=y0; z1=0;
run;

data B; set B;
strat='00000' || strat1;
nlarge=npop1;
nsmall=nresp1;
z0=0; z1=y1;
run;

data indat; set D B P;
proc sort; by strat;
run;
```

Förteckning över utkomna R&D Reports

R&D Reports är en för IT-enheten och Metodenheten gemensam publikationsserie, som 1988-01-01 ersatte de tidigare "gula" och "gröna" serierna. I serien ingick fram till årsskiftet 1992-93 även **Abstracts** (sammanfattning av metodrapporter från SCB).

Reports published during 1996 and onwards:

1996:1 (grön)	On Sampling with Probability Proportional to Size *(Bengt Rosén)*
1996:2 (grön)	Bortfallsbarometern nr 11 *(Antti Ahtiainen, Stefan Berg, Margareta Eriksson, Åsa Greijer, Dan Hedlin, Monica Rennermalm och Anita Ullberg)*
1996:3 (grön)	Regression Estimators in Theory and in Practice *(Tomas Garås)*
1996:4 (gul)	Quality Aspects of a Modern Database Service *(Pat Dean and Bo Sundgren)*
1996:5 (gul)	Metadata: A Quality Element in Official Statistics - the Swedish Approach *(Bo Sundgren and Pat Dean)*
1996:6 (gul)	Our Legacy to Future Generation - Using Databases for Better Availability and Documentation *(Gösta Guteland and Erik Malmborg)*
1997:1 (grön)	Bortfallsbarometern nr 12 *(Anti Ahtiainen, Stefan Berg, Mats Bergdahl, Fredrik Granström, Dan Hedlin, Lena Otterskog och Monica Rennermalm)*
1997:2 (grön)	Quality Concept for Official Statistics - Entry in the forthcoming update of the Encyclopedia of Statistical Sciences, Wiley & Sons *(Eva Elvers and Bengt Rosén)*
1998:1 (grön)	Preliminär statistik: Nybyggnadskostnader - en simuleringstudie *(Catarina Elffors)*
1998:2 (grön)	On Inclusion Probabilities for Order Sampling *(Bengt Rosén)*
1998:3 (grön)	On the Stratification of Highly Skewed Populations *(Dan Hedlin)*
1998:4 (grön)	Bortfallsbarometer nr 13 *(Per Nilsson, Antti Ahtiainen, Stefan Berg, Mats Bergdahl, Monica Rennermalm och Marcus Vingren)*
1998:5 (grön)	Estimation from Order πps Samples with Non - Response *(Bengt Rosén and Pär Lundqvist)*
1998:6 (grön)	On variance estimation for measures of change when samples are coordinated by a permanent random numbers technique *(Lennart Nordberg)*

ISSN 0283-8680

Tidigare utgivna **R&D Reports** kan beställas genom Katarina Klingberg, SCB, MET, Box 24 300, 104 51 STOCKHOLM (telefon 08-783 42 82, fax 08-783 45 99, e-post katarina.klingberg@scb.se).
R&D Reports from 1988-1995 can - in case they are still in stock - be ordered from Statistics Sweden, attn. Katarina Klingberg, MET, Box 24 300, SE-104 51 STOCKHOLM (telephone +46 8 783 42 82, fax +46 8 783 45 99, e-mail katarina.klingberg@scb.se).

SCB-Tryck, Örebro 1999.01