Published in Survey Methodology 34, issue 2, 173-181, 2008
which should be used for any reference to this work

Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added

Lionel Qualité and Yves Tillé

Abstract

We propose a method for estimating the variance of estimators of changes over time, a method that takes account of all the components of these estimators: the sampling design, treatment of non-response, treatment of large companies, correlation of non-response from one wave to another, the effect of using a panel, robustification, and calibration using a ratio estimator. This method, which serves to determine the confidence intervals of changes over time, is then applied to the Swiss survey of value added.

Key Words: Covariance; Stratified sampling; Panel.

1. Introduction

In longitudinal surveys, the precision of changes over time depends directly on the rate of overlap of the samples. We begin by reviewing known results for disjoint simple designs (on this subject, see Kish 1965; Sen 1973; Wolter 1985; Laniel 1988; Hidiroglou, Särndal and Binder (1995); Holmes and Skinner 2000; Nordberg 2000; Fuller and Rao 2001; Berger 2004). Next, we calculate the variance of such changes for simple designs in which the samples overlap. When the sampling ratios are very low, most of these results are well known and are described, for example, in Caron and Ravalet (2000). Results that take account of finite population corrections can be seen in Tam (1984).

We precisely calculated the variances of estimators for a larger class of sampling designs with a finite population. Finite population corrections can play a major role in business surveys, since large companies are sometimes selected with very high probabilities of inclusion. The calculations become much more complicated with a finite population for the following reason: if the size of the population is finite, two disjoint samples are not independent. If the population is infinite, two independent samples are disjoint. Several estimators are examined: the difference of the cross-sectional estimators; the difference estimated solely on the common portion; and relative changes. The calculations become even more complex when the population is dynamic (with births, deaths, changes of structure). The theory that we develop below is limited to the case in which the population does not change over time.

In the first part, we describe the two-dimensional simple random sampling design (on this subject, see Goga 2003) and we give the corresponding Horvitz-Thompson estimators. We calculate the variance of the estimator of changes that is based on this sampling design. In a second part, we give the variance of other simple estimators: the relative change or the totals quotient, and the difference estimator based on the overlap of the samples. We then describe how these results adapt to the presence of ignorable non-response and the use of more complex estimators, which introduce weights modified to obtain calibrated estimators, or variables modified by a robustification procedure.

These results for simple designs are easy to generalize to stratified designs, provided that companies do not change strata from one wave to the next. Lastly, we apply this method to the Swiss survey of value added, taking all components of the survey into account: stratification, the panel effect, non-response, correlation between non-responses from one wave to the next, calibration using a ratio estimator, and robustification.

2. Estimation of the difference in simple designs

Let there be a population $U = \{1, \ldots, k, \ldots, N\}$ of size N in which two samples are taken: s_1 and s_2 of respective sizes n_1 and n_2. These samples may have a common portion (see Figure 1).

Assume that s_1 and s_2 are samples taken according to a simple design without replacement, and sizes n_1 and n_2 are therefore not random. Samples s_1 and s_2 can be broken down into three parts $s_A = s_1 \setminus s_2, s_B = s_2 \setminus s_1,$ and $s_C = s_C = s_1 \cap s_2$. Let $n_A = |s_A|, n_B = |s_B|, n_C = |s_C|, n_1 = n_A + n_C, n_2 = n_B + n_C$. The sizes of $s_A, s_B,$ and $s_C,$ may be random. This design generalizes the following hypothetical cases:

- If samples s_1 and s_2 are selected independently, n_C is then a random variable;
- If sample s_1 is selected first, and sample s_2 is selected in the complement of s_1 in U, then s_C is empty and $n_C = 0$;
- if sample s_1 is selected first, and sample s_2 consists of the union of a subsample of fixed size of s_1 and a sample of fixed size of the complement of s_1 in U, then n_C is not random, and the situation is the same as in case A of Tam (1984).

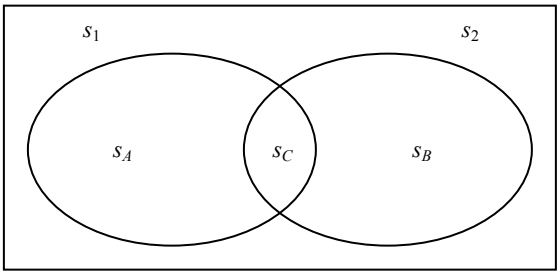

Figure 1 Overlapping samples

We make the additional hypothesis that conditional on $n_A, n_B,$ and $n_C,$ samples $s_A, s_B,$ and $s_C,$ are simple, without replacement and of fixed size. They come from the following sampling design:

Definition 1. Two-dimensional simple fixed-size sampling design (n_A, n_B, n_C):

$$p_{\text{simple}}(s_1, s_2 | n_A, n_B, n_C) = \begin{cases} \dfrac{n_A!\, n_B!\, n_C!\, (N - n_A - n_B - n_C)!}{N!} & \text{if } n_A = |s_A|, \\ & n_B = |s_B|, n_C = |s_C| \\ 0 & \text{otherwise,} \end{cases}$$

where $s_A = s_1 \setminus s_2, s_B = s_2 \setminus s_1$ and $s_C = s_1 \cap s_2$ (on this subject, see Goga 2003).

The law for drawing the pair $(s_1, s_2),$ which we do not know in general, is thus assumed to be of the form

$$p(s_1, s_2) = p_{\text{simple}}(s_1, s_2 | n_A, n_B, n_C) \Pr(|s_1 \cap s_2| = n_C).$$

Let there be two variables x and y whose values, taken on the units of U, are denoted respectively x_k and $y_k, k \in U$. Variables x and y may represent the same variable measured at two different times. Also assume that x can be observed only for s_1 and y for s_2. The objective is to estimate the totals

$$X = \sum_{k \in U} x_k \text{ and } Y = \sum_{k \in U} y_k,$$

as well as the difference $Y - X$. The Horvitz-Thompson estimators of X and Y are given by

$$\hat{X}_1 = \frac{N}{n_1} \sum_{k \in s_1} x_k \text{ and } \hat{Y}_2 = \frac{N}{n_2} \sum_{k \in s_2} y_k.$$

2.1 Natural estimation of the difference

2.1.1 Variance of the estimation of the difference

A first approach for estimating $\Delta = Y - X$ is to use the difference of the cross-sectional estimators $\hat{\Delta} = \hat{Y}_2 - \hat{X}_1,$ which is an unbiased estimator conditional on n_C according to the following simple design:

$$E(\hat{\Delta} | n_C) = Y - X,$$

and is therefore also unbiased under design p unconditional on n_C.

Proposition 1: The variance of $\hat{\Delta}$ is:

$$\text{var}(\hat{\Delta}) = N^2 \left(\frac{1}{n_1} - \frac{1}{N} \right) S_x^2 + N^2 \left(\frac{1}{n_2} - \frac{1}{N} \right) S_y^2 - 2N^2 \left(\frac{E(n_C)}{n_1 n_2} - \frac{1}{N} \right) S_{xy}, \quad (1)$$

where

$$S_x^2 = \frac{1}{N-1} \sum_{k \in U} (x_k - \bar{X})^2, S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{Y})^2,$$

$$S_{xy} = \frac{1}{N-1} \sum_{k \in U} (x_k - \bar{X})(y_k - \bar{Y}).$$

The demonstration of (1) is appended.

2.1.2 Specific cases and precision gain

Result (1) can be used to deal directly with the following specific cases of co-ordination:

- if the two samples form a panel, $n_C = n_1 = n_2,$ then

  $$\text{var}(\hat{\Delta}) = N^2 \left(\frac{1}{n_C} - \frac{1}{N} \right) (S_x^2 + S_y^2 - 2S_{xy});$$

- if the samples are disjoint (also see Ardilly and Tillé 2003, pages 24-28), $n_C = 0,$ and

  $$\text{var}(\hat{\Delta}) = N^2 \left(\frac{1}{n_1} - \frac{1}{N} \right) S_x^2 + N^2 \left(\frac{1}{n_2} - \frac{1}{N} \right) S_y^2 + 2NS_{xy}.$$

  Surprisingly, the covariance does not depend on the sizes of the samples. It is negative if x and y are positively correlated, and it becomes negligible in relation to the variance terms when the size of the population is large;

- if q is the set rate of overlap of the two samples and $n_1 = n_2 = n,$ we are back to case A developed by Tam (1984). We then obtain $n_C = qn,$ and

$$\text{var}(\hat{\Delta}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)(S_x^2+S_y^2) - 2N^2\left(\frac{q}{n}-\frac{1}{N}\right)S_{xy};$$

– if the two samples are independent, $E(n_C) = n_1 n_2 / N$, and we have

$$\text{var}_{\text{IND}}(\hat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)S_x^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)S_y^2.$$

If the size of the population is large and if the variables *x* and *y* have dispersions that are close to one another, the gain (or loss) of precision due to co-ordination in relation to the selection of two samples independently is

$$G = \frac{\text{var}(\hat{\Delta})}{\text{var}_{\text{IND}}(\hat{\Delta})} \approx 1 - \rho q, \qquad (2)$$

where ρ is the coefficient of correlation between *x* and *y*, $\rho = S_{xy}/S_x S_y$ and q is the overlap rate, $q = 2E(n_C)/(n_1+n_2)$. Expression (2) provides a simple multiplicative coefficient serving to take account of the effect of correlation and overlap.

2.1.3 Estimation of the variance of $\hat{\Delta}$

To estimate the variance, two cases must be considered:

– if $E(n_C)$ is known, which may be the case (for example, when the two samples are known to be independent), then

$$\widehat{\text{var}}(\hat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)s_{x1}^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)s_{y2}^2 - 2N^2\left(\frac{E(n_C)}{n_1 n_2}-\frac{1}{N}\right)s_{xyC}. \qquad (3)$$

where

$$s_{x1}^2 = \frac{1}{n_1-1}\sum_{s_1}(x_k-\bar{x}_1)^2,\ s_{y2}^2 = \frac{1}{n_2-1}\sum_{s_2}(y_k-\bar{y}_2)^2,$$

and

$$s_{xyC} = \frac{1}{n_C-1}\sum_{s_C}(x_k-\bar{x}_C)(y_k-\bar{y}_C).$$

This estimator is unbiased, but it can sometimes take on negative values;

– if $E(n_C)$ is not known, the only information concerning co-ordination is n_C.

$$\widehat{\text{var}}(\hat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)s_{x1}^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)s_{y2}^2 - 2N^2\left(\frac{n_C}{n_1 n_2}-\frac{1}{N}\right)s_{xyC}. \qquad (4)$$

This estimator is unbiased conditional on n_C and is therefore also unconditionally unbiased. It can also sometimes take on negative values. We will see further on that in some applications involving non-response, $E(n_C)$ is not known.

To use estimator (3), it is necessary to have at least two units in the overlap of the samples ($n_C \geq 2$), unless $E(n_C) = n_1 n_2 / N$. If $E(n_C) = n_1 n_2 / N$, which is the case where the two samples are independent, the third term of estimator (3) is nil. As to estimator (4), it is not defined when $n_C = 1$, unless $n_1 n_2 = N$.

2.2 Estimation using the common portion

The difference can also be estimated using only the common portion of the sample, which yields the estimator

$$\hat{\Delta}_C = N(\bar{y}_C - \bar{x}_C),$$

with $\bar{y}_C = 1/n_C \sum_{k\in s_C} y_k$ and $\bar{x}_C = 1/n_C \sum_{k\in s_C} x_k$. This estimator is unbiased unconditionally and conditionally on n_C.

2.2.1 Estimation of the variance of $\hat{\Delta}_C$

The conditional variance of $\hat{\Delta}_C$ is equal to

$$\text{var}(\hat{\Delta}_C | n_C) = N^2\left(\frac{1}{n_C}-\frac{1}{N}\right)(S_y^2 + S_x^2 - 2S_{xy}).$$

The unconditional variance is equal to

$$\text{var}(\hat{\Delta}_C) = N^2\left[E\left(\frac{1}{n_C}\right)-\frac{1}{N}\right](S_y^2 + S_x^2 - 2S_{xy}).$$

This unconditional variance may be difficult to calculate when n_C is random.

2.2.2 Comparison of the variances of $\hat{\Delta}$ and $\hat{\Delta}_C$

If we want to compare the two estimators of the difference, we can calculate

$$\text{var}(\hat{\Delta}) - \text{var}(\hat{\Delta}_C) = N^2\left[\frac{1}{n_1} - E\left(\frac{1}{n_C}\right)\right]S_y^2 + N^2\left[\frac{1}{n_2} - E\left(\frac{1}{n_C}\right)\right]S_x^2 - 2N^2\left[\frac{E(n_C)}{n_1 n_2} - E\left(\frac{1}{n_C}\right)\right]S_{xy}.$$

If $n_1 = n_2 = n, S_x^2 = S_y^2 = S^2$, and $E(1/n_C) \approx 1/E(n_C)$, then we obtain

$$\text{var}(\hat{\Delta}) - \text{var}(\hat{\Delta}_C) \approx \frac{1}{qn}[q-1]\,2N^2S^2 - 2\frac{1}{qn}[q^2-1]\,\rho N^2 S^2 = \frac{2N^2S^2}{qn}(1-q)\,[\rho(1+q)-1],$$

where $q = 2E(n_C)/(n_1 + n_2)$ is the overlap rate. The estimator $\hat{\Delta}_C$ is therefore more precise than $\hat{\Delta}$ if

$$\rho \geq \frac{1}{1+q}.$$

For example, if $q = 0.7$, it is preferable to use only the common portion once $\rho \geq 1/(1+0.7) \approx 0.588$ (on this subject, see Caron and Ravalet 2000, page 346). In cases where the overlap is sizable and the correlation is high, the estimator based on the difference of the cross-sectional estimators is therefore not very relevant.

3. Taking unit non-response into account

Non-response is considered to be independent of the selection design. According to the model, each unit decides randomly whether or not to respond, and the probabilities of response are equal between units. This is the most elementary model. However, if a unit does not respond in the first wave, it is highly probable that it will also not respond in the second wave. The model takes this dependency into account by considering separately four cases:

- the unit responds in both the first wave and the second;
- the unit responds in the first wave but not in the second;
- the unit does not respond in the first wave but it responds in the second;
- the unit responds in neither the first wave nor the second.

Non-response is commonly modelled by a multivariate Bernoullian design, which means that the probability of responding is the same for all statistical units and also that one unit decides to respond independently of the response of the other units. The non-response design is as follows:

$$q(r_A, r_B, r_C, r_D) = \phi_A^{\text{card}r_A} \phi_B^{\text{card}r_B} \phi_C^{\text{card}r_C} \phi_D^{\text{card}r_D},$$

where $r_A, r_B, r_C, r_D \subset U$, and r_A, r_B, r_C, r_D are mutually exclusive, and where

- $\phi_A^{\text{card}r_A}$ is the probability of responding in wave 1 but not in wave 2;
- $\phi_B^{\text{card}r_B}$ is the probability of responding in wave 2 but not in wave 1;
- $\phi_C^{\text{card}r_C}$ is the probability of responding in both wave 1 and wave 2;
- $\phi_D^{\text{card}r_D}$ is the probability of responding in neither wave 1 nor wave 2.

The modelled non-response phase thus consists in selecting four disjoint samples according to Bernoullian designs with different intensities. Since it is assumed to be independent of the sampling design, conditional on the sample sizes observed, the design resulting from the selection and the non-response is a simple multivariate design. If inference is conducted conditional on the sample sizes, the estimation of probabilities $\phi_A, \phi_B, \phi_C, \phi_D$ is not necessary and an unbiased inference can be conducted, as if dealing with a simple design. The theory of the preceding section therefore applies directly to the respondents, and all the information on the overlap of the two samples is found in $|s_C|$, regardless of whether this overlap is due to the design or to the link that exists between non-responses in the two waves. Note that even if the model is fairly simple, it takes account of the fact that if a unit has not responded in one wave, it will probably be less likely to respond in the following wave. Also, this model will be applied in relatively small, homogeneous strata.

4. Other measures of changes over time

The measurement of change over time is not always expressed in terms of differences. Such change is often measured in the form of a quotient or a relative difference. We therefore consider the following three measures:

- the difference $\hat{\Delta} = \hat{Y}_2 - \hat{X}_1$;
- the relative change $\hat{\Delta}_R = (\hat{Y}_2 - \hat{X}_1)/\hat{X}_1 = \hat{Y}_2/\hat{X}_1 - 1$;
- the quotient $\hat{Q} = \hat{Y}_2 / \hat{X}_1$.

The variance of $\hat{\Delta}$ may be expressed simply as a function of the estimators of variance of $\hat{Y}_2$ and $\hat{X}_1$ and the estimator of their covariance (see expression 4). The variance of $\hat{\Delta}_R$ is equal to the variance of $\hat{Q}$. They may be approached and then estimated using a residuals technique (on this subject, see Woodruff 1971; Binder and Patak 1994; Deville and Särndal 1992; Deville 1999),

$$\begin{aligned}\widehat{\text{var}}(\hat{\Delta}_R) &= \widehat{\text{var}}(\hat{Q}) \\ &= \frac{1}{\hat{X}_1^2}\left[\widehat{\text{var}}(\hat{Y}_2) + \hat{Q}^2 \widehat{\text{var}}(\hat{X}_1) - 2\hat{Q}\widehat{\text{cov}}(\hat{X}_1, \hat{Y}_2)\right].\end{aligned}$$

This variance can thus be simply estimated once we have estimators of $\text{var}(\hat{Y}_2)$, $\text{var}(\hat{X}_1)$ and $\text{cov}(\hat{X}_1, \hat{Y}_2)$.

5. Ratio estimation and robustification

Two techniques are commonly used for estimating the results of sample surveys: the use of a ratio estimator to

calibrate on the total of a dummy variable, and robustification of the estimators. These techniques must be taken into account in determining the precision of the final results.

5.1 Calibration

If an estimator is calibrated on known totals, the variance may be estimated simply by a residuals technique (see Woodruff 1971; Binder and Patak 1994; Deville and Särndal 1992; Deville 1999). For example, if $\mathbf{z}_{k1}$ and $\mathbf{z}_{k2}$ are column vectors of dummy variables on which the estimators $\hat{X}_{1\text{Cal}}$ and $\hat{Y}_{2\text{Cal}}$ are calibrated in waves 1 and 2, then the variances can be estimated by a residuals technique: $\text{var}(\hat{X}_{1\text{Cal}}) \approx \text{var}(\hat{E}_1)$ and $\text{var}(\hat{Y}_{2\text{Cal}}) \approx \text{var}(\hat{E}_2)$, where $\hat{E}_1$ et $\hat{E}_2$ are Horvitz-Thompson estimators of the totals of the residuals, with the latter being given for a simple design and for the generalized regression estimator by:

$$e_{k1} = x_k - \mathbf{z}'_{k1}\hat{\mathbf{B}}_1,$$
$$e_{k2} = y_k - \mathbf{z}'_{k2}\hat{\mathbf{B}}_2,$$

with

$$\hat{\mathbf{B}}_1 = \left(\sum_{k\in s_1} q_{k1}\mathbf{z}_{k1}\mathbf{z}'_{k1}\right)^{-1} \sum_{k\in s_1} q_{k1}\mathbf{z}_{k1}x_{k1},$$
$$\hat{\mathbf{B}}_2 = \left(\sum_{k\in s_2} q_{k2}\mathbf{z}_{k2}\mathbf{z}'_{k2}\right)^{-1} \sum_{k\in s_2} q_{k2}\mathbf{z}_{k2}y_{k2},$$

where $q_{kj}, j = 1, 2,$ is a coefficient that serves to take account of possible heteroscedasticity.

In the case of a sampling design with unequal probabilities, *e.g.*, a stratified sampling design such as in the Swiss survey of value added, the residuals are obtained by using a weighted regression. It is sufficient to replace $\hat{\mathbf{B}}_1$ and $\hat{\mathbf{B}}_2$ respectively by

$$\hat{\mathbf{B}}_1 = \left(\sum_{k\in s_1} \frac{q_{k1}\mathbf{z}_{k1}\mathbf{z}'_{k1}}{\pi_{k1}}\right)^{-1} \sum_{k\in s_1} \frac{q_{k1}\mathbf{z}_{k1}x_{k1}}{\pi_{k1}}, \quad \text{and} \qquad (5)$$

$$\hat{\mathbf{B}}_2 = \left(\sum_{k\in s_2} \frac{q_{k2}\mathbf{z}_{k2}\mathbf{z}'_{k2}}{\pi_{k2}}\right)^{-1} \sum_{k\in s_2} \frac{q_{k2}\mathbf{z}_{k2}y_{k2}}{\pi_{k2}}, \qquad (6)$$

where π_{kj} is the probability of inclusion of unit k in the sample for wave j, $j = 1, 2$.

5.2 Robustification

It is often useful to apply a robustification technique which offers a way to treat outliers. Simply consider that outliers have been detected and the weights of the individuals whose values are considered outliers have been modified by a factor $u_{kj}(s)$ in wave j. This factor is included between 0 and 1 and is equal to 1 for units that have values considered normal. The variance of the robustified estimator can be approached by advancing the classical hypothesis that weights $u_{kj}(s)$ depend only slightly on the sample s that was drawn (see Hulliger 1999). All that is needed, then, is to replace the variables x_k and y_k observed by $u_{k1}x_k$ and $u_{k2}y_k$ in the variance estimators.

By bringing together all the components of the mean square error of a change over time so as to take account of all components of that variance - namely the design, the panel effect, non-response, calibration and robustification - we obtain, for the relative change in a stratum,

$$\widehat{\text{MSE}}(\hat{\Delta}_R) = \widehat{\text{MSE}}(\hat{Q}) =$$
$$\frac{1}{\hat{X}_1}\left[\widehat{\text{var}}(\widehat{EU}_1) + \hat{Q}^2\,\widehat{\text{var}}(\widehat{EU}_1) - 2\hat{Q}\widehat{\text{cov}}(\widehat{EU}_1, \widehat{EU}_2)\right], \quad (7)$$

where

$$\hat{X}_1 = \frac{N}{m_1}\sum_{R_1} x_k, \;\; \hat{Y}_2 = \frac{N}{m_2}\sum_{R_2} y_k, \;\; \hat{Q} = \frac{\hat{Y}_2}{\hat{X}_1},$$

$$eu_{k1} = u_{k1}x_k - u_{k1}\mathbf{z}'_{k1}\hat{\mathbf{B}}_1,$$
$$eu_{k2} = u_{k2}y_k - u_{k2}\mathbf{z}'_{k2}\hat{\mathbf{B}}_2,$$

$$\widehat{EU}_j = \frac{N}{m_j}\sum_{R_j} eu_{kj}, \;\; \overline{EU}_j = \frac{\widehat{EU}_j}{N}, \;\; j = 1, 2,$$

$$\hat{\mathbf{B}}_1 = \left(\sum_{k\in D_1} \frac{q_{k1}u_{k1}^2\mathbf{z}_{k1}\mathbf{z}'_{k1}}{\pi_{k1}}\right)^{-1} \sum_{k\in D_1} \frac{q_{k1}u_{k1}^2\mathbf{z}_{k1}x_k}{\pi_{k1}},$$

$$\hat{\mathbf{B}}_2 = \left(\sum_{k\in D_2} \frac{q_{k2}u_{k2}^2\mathbf{z}_{k2}\mathbf{z}'_{k2}}{\pi_{k2}}\right)^{-1} \sum_{k\in D_2} \frac{q_{k2}u_{k2}^2\mathbf{z}_{k2}y_k}{\pi_{k2}}.$$

$$\widehat{\text{var}}(\widehat{EU}_j) =$$
$$N^2\left(\frac{1}{m_j} - \frac{1}{N}\right)\frac{1}{m_j - 1}\sum_{R_j}(eu_{kj} - \overline{EU}_j)^2, \; j = 1, 2,$$

$$\widehat{\text{cov}}(\widehat{EU}_1, \widehat{EU}_2) =$$
$$N^2\left(\frac{m_C}{m_1 m_2} - \frac{1}{N}\right)\frac{1}{m_C - 1}\sum_{R_C}(eu_{k1} - \overline{EU}_1)$$
$$\times (eu_{k2} - \overline{EU}_2).$$

R_1 and R_2 designate the set of respondents in the first and the second waves in the stratum, $m_1 = |R_1|$, $m_2 = |R_2|$, $|R_C = R_1 \cap R_2$ and $m_C = |R_1 \cap R_2|$. D_1 and D_2 are the sets of respondents in the two waves in the domain in which the calibration was carried out.

6. The Swiss survey of value added

6.1 Description of survey

The Swiss survey of value added is a survey of companies, conducted annually. Its purpose is to provide estimators of the main parameters of output in Switzerland: the value of gross output, the amount of intermediate consumption, the value added created by companies, and the cost of labour. The sampling design used is a stratified sampling of companies. In 1999, a sample of 11,210 companies (employing at least two persons) was selected and surveyed. This sample was run again in 2000 and 2001. Over that period, then, this is a panel survey. In the absence of a business register making it possible to identify births and deaths, the population of companies was considered constant during this period. The only adjustment to the annual data is made using a ratio estimation on the total of full-time equivalents (FTEs) per activity domain, available from an external source.

Stratification is based on the first two digits of the Nomenclature Générale des Activités économiques (general classification of economic activities) (NOGA2) and the size of the company (see Renfer 2000). In each activity stratum, three size strata are created: small companies employing 2-19 persons in FTE, medium-size companies, from 20 to M FTE, and large companies of more than M FTE. The stratum containing large companies is a take-all stratum, while small and medium-size companies are selected randomly with different sampling rates. The boundary M is chosen differently in each activity stratum in order to obtain optimum precision. In these three waves, approximately 6,000 establishments responded. The response rate for large companies, which all had to be surveyed, was close to 71% and was higher than the rate for small and medium-size companies. It was decided after the fact to treat some very large companies separately according to the "surprise" stratum methodology of Hidiroglou and Srinath (1981), considering that the response rate for the largest companies may well be better because they have an administrative structure better suited to responding to the survey questions. If they were assigned a weight equal to that of other large companies, this would introduce a bias as well as excessive variability. The "surprise" poststrata contain the 5% largest companies in the survey file. The latter were then considered as having, in effect, all been surveyed, and they received a weight of 1. No other treatment (calibration, robustification) was applied to them. The take-some strata consisting of small, medium-size and large companies were updated and some strata (size classes) containing few companies were later collapsed. If we accept the hypothesis that the very large companies were all taken, then the resulting estimator is unbiased and the variance related to very large companies is nil. We can therefore calculate only the variance in the other, updated strata.

During the survey, companies were again asked their category of economic activity. The estimates are based on these reported NOGA2s not on the NOGA2s in the sample frame. A calibration on the number of full-time equivalents (FTEs) provided by the business register is then conducted using a quotient estimator for the "reported" NOGA2 domains.

Finally, a robustification technique was used to lop the distribution of certain variables in the sample of small, medium-size and large companies (see Hulliger 1999; Peters, Renfer and Hulliger 2001). The weights of establishments whose values are considered outliers were modified by a factor $u_{kj}(s)$ included between 0 and 1. This factor is equal to 1 for companies that have values that are considered normal.

6.2 Variance of the change in value added

The objective is to estimate correctly the variance of estimators of change in value added (see Renfer 2000; Peters *et al.* 2001). In computing variances according to the hypothesis of independence of the samples, we largely overestimate the variance of changes, because the "value added" variables in times t_1 and t_2 are positively correlated. Correctly taking account of all aspects of the sampling design and the adjustment should provide better variance estimates. The study focuses on the 1999, 2000 and 2001 waves of the survey. Between these three dates, the raw sample was not modified. The fact that the sample remained fixed should make it possible to reliably estimate changes, but a response rate hovering around 50% may cause us to lose the benefit of the panel, if the number of respondents common to successive waves is low. The case of change between two survey waves where the sample has been updated, and where there are therefore two different raw samples and reference populations, is an entirely different problem.

In the present case, the fact that low variances were obtained can be attributed to the combined effect of several factors:

1. *Optimal design*: The sampling design was optimized. According to the optimal stratification, large companies have higher probabilities of inclusion. The stratum of companies contributing the most to value added is a take-all stratum. For this reason, the cross-sectional estimators have a low variance.

2. *High response fraction*: In the take-all stratum of large companies, the response rate approaches 70%. The finite population correction $(N-n)/N$

can therefore divide the variance by 3 compared to the case of an infinite population.

3. *Panel effect*: The sample is a panel, which is the best strategy for estimating changes over time.
4. *Correlation of non-response*: The non-response in one wave is strongly related to the previous wave and therefore does not greatly degrade the panel.
5. *Correlation of variables between waves*: The value added variables in times t and $t+1$ are highly correlated, since they are the same variable estimated at two different points in time.
6. *Calibration*: The estimators are calibrated in the strata on a variable related to the variable of interest; the variance of the estimators can then be written as a residual variance.

Of the 11,210 companies selected in 1999, approximately 5,200 responded in 1999 and 2000, and 5,300 responded in the 2000 and 2001 waves. Thus the size of the panel is relatively modest, and the treatment of non-response will therefore have a major impact on the results. To make variance estimates, we have assumed that non-response is ignorable (missing completely at random) within the take-some strata.

In each wave, estimates are made in the reported NOGA2 domains. This implies the possibility of a change of domain on the part of companies, and it is necessary to try to factor this into longitudinal estimates. We decided to ignore the impact of these changes initially, and to consider for the estimation of covariance that the domains are fixed and given by the value reported in the first of the two consecutive waves. This simplification is not inappropriate, since only 30 companies changed domain between 1999 and 2000, and only 25 did so between 2000 and 2001, representing respectively less than 0.5% and 0.2% of the FTEs in the sample. Calibration is carried out each year, and it can be taken into account using a residuals technique. As with estimating the variance of the cross-sectional estimators, robustification is taken into account by reweighting the survey variables.

With realistic assumptions, all components of the variance may be taken into account by means of the general expression (7). This expression is applied within each stratum and it covers all the components of the survey of value added: the panel effect, non-response, stratification, calibration and robustification. The estimators for the survey of value added are ratio estimators, and in this case the calculation of residuals is simplified. This is because in the case of the ratio, the regression coefficients given in (5) and (6) are calculated having only one dummy variable, and therefore $\mathbf{z}_{kj} = z_{kj}$ is scalar. Also, we take $q_{kj} = 1/z_{kj}$, for $j = 1, 2,$ and with robustification taken into account, we thus obtain:

$$eu_{k1} = u_{k1}x_k - \hat{B}_1 u_{k1} z_{k1},$$
$$eu_{k2} = u_{k2}y_k - \hat{B}_2 u_{k2} z_{k2},$$

where

$$\hat{B}_1 = \frac{\sum_{D_1} u_{k1}x_k / \pi_{k1}}{\sum_{D_1} u_{k1}z_{k1} / \pi_{k1}},$$
$$\hat{B}_2 = \frac{\sum_{D_2} u_{k2}y_k / \pi_{k2}}{\sum_{D_2} u_{k2}z_{k2} / \pi_{k2}}.$$

6.3 Variance estimation of changes

We made estimates of the standard deviations of changes in gross output values and value added figures calculated by the Swiss Federal Statistical Office. These estimates take into consideration all the aspects described above. We compared them with the estimated standard deviations that would have been obtained under the assumption that the draws for the different waves are independent. Over the various activity strata, the standard deviations that take account of the correlation between the survey waves are 41% lower than those based on the assumption of independence. This makes it possible to have much smaller confidence intervals than those calculated before this study, which were more quickly obtained but less precise. However, the gain is not the same in all activity strata. The following tables show standard deviations (SDs), calculated for the five largest activity strata (NOGA), of changes over time in the value of gross output (ΔOV) and of value added (ΔVA) between 1999 and 2000. The standard deviation that would have been obtained by ignoring the correlation between samples (SD_{ind}) is also included in the tables, along with the "gain" in precision realized by taking this correlation into account.

Table 1
Change in gross output value between 1999 and 2000 and standard deviations (in billions of Swiss francs)

Stratum	ΔOV	SD_{ind}	SD	Gain (%)
1	3.31	2.35	0.87	63
2	-0.77	4.38	1.98	55
3	3.07	2.11	0.94	56
4	4.33	1.10	1.00	09
5	-0.09	0.81	0.53	35

Table 2
Change in value added between 1999 and 2000 standard deviations (in billions of Swiss francs)

Stratum	ΔVA	SD_{ind}	SD	Gain (%)
1	1.96	0.91	0.32	65
2	0.68	2.99	1.04	65
3	1.90	1.47	0.72	51
4	0.36	0.47	0.45	05
5	-0.36	0.59	0.43	27

Acknowledgements

This study was carried out under an agreement between the University of Neuchâtel and the Swiss Federal Statistical Office. The findings published in this article are those of the authors alone and in no case do they commit the Federal Statistical Office. We wish to thank Paul-André Salamin for his contribution to this study.

Appendix

Demonstration of proposition 1

It is well known that

$$\text{var}(\hat{X}_1) = N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)S_x^2$$

and

$$\text{var}(\hat{Y}_2) = N^2\left(\frac{1}{n_2} - \frac{1}{N}\right)S_y^2.$$

It is thus sufficient to calculate $\text{cov}(\hat{X}_1, \hat{Y}_2)$. We note

$$\bar{x}_A = \frac{1}{n_A}\sum_{k\in s_A} x_k, \qquad \bar{x}_C = \frac{1}{n_C}\sum_{k\in s_C} x_k,$$

$$\bar{y}_B = \frac{1}{n_B}\sum_{k\in s_B} y_k, \qquad \bar{y}_C = \frac{1}{n_C}\sum_{k\in s_C} y_k,$$

$$\bar{x}_1 = \frac{n_A\bar{x}_A + n_C\bar{x}_C}{n_1}, \quad \bar{y}_2 = \frac{n_B\bar{y}_B + n_C\bar{y}_C}{n_2},$$

and therefore $\hat{X}_1 = N\bar{x}_1$ and $\hat{Y}_2 = N\bar{y}_2$. We must still calculate

$$\text{cov}(\bar{x}_1, \bar{y}_2) = E\,\text{cov}(\bar{x}_1, \bar{y}_2 \mid n_A, n_B, n_C) + \text{cov}\left[E(\bar{x}_1 \mid n_A, n_B, n_C), E(\bar{y}_2 \mid n_A, n_B, n_C)\right].$$

Since $\bar{x}_1$ and $\bar{y}_2$ are unbiased conditional on $n_A, n_B,$ and n_C,

$$\text{cov}\left[E(\bar{x}_1|n_A, n_B, n_C), E(\bar{y}_2|n_A, n_B, n_C)\right] = \text{cov}(\bar{X}, \bar{Y}) = 0.$$

We therefore obtain

$$\text{cov}(\bar{x}_1, \bar{y}_2) = E\,\text{cov}(\bar{x}_1, \bar{y}_2|n_A, n_B, n_C).$$

Conditional on $n_A, n_B,$ and $n_C,$ we are in case A of Tam (1984, theorem 1). The conditional variance is equal to

$$\text{cov}(\bar{x}_1, \bar{y}_2|n_A, n_B, n_C) = \left(\frac{n_C}{n_1 n_2} - \frac{1}{N}\right)S_{xy}$$

and therefore

$$\text{cov}(\bar{x}_1, \bar{y}_2) = \left(\frac{E(n_C)}{n_1 n_2} - \frac{1}{N}\right)S_{xy}.$$

Now,

$$\text{cov}(\hat{X}_1, \hat{Y}_2) = N^2\,\text{cov}(\bar{x}_1, \bar{y}_2),$$

enabling us to obtain the result (1).

References

Ardilly, P., and Tillé, Y. (2003). *Exercices corrigés de méthodes de sondage*. Paris: Ellipses.

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, 32, 4, 451-467.

Binder, D.A., and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

Caron, N., and Ravalet, P. (2000). Estimation dans les enquêtes répétées : application à l'enquête Emploi en continu. Technical report, 0005. Méthodologie Statistique, INSEE, Paris.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fuller, W.A., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.

Goga, C. (2003). Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliaire par des modèles nonparamétriques. Ph.D. Dissertation, Université de Rennes II, Haute Bretagne, France.

Hidiroglou, M., Särndal, C.-E. and Binder, D. (1995). Weighting and Estimation in Business Surveys. *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M. Colledge and P.S. Kott), New York: John Wiley & Sons, Inc., 477-502.

Hidiroglou, M.A., and Srinath, K.P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.

Holmes, D.J., and Skinner, C.J. (2000). Variance Estimation for Labour Force Survey Estimates of Level and Change. Technical report, Government Statistical Service Methodology Series, 21, London, England.

Hulliger, B. (1999). Simple and robust estimators for sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 54-63.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Laniel, N. 1988. (1988). Variances for a rotating sample from a changing population. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 246-250.

Nordberg, L. (2000). On variance estimation for measure of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

Peters, R., Renfer, J.-P. and Hulliger, B. (2001). Statistique de la valeur ajoutée : procédure d'extrapolation des données. Technical report, Swiss Federal Statistical Office.

Renfer, J.-P. (2000). Enquête sur la production et la valeur ajoutée : échantillonnage complémentaire. Technical report, Swiss Federal Statistical Office.

Sen, A.R. (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics*, 29, 381-385.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, (4), 288-289.

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

Woodruff, R.S. (1971). A simple method for approximating de variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.