Estimation of change in a rotation panel design

Andersson, Claes
Statistics Sweden
S-701 89 Örebro, Sweden
E-mail: claes.andersson@scb.se

Andersson, Karin
Statistics Sweden
P.O. 24 300
S-104 51 Stockholm, Sweden
E-mail: karin.andersson@scb.se

Lundquist, Peter
Statistics Sweden
P.O. 24 300
S-104 51 Stockholm, Sweden
E-mail: peter.lundquist@scb.se

Introduction

Repeated sample surveys are a rule rather than an exception within National Statistical Institutes (NSIs). It is of importance to estimate the value of different parameters, e.g. the number of unemployed, at the reference time point but also to get reliable estimates of change in the parameters between the reference times. A common design for this is to use rotating panel designs meaning that a part of the sample at time point 0 is retained at time point 1 and a complementary sample is taken at time point 1. The variance of an estimated difference between parameters at time 0 and 1 is a function of the variances at each time point and the covariance between the estimated parameters. The estimation of the variance of a difference between two totals based on partially overlapping samples is described in several papers (Qualité & Tillé 2008; Wood 2008; Berger 2004; Tam 1984). The case when the sampling rates are low and simple random sampling is used was treated by Kish (1965).

In this paper we will treat the situation with a fix population and a rotating sample design that means fix sample sizes at both occasions and a fix overlapping rate. The assumption of a fix population is an approximation that is valid in many situations where individuals are sampled and the time lag is short between the reference times of the measurements, for example in a monthly or quarterly LFS. The focus will be on methods for variance estimation that can be used in practice in an NSI when auxiliary information is used in the estimation process and where the number of parameters is large and typically for domains that as a rule do not coincide with the stratification.

Different methods of doing the variance estimation of the estimated changes are investigated; the relatively simple method used for LFS within Statistics Sweden is compared with two other methods.

The Sample Design

Let $U=(1, \ldots, k, \ldots, N)$ be a population of size N in which two samples s_0 and s_1 of size n_0 and n_1 are taken. At time point 0, the possibly stratified sample s_0 is taken according to the (without replacement) sampling design $p_0(s_0)$ with first order inclusion probabilities π_{0k}. At time point 1, let s_{01} be a randomly selected fraction $g=n_{01}/n_0$ of s_0, which is joined by the sample $s_{1|0}$ of size $n_{1|0}$ which is taken from U-s_0 to form the sample s_1 of size n_1, i.e. $s_1=s_{01}\cup s_{1|0}$ and $s_{01}=s_0\cap s_1$.

In this paper we will only treat the case when s_0 is a stratified simple random sample (srswor), the stra-

tification is the same at time point 0 and 1, the remaining (overlapping) part s_{01} and the complement $s_{1|0}$ are taken as srswor within strata. This procedure is a special case of the situation described in Berger (2004) and it is in accordance with the design for the Swedish LFS with overlapping fraction g=7/8 between two adjacent quarters.

The first order inclusion probabilities in case described in this paper is $\pi_{0k}=n_h/N_h$, $k\in h$ (stratum h), $\pi_{1k}=g_h\pi_{0k}+q_h(1-\pi_{0k})$, $k\in h$, where $g_h=n_{01h}/n_{0h}$, $q_h=n_{1|0h}/(N_h-n_{0h})$. The reason why we use stratum indices for g and q is the fact that in practice we will get nonresponse and will condition on the number of respondents within each stratum.

The Estimation problem

The parameter of interest here is $\theta=t_1-t_0$, where $t_0=\sum_U y_{0k}$, $t_1=\sum_U y_{1k}$ and y_{0k}, y_{1k} are the values of variable y at time 0 and 1 for unit k. The parameter is estimated by $\hat{\theta}=\hat{t}_1-\hat{t}_0$, where $\hat{t}_i$, i=0, 1 is some estimator of t_i, like the Horvitz-Thompson (H-T) estimator or a generalized regression (GREG) estimator.

The variance of $\hat{\theta}$ is $V(\hat{\theta})=V(\hat{t}_1)+V(\hat{t}_0)-2C(\hat{t}_1,\hat{t}_0)$, where $C(\hat{t}_1,\hat{t}_0)$ is the covariance between $\hat{t}_1$ and $\hat{t}_0$. The variance is estimated by $\hat{V}(\hat{\theta})=\hat{V}(\hat{t}_1)+\hat{V}(\hat{t}_0)-2\hat{C}(\hat{t}_1,\hat{t}_0)$, where $\hat{V}(\hat{t}_i)$, i=0, 1 is easily calculated from the sample s_i by standard methods.

The covariance may be estimated from the common part s_{01} of the sample, for example when stratified srswor is used together with the H-T-estimator and $n_{0h}=n_{1h}$, (Qualité & Tillé (2008), Tam (1984), although neither of them gave explicit results for stratification it should be straight forward),

$$\hat{C}_{01}(\hat{t}_1,\hat{t}_0)=\sum_h N_h^2(g_h/n_{1h}-1/N_h)s_{y01h},$$

where s_{y01h} is the simple covariance between y_0 and y_1 calculated from the sample s_{01h}.

It was proposed by Kish (1965) to estimate the covariance from a combination of the correlation estimated from s_{01} and the variances estimated from s_1 and s_0,

$$\hat{C}_K(\hat{t}_1,\hat{t}_0)=\hat{\rho}_{\hat{t}:01}\sqrt{\hat{V}_1(\hat{t}_1)\hat{V}_0(\hat{t}_0)};\quad \hat{\rho}_{\hat{t}:01}=\hat{C}_{01}(\hat{t}_1,\hat{t}_0)\Big/\sqrt{\hat{V}_{01}(\hat{t}_1)\hat{V}_{01}(\hat{t}_0)},$$

where $\hat{V}_i(\cdot)$ is calculated from the sample s_i, i=0, 1 and $\hat{V}_{01}(\cdot)$ is calculated from the common part s_{01}. Although Kish used the standard with replacement estimator, the same principle can be used for other types of estimators.

If $V(\hat{t}_0)\approx V(\hat{t}_1)$ then the variance of $\hat{\theta}$ may be approximated by $AV(\hat{\theta})=2V(\hat{t}_1)(1-\rho_{\hat{t}:01})$ with an obvious simple estimator $2V(\hat{t}_1)(1-\rho_{\hat{t}:01})$.

In most, if not all, surveys we will get nonresponse, it was noted by Qualité & Tillé (2008) that by conditioning on the respondents we can use the response sets instead of the sample sets in the formulas. However, the risk for nonrespons bias is always there and a positive bias in $\hat{\rho}_{01}$ based on the response set r_{01} is likely in practice as the correlation tends to be stronger among units responding at both occasions. It was shown by Berger (2004) that even a small positive bias in $\hat{\rho}_{01}$ implies a large negative bias in $\hat{V}(\hat{\theta})$ when the correlation is large.

It should be noted here that even if other parameters for measuring change than the difference between totals may be of interest, for example the ratio, the problem in the estimation of the covariance in a rotating design will be the same.

Berger (2004) suggested an estimator of the covariance by assuming that the vector $(\hat{\mathbf{t}}',\mathbf{n}')=(\hat{t}_1,\hat{t}_0,\mathbf{n}_1',\mathbf{n}_0',\mathbf{n}_{01}')$ has a multivariate normal distribution under the Poisson sampling scheme with covariance matrix,

$$\boldsymbol{\Sigma}_{\mathbf{u}}=\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{tt}} & \boldsymbol{\Sigma}_{\mathbf{nt}}'\\ \boldsymbol{\Sigma}_{\mathbf{nt}} & \boldsymbol{\Sigma}_{\mathbf{nn}}\end{bmatrix}$$

where $\mathbf{n} = (\mathbf{n}_1', \mathbf{n}_0', \mathbf{n}_{01}')'$ is the vector of sample sizes within each stratum for the samples s_0, s_1 and s_{01}.

The covariance matrix of $\hat{\mathbf{t}}$ is obtained by conditioning on the vector $\mathbf{n}$ using standard theory, $\boldsymbol{\Sigma}_{t|n} = \boldsymbol{\Sigma}_{\mathbf{tt}} - \boldsymbol{\Sigma}_{\mathbf{nt}}' \boldsymbol{\Sigma}_{\mathbf{nn}}^{-1} \boldsymbol{\Sigma}_{\mathbf{nt}}$, which is estimated by, $\hat{\boldsymbol{\Sigma}}_{t|n} = \hat{\boldsymbol{\Sigma}}_{\mathbf{tt}} - \hat{\boldsymbol{\Sigma}}_{\mathbf{nt}}' \hat{\boldsymbol{\Sigma}}_{\mathbf{nn}}^{-1} \hat{\boldsymbol{\Sigma}}_{\mathbf{nt}}$, see Berger (2004).

The method is quite general and can handle situations where s_0 does not involve srs and where s_1 need not be taken within each stratum but can be taken independent of the stratification at time point 0. It also uses all information in the two samples rather than just the common part of the two samples.

Berger (2004) suggests that the covariance is calculated by the same principle as $\hat{C}_K$ above, i.e. $\hat{\rho}_{01}$ is calculated from $\hat{\boldsymbol{\Sigma}}_{t|n}$ as defined above while the variances are estimated at time points 0 and 1 by conditioning on $\mathbf{n}_0$ and $\mathbf{n}_1$ respectively, leading to the Hajek (1964) estimator of the variance. He also suggests a simple way to do the estimation of the covariance by using standard multiple regression software.

Normally, a large number of differences are of interest for an NSI, each regarding a domain of study that not necessarily coincides with a stratum in the design. This means that the vector $\hat{\mathbf{t}}$ normally is of length >2. Further, when auxiliary information is used in the estimation of $\hat{\mathbf{t}}$ by regression estimators or calibration against know totals, Qualité & Tillé (2008) pointed out that the variables of interest in calculating the covariance are not y_0 and y_1 but the residuals e_0 and e_1.

The residuals used in domain d may be defined as $e_{idk} = y_{idk} - \mathbf{x}_{ik}' \hat{\boldsymbol{\beta}}_{id}$, i=0,1 where y_{idk}=y_{ik} when unit k belongs to domain d, and y_{idk}=0 otherwise. The auxiliary vector $\mathbf{x}$ does not have to be the same or to have the same values for unit k at both occasions.

The Swedish LFS used in a simulation study

The Swedish LFS is a monthly survey where each selected individual attend once each quarter during two years. The sample design is stratified srswor of about 29 500 individuals each month where one eights of the sample is replaced each month. This means that g=7/8 for the estimate of change between two quarters and g=4/8 for the estimate of change between a year. Note that there is no overlap between the samples for two adjacent months.

The GREG-estimator is used in the estimation where the auxiliary information is updated continuously.

Data from the Swedish LFS will be used to illustrate three different suggested methods for the estimation of change as defined in this paper. Two sets of data will be used, 1) "Quarter", the common part, 39 131 individuals, between two quarters in two different, adjacent quarters, 2) "Year", the common part, 21 671 individuals, between the same quarters two adjacent years.

It is of interest to find reliable variance estimators that can be used in a large scale production of statistics, especially for the LFS which is produced monthly.

Estimators used in the simulation study

The variance estimator of the estimated change in the study is $\hat{V}(\hat{\theta}) = \hat{V}(\hat{t}_1) + \hat{V}(\hat{t}_0) - 2\hat{C}(\hat{t}_1, \hat{t}_0)$, where the covariance is estimated by $\hat{\rho}_{\hat{t}:01} \sqrt{\hat{V}_1(\hat{t}_1)\hat{V}_0(\hat{t}_0)}$. The correlation $\hat{\rho}_{\hat{t}:01}$ between the estimated totals is calculated in three different ways.

In the old days when the regression estimator was not used and the calculation of standard errors was cumbersome the estimation of change was done in a rather naive way. The correlation $\hat{\rho}_{y;01}$ between the y-values at time points 0 and 1 was calculated, based on the common part s_{01}, and not using the design, i.e. by treating s_{01} as generated by srs. This made sense to some extent since the allocation of the sample was proportional to stratum sizes the nonresponse rates were low and the HT-estimator was used. It was also observed that the changes in the calculated correlations were small over time meaning that a set of correlations or "constants" could be used over a longer period, typically several years. Our first estimator uses the correlation $\hat{\rho}_{\hat{t};01} = g\hat{\rho}_{y;01}$.

The second estimator in our study uses the residuals that come from the GREG estimator at the two time points. The correlation is calculated using the residuals $e_{ik} = y_{ik} - \mathbf{x}'_{ik}\hat{\boldsymbol{\beta}}_i$ and using the design weights in the calculations,

$$\hat{\rho}_{e:01} = s_{e:01} \Big/ \sqrt{s^2_{e:0} s^2_{e:1}} ,$$

with $s_{e:01} = (\sum_{s01} w_k e_{0k} e_{1k} - \sum_{s01} w_k e_{0k} \sum_{s01} w_k e_{1k} / \sum_{s01} w_k) \big/ ((\sum_{s01} w_k) - 1)$; $w_k = N_h / m_h$, $k \in h$, N_h is the stratum size and m_h is the number of units in s_{01h}. The residuals are calculated from s_0 and s_1, i.e. the whole samples are used at time 0 and 1. The second estimator uses $\hat{\rho}_{\hat{t}:01} = g\hat{\rho}_{e:01}$.

The third estimator is the one suggested by Berger (2004). In the calculation of $\hat{\rho}_{\hat{t}:01}$ the inversion of the matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{nn}}$ may be cumbersome when the number of strata is large, it is a $3H \times 3H$ symmetric matrix, where H is the number of strata. However, it turns out that the matrix contains a number of diagonal sub matrices which simplifies the storage and inversion by using the theory for partitioned matrices.

The idea behind the second and third estimator is to find a better way to calculate the "constants" (correlations) to be used in the present production framework within Statistics Sweden.

The simulation study

A number of cases were setup for the simulation study but only a limited number will be reported here.

The two populations were stratified into 3 strata by age. From each stratum a srswor of 500 individuals were taken at time point 0, and at time point 1 two different g were used, 7/8 and 4/8. Four parameters were estimated, the difference of the number of individuals employed, unemployed, not in the labor force, total number of hours worked by employed individuals. These parameters were also estimated for different domains but the results will not be reported here.

The population parameters and their variances are shown for some of the cases in Table 1. The variances and correlations between $\hat{t}_1$ and $\hat{t}_0$ were calculated, both for the HT- and the GREG-estimators. The auxiliary vectors contained two categorical variables, 1) registered as job seekers or not, and 2) six age classes. Note that the use of auxiliary variables decreases the variances of the estimated totals but also the correlation between them, which is to be expected. The effect on the estimated difference is then dependent on the relation between these decreases. For the quarter data the decrease in $V(\hat{t}_0)$ is for example about 22% for the total "employed" while the variance of the difference decreases by about 11% when GREG is used instead of HT. For the year data the figures are 25% and 20%. The decreases in the correlations are 6% and 10%.

For each setup 1 000 random samples were generated according to the design and for each sample the point estimate of the change was calculated together with the three different estimates of the variance and the related indicators of the coverage of a calculated 95% confidence interval (CI95).

Some results

In table 2 the results from the simulation study are shown. In column A is the variance of the differences calculated from the population, these are the theoretical values that the three different estimators are supposed to be estimators of.

In columns B, E and H are the relative differences, $R = (\bar{\hat{V}} / V - 1)100$ between the means of the estimated variances, $\bar{\hat{V}}$ for each estimator. It is obvious that the first, naïve, estimator underestimate the variance, the bias is larger for g=7/8 than for g=4/8. This is not unexpected since the correlation between the y-values is larger than between the e-values, while the variances for the totals are estimated by GREG. The differences between the second and third estimators are small in all cases.

In columns D, G and J (CI95) are the coverage rates for nominal 95% confidence intervals calculated for each sample. As expected the first estimator gives too short intervals, although the coverage rates are not

that low, considering the underestimate of the variances. For the second and third estimator the coverage rates are close to the nominal value, 95%.

Table 1

	Estimator	HT				GREG				
		A	B	C	D	E	F	G	H	I
g=7/8		$V(\hat{t}_0)$	$V(\hat{t}_1)$	$\rho_{\hat{t}:01}$	$V(\hat{t}_1-\hat{t}_0)$	$V(\hat{t}_0)$	$V(\hat{t}_1)$	$\rho_{\hat{t}:01}$	$V(\hat{t}_1-\hat{t}_0)$	(H/D-1)%
Quarter	Employed	82 446	82 990	0.73	45 169	64 241	64 065	0.69	40 260	-10.9
N=39 131	Unemployed	22 619	20 996	0.47	23 200	17 674	16 823	0.38	21 375	-7.9
	Not in LF	69 578	71 850	0.73	38 895	56 591	57 347	0.69	35 621	-8.4
	Hours worked	186.1	174.3	0.39	218.5	176.5	154.7	0.36	212.4	-2.8
g=4/8										
Year	Employed	24 214	23 049	0.30	32 890	18 119	17 981	0.27	26 255	-20.2
N=21 671	Unemployed	6 882	5 922	0.14	11 012	4 776	4 757	0.10	8 556	-22.3
	Not in LF	20 611	19 710	0.30	28 209	16 650	15 984	0.27	23 839	-15.5
	Hours worked	50.5	49.9	0.27	73.4	44.5	44.7	0.25	66.7	-9.1

Table 2

	Estimator		1			2			3		
		A	B	C	D	E	F	G	H	I	J
g=7/8		$V(\hat{t}_1-\hat{t}_0)$	$R^{(1)}$	$\bar{\hat{\rho}}_{\hat{t}:01}$	CI95$^{(1)}$	$R^{(2)}$	$\bar{\hat{\rho}}_{\hat{t}:01}$	CI95$^{(2)}$	$R^{(3)}$	$\bar{\hat{\rho}}_{\hat{t}:01}$	CI95$^{(3)}$
Quarter	Employed	40 260	-20.8	0.75	92.0	-2.8	0.69	94.6	-3.1	0.70	94.7
N=39 131	Unemployed	21 375	-14.6	0.47	93.1	-0.8	0.39	94.8	-1.6	0.39	94.9
	Not in LF	35 621	-21.0	0.75	93.1	-2.9	0.70	95.3	-3.0	0.70	95.3
	Hours worked	212.4	-12.4	0.44	95.0	-2.6	0.38	96.1	-1.3	0.37	96.2
Year	Employed	15 969	-23.5	0.66	91.7	-2.9	0.57	95.1	-2.9	0.57	95.1
N=21 671	Unemployed	7 555	-9.9	0.29	94.4	0.1	0.21	95.4	-0.5	0.21	95.3
	Not in LF	14 623	-23.8	0.66	92.3	-2.6	0.56	95.8	-2.8	0.56	95.8
	Hours worked	43.7	-16.0	0.59	93.5	-2.8	0.52	94.8	-2.4	0.52	94.8
g=4/8											
Quarter	Employed	81 703	-10.7	0.43	94.1	-5.6	0.40	94.6	-5.7	0.40	94.7
	Unemployed	27 527	-8.9	0.27	93.1	-2.7	0.22	94.1	-3.2	0.23	93.9
	Not in LF	72 519	-10.8	0.43	94.3	-5.7	0.40	94.8	-5.7	0.40	94.9
	Hours worked	268.2	-7.7	0.25	93.8	-3.2	0.22	94.6	-2.6	0.21	94.6
Year	Employed	26 255	-14.4	0.38	92.9	-7.3	0.33	94.0	-7.2	0.33	93.9
	Unemployed	8 556	-6.8	0.16	94.8	-1.7	0.12	95.3	-2.0	0.12	95.3
	Not in LF	23 839	-14.6	0.38	93.7	-7.1	0.32	94.4	-7.2	0.32	94.4
	Hours worked	66.7	-11.3	0.34	93.7	-6.3	0.30	94.2	-6.2	0.30	94.3

Discussion

The way the first estimator is applied in the simulation study does not fully reflect the way it is used in the Swedish LFS. As mentioned before the correlation between the *y*-values are calculated for one year and then used for many years thereafter until it is assumed that new values are needed. To mimic that situation a fixed set of correlation should be calculated and then be used for all samples.

The second estimator is a bit more advanced than the first in the sense that it takes the design into account as well as the fact that the GREG-estimator is used in the estimation of the totals at each time point. The estimator, $\hat{\rho}_{e:01}$ is basically an estimator of the correlation between the residuals at time point 0 and 1 in the population and it is not obvious how it relates to the correlation between the estimated totals under the present design, however, it seems to work, at least in the cases shown here. A design based estimator of the correlation between the estimated totals would have been constructed in a different way.

The third estimator takes both the design and the residuals into account and it has a theoretical basis, although some approximations are made. It is also more complicated to understand and maybe not very transparent for non-experts. However, a procedure based on a theory should be preferred to ad-hoc based procedures.

The adaption of the second and third estimators in a large scale production system for official statistics is a bit complicated but can be solved. One obvious way is to calculate a set of correlation for one year and use that set for a number of years like it is done today in the Swedish LFS.

REFERENCES

Berger, Y. G. (2004). Variance Estimation of Measures of Change in Probability Sampling. Canadian Journal of Statistics, 32, pp. 451-467.

Hajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics, 35, pp. 1491-1523.

Kish, L. (1965). Survey Sampling: Wiley, New York.

Qualité, L. and Tillé, Y. (2008). Variance Estimation of Changes in Repeated Surveys and its Application to the Swiss Survey of Value Added. Survey Methodology, 34, pp 173-181.

Tam, S. M. (1984). On Covariances from overlapping Samples. The American Statistician, 38, pp. 288-289.

Wood, J. (2008). On the Covariance Between Related Horvitz-Thompson Estimators. Journal of Official Statistics, 24, pp. 53-78.