LMD konf. darbai, 298–303
1997, Vilnius

On some estimators in cluster sampling

G. Klimavičius, D. Krapavickaitė, A. Plikusas (MII)

INTRODUCTION

The aim of the study is to investigate and compare the precision of some kinds of estimators in the Lithuanian labour force survey (LFS).

The distribution of the surveyed individuals in any sample is not exactly the same as the population distribution. Thus, reweighting of the sample using auxiliary information, or poststratification, may be used in order to "correct" the distribution of the sample. This method is used in Statistics Finland by P. Väisänen and K. Djerf ([4]), in Statistics Sweden by J. Hörngren ([3]). Lithuanian official statistics did not use any special methods for the increase of precision of the estimates before, because this survey was carried out only several times. We have studied the effects of poststratification on the accuracy of estimates using demographical data and information on the registered unemployment.

NOTATION

The Lithuanian LFS was carried out in September of 1996 selecting a simple random sample of persons belonging to the defind age category from the population register. All persons living together with the selected one and belonging to the surveyed age category are also included into the sample. Thus, the household which has more such individuals has a greater probability of being included into the sample.

There were 7258 individuals interviewed, living in 2686 households. A cluster sample with unequal selection probabilities was obtained. The nonresponse is ignored.

The poststratification of the sample in 13 age, 2 sex and 2 urban/rural groups, $13 \cdot 2 \cdot 2 = 52$ groups in total, is used. The other kind of poststratification is examined in order to "correct" the sample with respect to the registered unemployment (2 possibilities) in 6 age and 2 sex groups, $5 \cdot 2 \cdot 2 + 3 = 23$ groups in total (the eldest unemployed are not devided by sex).

Let us denote:

$L = 2897121$ – number of 14–74 year-old persons in the population register,

$M = 2802227$ – size of 14–74 year-old population according to the demographical data,

K – number of poststrata,

M_k – size of the k-th poststratum in the population known from the demographical data, $M_1 + M_2 + \cdots + M_K = M$,

N – number of households (clusters) in the population,

$n = 2686$ – number of households in the sample,

$m = 7258$ – number of persons in the sample,

m_i – number of persons in the i-th household, $m_1 + m_2 + \cdots + m_n = m$,

y_{ij} – value of the study variable y of for the j-th person in the i-th household, $j = 1, 2, \ldots, m_i$, $i = 1, 2, \ldots, n$ (for example, if y is a variable that determines the belonging of the person to the labour force, then y_{ij} equals 1 if the corresponding person from the domain under consideration belongs to the labour force and it equals 0, otherwise),

π_i – inclusion probability of the i-th household

$$\pi_i = \frac{nm_i}{L}, \quad i = 1, 2, \ldots, n,$$

π_{il} – joint inclusion into the sample probability of the i-th and l-th clusters,

$$\pi_{il} = \frac{nm_i}{L} \cdot \frac{(n-1)m_l}{L-1}, \quad i, l = 1, 2, \ldots, n,$$

w_i – design weight of the i-th household

$$w_i = \frac{1}{\pi_i} \cdot \frac{M}{L}, \quad i = 1, 2, \ldots, n,$$

$\widehat{M}_k$ – size estimator of the k-th poststratum,

$$\widehat{M}_k = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i \delta_{ij}(k), \quad k = 1, 2, \ldots, K,$$

$\delta_{ij}(k)$ equals 1, if the j-th person from the i-th household belongs to the k-th poststratum and 0, otherwise.

ESTIMATORS

The main parameters to be estimated in the LFS are labour force size and number of the unemployed in the population and its various domains. These parameters are totals. Let T be a population total of the variable y, $T = \sum_{i=1}^{M} y_i$. We need to find out the estimator for T. This parameter is estimated for four sampling strategies (e.g., sampling design and estimator): using Horvitz-Thompson (HT) estimator for cluster sampling, using poststratified estimator with respect to demographical data for cluster sampling, using poststratified estimator with respect to the registered unemployment for cluster sampling and assuming simple random sample (SI) of persons and HT estimator.

The estimators of totals for the cluster and simple random sample can be found, for example in ([1], [2]). The case of the cluster sample with unequal selection probabilities and poststratification is the most complicated one among those investigated because the clusters are destroyed by poststrata. The estimator for this case was not

found in the literature available for the authors, thus, it was derived by the authors themselves.

Let g_k be poststratification weights, $g_k = M_k/\widehat{M}_k$, $k = 1, 2, \ldots, K$, μ_k and $\hat{\mu}_k$ be mean and its estimator of the variable y in the k-th poststratum

$$\mu_k = \mu_{yk} = \frac{1}{M_k} \sum_{i=1}^{N} \sum_{j=1}^{m_i} \delta_{ij}(k) y_{ij},$$

$$\hat{\mu}_k = \hat{\mu}_{yk} = \frac{1}{\widehat{M}_k} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i \delta_{ij}(k) y_{ij}.$$

RESULT 1. *An unbiased poststratified estimator* $\widehat{T}^{\text{pos}}$ *of the population total* T *is*

$$\widehat{T}^{\text{pos}} = \widehat{T}_y^{\text{pos}} = \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i g_k \delta_{ij}(k) y_{ij}.$$

Its approximate variance and the variance estimator are

$$\mathbf{D}\widehat{T}^{\text{pos}} \approx \sum_{i=1}^{N} \sum_{l>1} \left[\sum_{k=1}^{K} (\pi_i \pi_j - \pi_{il}) \left(w_i \sum_{j=1}^{m_i} \delta_{ij}(k)(y_{ij} - \mu_k) - w_l \sum_{j=1}^{m_i} \delta_{lj}(k)(y_{lj} - \mu_k) \right) \right]^2$$

$$\widehat{\mathbf{D}}\widehat{T}^{\text{pos}} = \frac{1 - \frac{n}{L}}{n-1} \sum_{i=1}^{n} \sum_{l>1} \left[\sum_{k=1}^{K} g_k \left(\sum_{j=1}^{m_i} w_i \delta_{ij}(k)(y_{ij} - \hat{\mu}_k) - \sum_{j=1}^{m_l} w_l \delta_{lj}(k)(y_{lj} - \hat{\mu}_k) \right) \right]^2.$$

The other type of parameter to be estimated in the LFS is the ratio between totals of variables y and v: $R = T_y/T_v$. The unemployment rate may serve an example of estimator of the ratio.

RESULT 2. *A poststratified estimator of the ratio* R *is* $\widehat{R}^{\text{pos}} = \widehat{T}_y^{\text{pos}}/\widehat{T}_v^{\text{pos}}$. *Its approximate variance and the variance estimator are*

$$\begin{aligned} \mathbf{D}\widehat{R}^{\text{pos}} \approx{}& \frac{1}{T_v^2} \sum_{i=1}^{N} \sum_{l>1} (\pi_i \pi_l - \pi_{il}) \\ &\times \left[\sum_{k=1}^{K} \left(\omega_i \sum_{j=1}^{m_i} \delta_{ij}(k) \big(y_{ij} - \mu_{yk} - R(v_{ij} - \mu_{vk}) \big) \right.\right. \\ &\left.\left. - w_l \sum_{j=1}^{m_l} \delta_{lj}(k) \big(y_{lj} - \mu_{yk} - R(v_{lj} - \mu_{vk}) \big) \right) \right]^2, \end{aligned}$$

$$\widehat{\mathbf{D}}\widehat{R}^{\text{pos}} \approx \frac{1-n/L}{n-1}\frac{1}{(\widehat{T}_v^{\text{pos}})^2}\sum_{i=1}^{n}\sum_{l>1} \times \left[\sum_{k=1}^{K}\left(\omega_i\sum_{j=1}^{m_i}\delta_{ij}(k)\left(y_{ij}-\hat{\mu}_{yk}-\widehat{R}^{\text{pos}}(v_{ij}-\hat{\mu}_{vk})\right)\right.\right. \left.\left.-w_l\sum_{j=1}^{m_l}\delta_{lj}(k)\left(y_{lj}-\hat{\mu}_{yk}-\widehat{R}^{\text{pos}}(v_{lj}-\hat{\mu}_{vk})\right)\right)\right]^2.$$

The Taylor linearization of poststratified estimators is used for the calculation of their approximate variance, that is often met in the calculation of variance of the ratio estimator.

In practical calculations the quality of the sampling strategy is measured by the relative standard error $cv(\widehat{T}) = (\sqrt{\widehat{\mathbf{D}}_{\text{design}}(\widehat{T})}/\widehat{T})\cdot 100\%$, here $\widehat{T}$ is an estimator of the total, and $\widehat{\mathbf{D}}_{\text{design}}(\widehat{T})$ is its variance estimator. The quality of the sampling strategy with respect to the SI and HT estimator is measured by the design effect $\text{defff}(\widehat{T}) = \widehat{\mathbf{D}}_{\text{design}}(\widehat{T})/\widehat{\mathbf{D}}_{SI}(\widehat{T}^{HT})$. The measures of the quality of the ratio estimator are defined similarly.

Results of the LFS in September, 1996
Estimate
Relative standard error
Design effect

Strategy	Labour force	Number of unemployed	Unemployment rate
Cluster sampling	2037126 0.85% 1.40	303335 4.03% 1.42	0.1489 3.96% 1.42
Poststratified cluster sampling on demographical data (52 strata)	**2036685** **0.67%** **0.87**	316836 3.90% 1.45	0.1556 3.83% 1.45
Poststratified cluster sampling on registered unemployment (23 strata)	2023230 0.78% 1.14	**273788** **3.30%** **0.78**	**0.1353** **3.29%** **0.81**
SI	2022715 0.73% 1	306168 3.35% 1	0.1514 3.27% 1
Registered unemployment	1752600	112491	0.0642

This table shows the inefficiency of cluster sampling, well known from the theory ([1], [2]). The design effects in cluster sampling are greater than unit, it means that

the accuracy of estimates in pure cluster sampling is worse than in SI. But it does not remain the same when the poststratification is used. Poststratification by age, sex, urban/rural groups improved the labour force estimates substantially. Despite this, the accuracy of the estimates of unemployment remains insufficient. Poststratification on registered unemployment improved the unemployment estimates. The numbers on the registered unemployment are substantially smaller than the estimates obtained.

Figures 1–4 show the dependence of a relative standard error of the estimate on the size of the estimate (relative to the population size). The coefficient $cv = 5\%$ is assumed as the limit below which the accuracy of estimates is considered as sufficient. Small relative estimates are not precise enough.

Dependence of the relative standard error of all the estimates of totals on a relative estimate

Figure 2 in comparison to Figure 1 shows the effect of poststratification by demographical data on the accuracy of estimates. The limit $cv = 5\%$ level is reached under smaller estimates due to this poststratification.

Cluster sampling

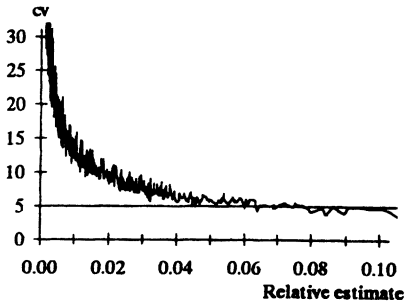

Figure 1

Poststratified cluster sampling by age, sex, urban/rural groups

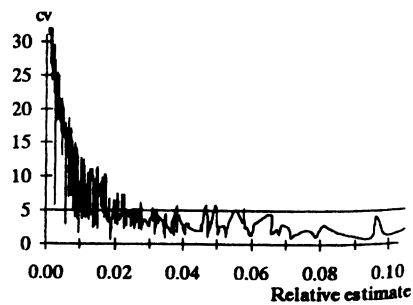

Figure 2

Dependence of the relative standard error of the number of unemployed on a relative estimate

The unemployed persons do not constitute a great part of population. Thus, many estimates of the number of unemployed are not accurate. Figure 3 in comparison to Figure 4 shows that the accuracy of the number of the unemployed is worse than the accuracy of estimates in general, using the poststratification by demographical data. But the poststratification by the registered unemployment makes them a little bit better. In the future the work has to be done in the direction of using more auxiliary information in the estimating stage as well as more modern ways of its accumulation in order to make LFS estimates more accurate.

Poststratified cluster sampling by age, sex, urban/rural groups

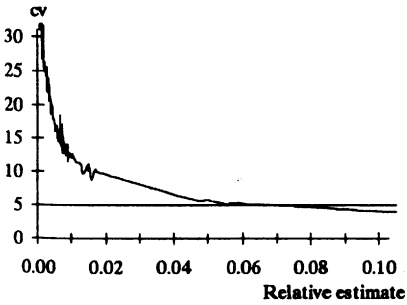

Figure 3

Poststratified cluster sampling by the registered unemployment

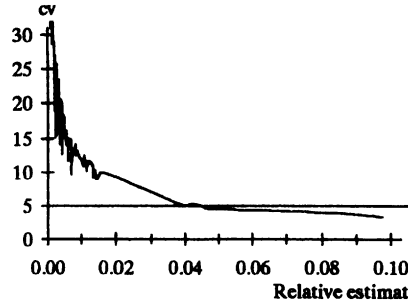

Figure 4

REFERENCES

[1] C.-E. Särndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, 1992.

[2] W. G. Cochran, *Sampling Techniques*, John Wiley & Sons, 1977.

[3] J. Hörngren, The use of registers as auxiliary information in the Swedish labour force survey. R&D Report, Statistics Sweden, 1992.

[4] K. Djerf, P. Väisänen, Effects of post-stratification on the estimates of the Finnish labour force survey. International Statistical Institute 49th session. Contributed papers, Bock 1. Firenze, ISI, (1993), 375–376.

Apie kai kuriuos įverčius, turint klasterinę imtį

G. Klimavičius, D. Krapavickaitė, A. Plikusas (MII)

Darbo tikslas yra ištirti kai kurių parametrų įverčių, kurie gali būti naudojami Lietuvos darbo jegos tyrime, tikslumą. Tam yra naudojami 1996 m. rugsėjo mėn. atlikto tyrimo duomenys. Iš gyventojų registro yra išrinkta namų ūkių klasterinė imtis su nevienodomis išrinkimo tikimybėmis. Po to nagrinėjama, kaip parametrų įverčių tikslumas priklauso nuo imties poststratifikacijos, gautos panaudojant demografinius duomenis ir valstybinės darbo biržos duomenis apie ten užsiregistravusius bedarbius. Pasirodo, kad imties poststratifikacija pagal amžiaus, lyties ir miesto/kaimo gyventojų grupes žymiai pagerina darbo jėgos, bet ne bedarbių skaičiaus įverčių tikslumą. Tuo tarpu imties poststratifikacija pagal registruotą bedarbystę patikslina bedarbių skaičiaus įverčius.