Variance Estimation Using Package vardpoor in R

Juris BREIDAKS (e-mail: juris.breidaks@csb.gov.lv)
Central Statistical Bureau of Latvia

ABSTRACT

The paper is devoted to the R package vardpoor. The Central Statistical Bureau of Latvia in 2012 developed R package vardpoor. The package vardpoor was developed with the objective to modernise the sample error estimation in sample surveys. Sampling errors can be estimated for household, agricultural and business surveys using the package. The main advantage of the proposed package is its simplicity and flexibility. R package vardpoor is implemented in practice. Sampling error estimation mechanism, calculation of the domain-specific study variables, variable linearization, calculation of regression residual, and variance estimation with the ultimate cluster method, variance estimation of the simple random sampling is briefly explained in the paper.

Keywords: *Survey sampling, variance estimation, domain estimation, linearization, R.*

JEL Classification: *C400; C420*

INTRODUCTION

The Central Statistical Bureau of Latvia (CSB) in 2012 developed R (R Core Team, 2014) package *vardpoor* (Breidaks, Liberts and Ivanova, 2015) (a set of functions for statistical calculation in programme R). The package *vardpoor* was developed with the objective to modernise the sample error estimation in sample surveys.

Before the package was developed, sampling errors were estimated using the chargeable programme SUDAAN (www.rti.org/sudaan). Use of SUDAAN had several shortcomings:

- Only obsolete SUDAAN version was available at CSB, which had to be updated;
- Updating of SUDAAN version would require financial resources;
- It is difficult to integrate SUDAAN into work with other data processing programmes (IBM SPSS Statistics or R);
- With the help of SUDAAN it was possible to linearize only non-linear statistics, as the ratio of two totals, but in the EU-SILC survey there were several other non-linear statistics, which had to be linearized separately;
- SUDAAN sampling error estimation did not include the effect of weight calibration.

Given the above shortcomings, it was decided to develop the *vardpoor* package, which would be designed as R extension. First of all, R is an open-source

free statistical calculation environment; secondly, R is currently the most popular computing environment among statisticians; and thirdly R environment is very convenient and suitable for development of such solutions. It should also be mentioned that, upon developing *vardpoor* package as R extension, it was easily integrated in the statistical production processes.

The theoretical basis of *vardpoor* was borrowed from Guillaume Osier article "The Linearisation approach implemented by Eurostat for the first wave of EU-SILC: what could be done from the second wave onwards?" (Osier, Meglio, 2012), which was presented at the workshop devoted to the evaluation of the standard errors and other issues related to the EU-SILC survey in March 2012.

SAMPLING ERROR ESTIMATION MECHANISM

Sampling error estimation mechanism consists of a sequence of procedures:

1. Calculation of the domain-specific study variables, if the sampling errors have to be estimated for population domains;
2. Variable linearization, if the sampling error has to be estimated for nonlinear type estimators: the ratio of two totals, at-risk-of-poverty rate, at-risk-of-poverty threshold, Gini coefficient, gender pay gap, median income of individuals below at-risk-of-poverty gap, income quintile share ratio, relative median at-risk-of-poverty gap (Osier, 2009);
3. Calculation of regression residual if the weights are calibrated;
4. Variance estimation with the ultimate cluster method (Hansen, Hurwitz and Madow, 1953);
5. Variance estimation under the assumption of the simple random sampling with the same number of respondents for design effect estimation.

Calculation of the domain-specific study variables

Often separate estimates for subpopulations are needed. Subpopulations are called domains. The domains concerned are denoted as $(U_1,...,U_d,...,U_D)$. It is assumed that y total value in each domain must be estimated. The aim is to estimate $(Y_1,...,Y_d,...,Y_D)$, where

$$Y_d = \sum_{k \in U_d} y_k, \ d = 1,...,D \tag{1}$$

The domain total can be expressed with a new variable y_{dk}, constructed from y specifically for domain U_d (Lundström, Särndal, 2001). The new variable is denoted with y_{dk} and its values for each element k are defined as

$$'_{dk} = \begin{cases} y_k, \ if \ k \in U_d, \\ 0, \quad if \ k \in U_d. \end{cases} \tag{2}$$

Then Y_d can be expressed as a total from the new variable y_{dk} for the whole population:

$$Y_d = \sum_{k \in U} y_{dk}. \tag{3}$$

Linearization approach

The linearisation method (Särndal, Swensson and Wretman, 1992; Deville, 1999; Wolter, 2007; Osier, 2009) uses Taylor-like series approximation to reduce non-linear statistics to a linear form, justified by asymptotic properties of the estimator (Verma and Betti, 2005). The method based on influence functions (Deville, 1999) is general enough to handle all the complex non-linear indicators of poverty and inequality based on EU-SILC such as the at-risk-of-poverty rate, the income quintile share ratio (S80/S20) and the Gini coefficient.

The estimated variance of the estimator $\hat{\theta}$ can be approximated by a linear function of the sample observations:

$$\hat{Var}\left(\hat{\theta}\right) \cong \hat{Var}\left(\sum_{k \in s} w_k \cdot \hat{u}_k\right) \tag{4}$$

where the value of the estimated linearized variable $\hat{u}_k$ is determined by calculating the following functional derivative:

$$\hat{u}_k = \lim_{t \to 0} \frac{T\left(\hat{M} + t\delta_k\right) - T\left(\hat{M}\right)}{t} \tag{5}$$

where the estimated population parameter $\hat{\theta}$ is expressed as a functional T of the measure $\hat{M}$, i.e.,

$$\hat{\theta} = T\left(\hat{M}\right) \tag{6}$$

and the measure $\hat{M}$ allocates the sample weight w_k to each unit k in the sample s:

$$\hat{M}\left(k\right) = \hat{M}_k = w_k, \quad k \in s \tag{7}$$

δ_k is the Dirac measure at k: for each unit k in the sample, $\delta_k\left(i\right) = 1$ if and only if $k=i$. The functional derivative (5) is called the influence function.

Variable linearization for the ratio of two totals

If sampling errors are to be assessed for the ratio of two totals, then the estimate of the ratio is calculated as

$$\hat{R} = \frac{\hat{Y}}{\hat{Z}} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k z_k} \tag{8}$$

where y_k is the value of variable y for element k, z_k is the value of variable z for element k, w_k is the sampling weight. Using Taylor linearization approach linearized variable for the ratio estimator is calculated as

$$\hat{u}_k = \frac{1}{\hat{Z}}\left(y_k - \hat{R}z_k\right) \tag{9}$$

Calculation of the Gini coefficient I in domain and its linearization

The values y_k are sorted and distinct. The estimate of the Gini coefficient I (Osier, 2012) in the domain D is calculated as

$$\hat{G}_D^I = 100 \cdot \left(\frac{\sum_{k \in s_D} w_k y_k \left(2\hat{N}_{D;k}^I - 1\right)}{\hat{N}_D \hat{Y}_D} - 1 \right) \tag{10}$$

where $\hat{N}_{D;k}^I$ are the cumulative sum of weights according to the y_k ordered non-decreasingly in the domain D of the sample s, i.e.

$$\hat{N}_{D;k}^I = \sum_{i \in s_D} w_i \cdot 1_{[i \le k]} \tag{11}$$

$$1_{[a]} = \begin{cases} 1, \textit{ if a is TRUE,} \\ 0, \textit{ if a is FALSE.} \end{cases} \tag{12}$$

$$\hat{Y}_D = \sum_{k \in s_D} w_k y_k, \tag{13}$$

$$\hat{N}_D = \sum_{k \in s_D} w_k, \tag{14}$$

The estimate of the Gini coefficient in domain D by Eurostat (2004, 2009) is calculated as

$$\hat{G}_D^{Eurostat} = 100 \cdot \left(\frac{\sum_{k \in s_D} w_k y_k \left(2\hat{N}_{D;k}^I - w_k\right)}{\hat{N}_D \hat{Y}_D} - 1 \right) \tag{15}$$

The estimated linearized variable of the Gini coefficient in the domain D is defined by Osier (2009):

$$\hat{u}_{D;k}^{GINI\_I} = 100 \cdot \frac{1_{[k \in D]}}{\hat{N}_D \hat{Y}_D} \cdot \left[2 \cdot \left(\hat{Y}_D - \hat{G}_{D;k} + w_k y_k + y_k \hat{N}_{D;k}^I \right) - y_k - \left(\hat{G}_D^I + 1 \right) \left(\hat{Y}_D + y_k \hat{N}_D \right) \right] \tag{16}$$

where $$\hat{G}_{D;k} = \sum_{i \in s_D} y_i w_i \cdot 1_{[i \le k]} \tag{17}$$

are weighted partial sums.

Calculation of the Gini coefficient II in domain and its linearization

The estimate of the Gini coefficient II (Langel and Tillé, 2012, Graf and Tillé, 2014) in the domain D is calculated as

$$\hat{G}_D^{II} = 100 \cdot \left(\frac{\sum_{k \in s_D} w_k y_k \left(2\hat{N}_{D;k}^{II} - w_k \right)}{\hat{N}_D \hat{Y}_D} - 1 \right) \tag{18}$$

where $\hat{N}_{D;k}^{II}$ are the cumulative sum of weights according to the y_k ordered non-decreasingly in the domain D of the sample s, i.e.

$$\hat{N}_{D;k}^{II} = \sum_{k \in s_D} w_i \cdot 1_{[y_i \le y_k]} \tag{19}$$

$\hat{Y}_D$ is estimated total in the domain D, and $\hat{N}_D$ is estimated size of the population in the domain D.

Langel and Tillé (2012) combine the various approaches to obtain the same estimated linearized variable of the Gini coefficient II for the sample in the domain D:

$$\hat{u}_{D;k}^{GINI\_II} = 100 \cdot \frac{1_{[k \in D]}}{\hat{N}_D \hat{Y}_D} \cdot \left[2\hat{N}_{D;k}^{II} \left(y_k - \hat{\bar{Y}}_{D;k} \right) + \hat{Y}_D - \hat{N}_D y_k - \hat{G}_D^{II} \left(\hat{Y}_D + \hat{N}_D y_k \right) \right] \tag{20}$$

where

$$\hat{\bar{Y}}_{D;k} = \frac{1}{\hat{N}_{D;i}} \sum_{i \in s_D} w_i y_i \cdot 1_{[i \le k]} \tag{21}$$

Weighted quantile estimation in the domain

Quantiles are defined as $Q_{D;p} = F_D^{-1}(p)$, where F_D is the income distribution function on the population in the domain D, i.e.,

$$F_{D;y}(x) = \frac{1}{N_D} \sum_{k \in U_D} 1_{[y_k \le x]} \tag{22}$$

and $0 \le p \le 1$. The median is given by $p=0.5$. For the following definitions, let n_D be the number of observations in the domain D of the sample, let $x_D := (x_1, ..., x_{n_D})$, denote the equalized disposable income with $x_1 \le ... \le x_{n_D}$, and let $w_D := (w_1, ..., w_{n_D})$ be the corresponding personal sample weights. Weighted quantiles for the estimation of the population values in the domain D according are then given (Alfons, Templ, 2014) by

$$\hat{Q}_{D;p} = \hat{Q}_{D;p}(x_D; w_D) := \begin{cases} \frac{1}{2}(x_j + x_{j+1}), & if \sum_{i=1}^{j} w_i = p \sum_{i=1}^{n_D} w_i \\ x_{j+1}, & if \sum_{i=1}^{j} w_i < p \sum_{i=1}^{n_D} w_i < \sum_{i=1}^{j+1} w_i \end{cases} \tag{23}$$

***Calculation of the* at-risk-of-poverty threshold in domain and its linearization**

The at-risk-of-poverty threshold (ARPT) in the domain *D* is defined as 60% of the median income in the domain *D*:

$$ARPT_D = 0.6 \cdot F_D^{-1}(0.5), \tag{24}$$

$$\hat{ARPT}_D = 0.6 \cdot \hat{Q}_{D;0.5}(0.5). \tag{25}$$

The linearized variable of the ARPT in the domain *D* is defined by Osier (2009):

$$\hat{u}_{D;k}^{ARPT} = I(ARPT_D)_k = 0.6 \cdot I\left(\hat{Q}_{D;0.5}\right)_k = -\frac{0.6}{f\left(\hat{Q}_{D;0.5}\right)} \cdot \frac{1_{[k \in D]}}{\hat{N}_D}\left[1_{[y_k \le \hat{Q}_{D;0.5}]} - 0.5\right], \tag{26}$$

where y_i is i-th equalized disposable income, $\hat{N}_D$ is estimated size of the population in the domain *D*. Deville (2000) and Osier (2009) suggest using Gaussian kernel estimation

$$K(o) = -\frac{1}{h_D\sqrt{2\pi}} e^{-\frac{o^2}{2}}, \tag{27}$$

to estimate the income distribution function $F_d(x)$

$$f_D(x) = \frac{1}{h_D\sqrt{2\pi}} \cdot \frac{1_{[k \in D]}}{\hat{N}_D} \sum_{i \in s_D} w_i K\left(\frac{x - y_i}{h_D}\right) = \frac{1}{h_D\sqrt{2\pi}} \cdot \frac{1_{[k \in D]}}{\hat{N}_D} \sum_{i \in s_D} w_i \exp\left[-\frac{(x - y_i)^2}{2h_D^2}\right], \tag{28}$$

where h_D $\boldsymbol{h_D}$ is the bandwidth in the domain *D*, which is estimated using the "plug-in" estimator (Silverman, 1986):

$$\hat{h}_D = \hat{\sigma}_D \hat{N}_D^{-0.2} \tag{29}$$

and $\hat{\sigma}_D$ is the estimated standard deviation of the empirical income distribution:

$$\hat{\sigma}_D = \frac{1}{\hat{N}_D}\sqrt{\hat{N}_D \sum_{k \in s_D} w_k y_k^2 - \left(\sum_{k \in s_D} w_k y_k\right)^2}, \tag{30}$$

***Calculation of the* at-risk-of-poverty rate in domain and its linearization**

The at-risk-of-poverty rate in domain *D* is the share of persons in domain *D* with an income below the at-risk-of-poverty threshold ARPT:

$$ARPR_D = F_D(ARPT). \tag{31}$$

The estimated at-risk-of-poverty rate (ARPR) in the domain (Osier, 2009) for a sample is

$$ARPR_D = \frac{\sum_{y_k < \hat{ARPT}} w_k}{\hat{N}_D}. \tag{32}$$

The linearized variable of the ARPR in the domain D is defined by Osier (2009):

$$\hat{u}_{D;k}^{ARPR} = \frac{1_{[k\in D]}}{\hat{N}_D}\left(1_{[y_k \le A\hat{R}PT]} - A\hat{R}PR\right) - \frac{f\left(A\hat{R}PT_D\right)}{f\left(\hat{Q}_{ALL;0.5}\right)} \cdot \frac{0.6}{\hat{N}_D}\left[1_{[y_k \le \hat{Q}_{ALL;0.5}]} - 0.5\right], \tag{33}$$

$$\hat{u}_{D;k}^{ARPR} = \frac{1_{[k\in D]}}{\hat{N}_D}\left(1_{[y_k \le A\hat{R}PT]} - A\hat{R}PR\right) - f\left(A\hat{R}PT_D\right) \cdot \hat{u}_{ALL;k}^{ARPT}, \tag{34}$$

where $\hat{Q}_{ALL;0.5}$ is the estimated median income of the sample.

***Calculation of the median income of individuals below the* ARPT in domain and its linearization**

The indicator is defined as the difference between the ARPT and the median income of the "poor" people, taken relatively to the ARPT: The median income of individuals below the ARPT (MED^p) in the domain D is estimated

$$M\hat{E}D_D^p = F_D^{-1}\left(\frac{1}{2}F_D\left(A\hat{R}PT\right)\right) = F_D^{-1}\left(\frac{1}{2}A\hat{R}PR_D\right) \tag{35}$$

where F_D is the cumulative income distribution function.

The linearized variable of $M\hat{E}D_D^p$ is defined by Osier (2009):

$$\hat{u}_{D;k}^{MED^p} = \frac{1}{f\left(M\hat{E}D_D^p\right)} \cdot \left(\frac{\hat{u}_{D;k}^{ARPR}}{2} - \frac{1_{[k\in D]}}{\hat{N}_D}\left(1_{[y_k \le M\hat{E}D_D^p]} - F\left(M\hat{E}D_D^p\right)\right)\right), \tag{36}$$

$$\hat{u}_{D;k}^{MED^p} = \frac{1}{f\left(M\hat{E}D_D^p\right)} \cdot \left(\frac{\hat{u}_{D;k}^{ARPR}}{2} - \frac{1_{[k\in D]}}{\hat{N}_D}\left(1_{[y_k \le M\hat{E}D_D^p]} - \frac{1}{2}A\hat{R}PR_D\right)\right), \tag{37}$$

where $f(\cdot)$ is influence function (25).

***Calculation of the relative median poverty gap* in domain and its linearization**

The relative median at-risk-of-poverty gap (RMPG) is the relative difference between the ARPT and the median income of individuals below the ARPT. RMPG equals to 0 if the income of all "poor" individuals is equal to the ARPT, and RMPG equals to 1 if the income of all these individuals is zero (Graf and Tillé, 2014). The RMPG in domain D is estimated:

$$Q\hat{S}R_D = \frac{\sum_{y_k \ge \hat{Q}_{D;0.8}} w_k \cdot y_k \cdot 1_{[k\in D]}}{\sum_{y_k \le \hat{Q}_{D;0.2}} w_k \cdot y_k \cdot 1_{[k\in D]}}, \tag{38}$$

The linearized variable of the RMPG in the domain D is defined by Osier (2009):

$$\hat{u}_{D;k}^{RMPG} = \frac{\hat{MED}_D^p \cdot \hat{u}_{D;k}^{ARPT} - \hat{ARPT}_D \cdot \hat{u}_{D;k}^{MED^p}}{\left(\hat{ARPT}_D\right)^2}, \tag{39}$$

Calculation of the Quantile Share Ratio in domain and its linearization

The Quantile Share Ratio (QSR or S80 / S20) is the ratio of the total income of the 20% of the population with the highest income to the total income of the 20% of the population with the lowest income. The RMPG in domain D by G. Osier (2009) is estimated:

$$\hat{QSR}_D = \frac{\sum_{y_k \geq \hat{Q}_{D;0.8}} w_k \cdot y_k \cdot 1_{[k \in D]}}{\sum_{y_k \leq \hat{Q}_{D;0.2}} w_k \cdot y_k \cdot 1_{[k \in D]}}, \tag{40}$$

The estimate of the QSR in domain *D* by Eurostat (2004, 2009) is calculated as

$$\hat{QSR}_D^{Eurostat} = \frac{\frac{\sum_{y_k \geq \hat{Q}_{D;0.8}} w_k \cdot y_k \cdot 1_{[k \in D]}}{\sum_{y_k \geq \hat{Q}_{D;0.8}} w_k \cdot 1_{[k \in D]}}}{\frac{\sum_{y_k \leq \hat{Q}_{D;0.2}} w_k \cdot y_k \cdot 1_{[k \in D]}}{\sum_{y_k \leq \hat{Q}_{D;0.2}} w_k \cdot 1_{[k \in D]}}}, \tag{41}$$

The estimated linearized variable of the QSR in the domain *D* is defined by Osier (2009):

$$\hat{u}_{D;k}^{QSR} = \frac{\left(\sum_{y_k \leq \hat{Q}_{D;0.2}} w_k \cdot y_k \cdot 1_{[k \in D]}\right) \cdot lin_{top} - \left(\sum_{y_k \geq \hat{Q}_{D;0.8}} w_k \cdot y_k \cdot 1_{[k \in D]}\right) \cdot lin_{bottom}}{\left(\sum_{y_k \leq \hat{Q}_{D;0.2}} w_k \cdot y_k \cdot 1_{[k \in D]}\right)^2}, \tag{42}$$

where

$$lin_{top} = y_k - y_k \cdot 1_{[y_k \leq \hat{Q}_{D;0.8}]} + \frac{s\left(\hat{Q}_{D;0.8}\right)}{f\left(\hat{Q}_{D;0.8}\right) \cdot \hat{N}_D} \cdot \left(1_{[y_k \leq \hat{Q}_{D;0.8}]} - 0.8\right), \tag{43}$$

$$lin_{bottom} = y_k \cdot 1_{[y_k \leq \hat{Q}_{D;0.2}]} + \frac{s\left(\hat{Q}_{D;0.2}\right)}{f\left(\hat{Q}_{D;0.2}\right) \cdot \hat{N}_D} \cdot \left(1_{[y_k \leq \hat{Q}_{D;0.2}]} - 0.2\right), \tag{44}$$

Osier (2009) suggest using Gaussian kernel estimation to estimate the function *s(x)*

$$K(o) = -\frac{1}{h_D \sqrt{2\pi}} e^{-\frac{o^2}{2}}, \tag{45}$$

$$s(x) = \frac{1}{h_D \sqrt{2\pi}} \cdot \frac{1_{[k \in D]}}{\hat{N}_D} \sum_{i \in s_D} y_i w_i K\left(\frac{x - y_i}{h_D}\right) = \frac{1}{h_D \sqrt{2\pi}} \cdot \frac{1_{[k \in D]}}{\hat{N}_D} \sum_{i \in s_D} y_i w_i \exp\left[-\frac{(x - y_i)^2}{2h_D^2}\right], \tag{46}$$

Regression residual calculation

If the weights are calibrated, then calibration residual estimates $\hat{e}_k$ are calculated (Lundström, Särndal, 2001) by formula

$$\hat{e}_k = y_k - x_k^{'}\hat{B}, \tag{47}$$

where

$$\hat{B} = \left(\sum_{k\in s} d_k q_k x_k x_k^{'}\right)^{-1}\left(\sum_{k\in s} d_k q_k x_k y_k\right). \tag{48}$$

Variance estimation with the ultimate cluster method

If we assume that $n_h \geq 2$ for all h, that is, two or several primary sampling units (PSUs) are sampled from each stratum, then variance of $\hat{\theta}$ can be estimated from the variation among the estimated PSU totals of y (Hansen, Hurwitz, Madow, 1953; Osier, Meglio, 2012; Berger, Goedemé, Osier, 2013):

$$\hat{V}\left(\hat{\theta}\right) = \sum_{h=1}^{H}\left(1-f_h\right)\frac{n_h}{n_h-1}\sum_{k=1}^{n_h}\left(y_{hk\bullet}-\bar{y}_{h\bullet\bullet}\right)^2, \tag{49}$$

where

- $y_{hk\bullet} = \sum_{j=1}^{m_{hk}} w_{hkj} y_{hkj}$,
- $\bar{y}_{h\bullet\bullet} = \frac{\sum_{j=1}^{m_{hk}} y_{hk\bullet}}{n_h}$, ,
- f_h is a sampling fraction of PSUs for stratum h,
- h is the stratum number, with a total of H strata,
- k is the number of PSU within the sample of stratum h, with a total of n_h PSUs,
- j is the household number within PSU k of stratum h, with a total of m_{hi} households,
- w_{hkj} is the sampling weight for household j in PSU k of stratum h,
- y_{hkj} denotes the observed value of study variable y for household j in PSU k of stratum h.

Variance estimation of the simple random sampling

The variance $V_{SRS}\left(\hat{\theta}_{SRS}\right)$ under assumption of the simple random sampling is estimated by formula (Ardilly and Tillé, 2005; Ardilly and Osier, 2007):

$$\hat{V}_{SRS}\left(\hat{\theta}_{SRS}\right) = N^2\frac{1-f}{n}\frac{\sum_{k\in s} w_k\left(y_k-\bar{y}_w\right)^2}{\sum_{k\in s} w_k - 1}, \tag{50}$$

where

- k is the number of unit in the sample s, with a total of n units,
- y_k denotes the observed value of study variable y for unit k,
- w_k is the sampling weight for unit i,
- $N = \sum_{k \in s} w_k$,
- $f = \frac{n}{N}$ is the sampling fraction, $1-f$ the finite population correction factor,
- $\bar{y}_k = \frac{\sum_{k \in s} y_k w_k}{\sum_{k \in s} w_k}$ is the weighted sample mean of y.

***The design effect estimation and* effective sample size**

The design effect of sampling is estimated by

$$D\hat{e}ff_{sam}\left(\hat{\theta}\right) = \frac{V\hat{a}r_{CUR,HT}\left(\hat{\theta}\right)}{V\hat{a}r_{SRS,HT}\left(\hat{\theta}\right)}, \tag{51}$$

where $V\hat{a}r_{SRS,HT}\left(\hat{\theta}\right)$ is the variance of HT estimator under SRS, $V\hat{a}r_{CUR,HT}\left(\hat{\theta}\right)$ is the variance of HT estimator under current sampling design.

The design effect of estimator is estimated by

$$\hat{e}ff_{est}\left(\hat{\theta}\right) = \frac{V\hat{a}r_{CUR,CAL}\left(\hat{\theta}\right)}{V\hat{a}r_{CUR,HT}\left(\hat{\theta}\right)}, \tag{52}$$

where $V\hat{a}r_{CUR,HT}\left(\hat{\theta}\right)$ is the variance of calibrated estimator under current sampling design.

The overall design effect of sampling and estimator is estimated by

$$D\hat{e}ff\left(\hat{\theta}\right) = D\hat{e}ff_{samp}\left(\hat{\theta}\right) \cdot \hat{e}ff_{est}\left(\hat{\theta}\right), \tag{53}$$

The effective sample size is estimated by

$$\hat{n}_{eff}\left(\hat{\theta}\right) = \frac{n}{D\hat{e}ff\left(\hat{\theta}\right)}, \tag{54}$$

where n is the sample size or the number of respondents (in case of non-response).

***Function** vardom* **description**

Function *vardom* is used to estimate sampling errors for total or the ratio of two totals. At the beginning of the function execution a range of tests is performed in order to test if there are any mistakes in data. If Dom is entered, the variable concerned is broken down by domain by formula (2). If Z is entered, the ratio of two totals is calculated. If the calculation is performed by domain, also Z is broken down by domain. By means of the function the matrix of linearized values is calculated. If calibration matrix X and g weights are used, function calculates the residuals.

Function *vardom* outputs several results:

- count of the respondent,
- the count of the respondents with non-zero values,
- the population size,
- point estimates for statistics,
- variance estimates,
- relative standard error,
- absolute margin of error,
- relative margin of error,
- lower and upper bound of the confidence interval at α significance level,
- variance of HT estimator under current design,
- variance of calibrated estimator under SRS,
- the sample design effect,
- the estimated design effect of estimator,
- the overall design effect of sample design and estimator,
- the effective sample size.

vardom **function testing results**

Function was tested on Latvia data of survey for average purchase prices of round timber in the 1st half of 2014. To estimate the quality indicators for this survey, the function *vardom*() is available

```
> variables <- c("Pine_sawlogs_up_to_14_cm_in_diameter", "Pine_
sawlogs_14-18_cm_in_diameter", "Pine_sawlogs_18-26_cm_in_diameter",
"Pine_sawlogs_more_than_26_cm_in_diameter")
>rez <- vardom(Y=variables, H="strata", PSU="resp_code", w_final="weights",
N_h=NULL, dataset=woodstock2014h1)$all_results
```

Quality measure for survey for average purchase prices of round timber using function *vardom*

Table1

variable	estim	se	cv	CI_lower	CI_upper	deff_sam	deff_est	deff
Pine_sawlogs_up_to_14_cm_in_diameter	50.16	1.33	2.66	47.55	52.78	0.153	1	0.153
Pine_sawlogs_14-18_cm_in_diameter	69.68	0.19	0.27	69.31	70.04	0.005	1	0.005
Pine_sawlogs_18-26_cm_in_diameter	73.60	0.22	0.30	73.17	74.04	0.027	1	0.027
Pine_sawlogs_more_than_26_cm_in_diameter	78.48	0.33	0.42	77.84	79.11	0.082	1	0.082

In Table 1 estimates are made as ratio type estimation, estimated standard error estimation and coefficient of variance is shown, that our data quality is well in real Latvia data.

***Function** vardomh* **description**

Function ***vardomh*** is used to estimate sampling errors for total or the ratio of two totals, when data is given at the person level, but information for the calibration is given at the household level. At the beginning of the function execution a range of tests is performed in order to test if there are any mistakes in data. If Dom is entered, the variable concerned is broken down by domain at the person level. If *Z* is entered, the ratio of two totals is calculated. If the calculation is performed by domain, also Z is broken down by domain at the person level. By means of the function *lin.ratio* the matrix of linearized values at the person level is calculated. If calibration matrix X and g weights are used, function calculates the residuals at the household level.

Function *vardomh* outputs the same output as function *vardom*.

vardomh **function testing results**

Function was tested on Latvia data of EU – SILC 2014. Results are given at individual and household level. Data were prepared by SAS programme. Ratio is AROPE (poverty or social exclusion indicator), DEP (the severe material deprivation rates), LWI (the share of individuals aged less than 60 living in households with very low work intensity) and POV (at-risk-of-poverty rate). The standard error is the square root of the variance.

To estimate the quality of POV, AROPE, LWI, DEP in EU-SILC survey, the function *vardomh*() is used:

```
>N_h = data.frame(db050 = 1:4, pop = c(1079, 719, 684, 1412))
>rez <- vardomh(Y=c("POV", "AROPE", "DEP", "LWI"), H="db050",
PSU="db060",
               w_final="rb050a", ID_household = "db030", Z ="zz", dataset
= dataset2,
               X = Xm, X_ID_household = X_ID_household, ind_gr = ind_gr,
g = g, q=q)
```

Quality for EU-SILC in Latvia for 2014

Table 2

variable	estim	se	cv	CI_lower	CI_upper	deff_sam	deff_est	deff
POV	19.38	0.489	2.52	18.42	20.34	0.95	0.812	0.773
AROPE	35.14	0.585	1.66	33.99	36.28	0.90	0.815	0.734
DEP	24.01	0.561	2.34	22.91	25.11	0.97	0.854	0.829

variable	estim	se	cv	CI_lower	CI_upper	deff_sam	deff_est	deff
LWI	9.92	0.393	3.96	9.15	10.69	0.78	0.808	0.629

In function *vardomh* example is shown, that is used calibration matrix Xm and g weights for each independent group separately using ind_gr. In table 2 has calculated standard errors, coefficient of the variance, confidence interval and design effect.

***Function** varpoord* **description**

Function *varpoord* is used to estimate sampling errors for indicators on social exclusion and poverty. Data is given at the person level, but information for the calibration is given at the household level. At the beginning of the function execution a range of tests is performed in order to test if there are any mistakes in data. Function *varpoord* consist argument type, if it is chosen

- **linarpt**, then calculate the at-risk-of-poverty threshold (ARPT) in the domain and linearized values in the domain *D* by G. Osier (2009) methodology;
- **linarpr**, then calculate the at-risk-of-poverty rate (ARPR) in the domain and linearized values in the domain *D* by G. Osier (2009) methodology;
- **linpoormed**, then calculate the median income of individuals below the ARPT in the domain and linearized values in the domain *D* by G. Osier (2009) methodology;
- **linrmpg**, then calculate the relative median poverty gap (RMPG) in the domain by G. Osier (2009) methodology and linearized values in the domain *D* by G. Osier methodology;
- **lingini**, then calculate the Gini coefficient in domain by G. Osier (2009) and Eurostat methodology and linearized values of Gini coefficient in the domain by G. Osier methodology;
- **lingini2**, then calculate the Gini coefficient in domain (Langel and Tillé, 2012, Graf and Tillé, 2014) and Eurostat methodology and linearized values of Gini coefficient in domain by Langel and Tillé methodology;
- **linqsr**, then calculate the quantile share ratio (QSR) in domain by G. Osier (2009) and Eurostat methodology and linearized values of QSR in the domain by G. Osier methodology;
- **all_choices**, then calculate all previously descripted choices.

If calibration matrix X and g weights are used at household level, function calculates the residuals at the household level.

Function *varpoord* outputs several results:

- point estimates for statistics,
- variance estimates,
- relative standard error,
- absolute margin of error,
- relative margin of error,
- lower and upper bound of the confidence interval at α significance level,
- variance of HT estimator under current design,
- variance of calibrated estimator under SRS,
- the sample design effect, the estimated design effect of estimator,
- the overall design effect of sample design and estimator,
- the effective sample size.

***varpoord* function testing results**

Function was tested on Latvia data of EU – SILC 2012. In this function will test ARPT and ARPR quality indicators. To estimate the quality of POV, AROPE, LWI, DEP in EU-SILC survey, the function *varpoord*() is used:

```
>N_h = data.frame(db050 = 1:4, pop = c(1079, 719, 684, 1412))
alw <-  varpoord(inc="INC_ekv20", w_final="db090", income_thres = "INC_ekv20",
                 wght_thres = "db090", ID_household = "db030n", H="db050",
                 PSU="db060", N_h=N_h, sort = NULL, dataset =  dataset2, X = Xm,
                 X_ID_household = X_ID_household, ind_gr = ind_gr, g = g,
                 q = q, several.ok=T, type=c("linarpt", "linarpr"))
```

EU-SILC quality indicators and its quality in 2012

Table 3

variable	value	se	cv	CI_lower	CI_upper	deff_sam	deff_est	deff
ARPT	1876.67	21.45	1.143	1834.64	1918.71	0.997	1.12	1.12
ARPR	19.21	0.50	2.619	18.22	20.19	0.880	0.86	0.76

In function *varpoord* example is shown, that is used calibration matrix *X* and *g* weights for each independent group separately using *ind_gr*. In table 3 has calculated standard errors, coefficient of the variance, confidence interval and design effect.

CONCLUSION

R package *vardpoor is* implemented in practice. Sampling errors can be estimated for household, agricultural and business surveys using the package. Functions for sampling error estimation for extra cross-sectional and longitudinal measures and the measures of net changes have added to the package and descripted in paper (Breidaks, Veretjanovs, Ivanova, 2015) and it will test more on real data.

REFERENCES

1. ALFONS, A., TEMPL, M., (2014). Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken, R package version 0.4.6, http://cran.r-project.org/web/packages/laeken/.
2. BERGER, Y. G., OSIER, G., GOEDEMÉ, T., (2013), Net-SILC2 Handbook on Standard error estimation and other related sampling issues (ver. 29/07/2013), Second Network for the Analysis of EU-SILC. http://www.cros-portal.eu/content/handbook-standard-error-estimation-and-other-related-sampling-issues-ver-29072013.
3. BREIDAKS, J., Precision estimation using ultimate cluster method in software R, Workshop of BNU Network in survey statistics, August 25-28, 2014, Tallinn, Estonia, ISBN 978-9985-74-563-2, http://www.stat.ee/74630
4. BREIDAKS, J., LIBERTS, M., IVANOVA, S., (2015). vardpoor: Variance Estimation for Sample Surveys by the Ultimate Cluster, R package version 0.2.8, http://cran.r-project.org/web/packages/vardpoor/.
5. BREIDAKS, J., VERETJANOVS, V., IVANOVA, S., (2015). The variance estimation for the measures of change for multistage cluster sampling designs using software R, Proceedings of the 14th APLIMAT 2015 Conference (APLIMAT 2015), Bratislava (Slovakia), pp. 115-133.
6. COCHRAN, W. G. (1977). Sampling techniques. New York: John Wiley and Sons.
7. DEVILLE J. C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. Survey Methodology, 25, 193-203, http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?lang=eng&catno=12-001-X19990024882
8. DI MEGLIO, E., OSIER, G., GOEDEMÉ, T., BERGER, Y. G., and DI FALCO, E., (2013). Standard Error Estimation in EU-SILC – First Results of the Net-SILC2 Project, NTTS (New Techniques and Technologies for Statistics), http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_144.pdf.
9. EUROSTAT, (2004). Common cross-sectional EU indicators based on EU-SILC; the gender pay gap. EU-SILC 131-rev/04, Unit D-2: Living conditions and social protection, Directorate D: Single Market, Employment and Social statistics, Luxembourg.
10. EUROSTAT, (2009). Algorithms to compute social inclusion indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC). Doc. LC-ILC/39/09/EN-rev.1, Unit F-3: Living conditions and social protection, Directorate F: Social and information society statistics, Luxembourg.
11. GRAF E., TILLÉ, Y. (2014). Variance Estimation Using Linearization for Poverty and Social Exclusion Indicators, Survey Methodology 61 Vol. 40, No. 1, pp. 61-79, Statistics Canada, Catalogue no. 12-001-X, http://www.statcan.gc.ca/pub/12-001-x/12-001-x2014001-eng.pdf
12. HANSEN, M. H., HURWITZ, W. N., MADOW, W. G., (1953), Sample survey methods and theory Volume I Methods and applications, 257-258, Wiley.
13. LANGEL, M., and TILLÉ, Y. (2012). Variance estimation of the Gini index: Revisiting a result several times published. In Press in Journal of the Royal Statistical Society - Series A.
14. LUNDSTÖM, S., SÄRNDAL, C. E., (2001). Estimation in the presence of Nonresponse and Frame Imperfections. Statistics Sweden, 43-44.
15. SÄRNDAL, C. E., SWENSSON, B., WRETMAN, J., (1992). Model Assissted Survey Sampling, 176-181, Springer-Verlag.
16. OSIER, G., (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. Survey Research Methods, Vol. 3, No. 3, pp. 167-195. Available at: http://w4.ub.uni-konstanz.de/srm/article/view/369
17. OSIER G., MEGLIO E. Di (29-30 March 2012). The linearisation approach implemented by Eurostat for the first wave of EU-SILC: what could be done from second wave onwards? Net-SILC2 workshop on standard error estimation and other related sampling issues, EUROSTAT.
18. R CORE TEAM (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/