STATISTICAL OFFICE OF ESTONIA

VARIANCE OF LAEKEN INDICATORS IN COMPLEX SURVEYS

Elsa Leiten and Imbi Traat

Statistical Office of Estonia

PREFACE

On October 28, 2004, the ICON Institute Public Sector GmbH and the Statistical Office of Estonia (SOE) have signed a Contract No 1138 — ILC-SerCon-Estonia-1 „Income and Living Conditions Statistics“. The Statistical Office of Estonia as a contractor has to carry out studies on three specified fields (three outputs) and provide results to Eurostat in Final Reports.

The current study concerns Output 2:

Variance coefficients in complex statistics like Gini coefficient, at-risk-of-poverty rate, etc.:

theoretical study (resampling and linearization methods, etc.), existing results;

programs in SAS to calculate variance coefficients.

Two persons, Elsa Leiten and Imbi Traat have taken responsibility on the studies and results on Output 2. In the subcontracts with the Statistical Office of Estonia they have agreed in the amount and terms of the work (the Intermediate Report March 31, the Final Report August 25). Basically, Elsa Leiten focuses on the Gini coefficient and Imbi Traat on the at-risk-of-poverty rate and associated indicators like median, poverty threshold and other quantiles. The main concern is variability of the listed indicators.

Elsa Leiten is a methodologist-mathematician at the Methodology Department of the Statistical Office of Estonia and a master student at the Institute of Mathematical Statistics of the University of Tartu under supervision of Imbi Traat.

Imbi Traat is Docent at the Institute of Mathematical Statistics of the University of Tartu with research interests in survey sampling theory and methodology.

CONTENTS

1. INTRODUCTION

The Gini coefficient and the at-risk-of-poverty rate are the items among many other similar parameters, commonly called Laeken indicators (see Eurostat [1]). Therefore we have specified the field of our research more generally as concerning Laeken indicators.

The formulae for calculating Laeken indicators are well known, also available for complex sampling designs. Of course, for each specified sampling design the correct sampling weights need to be used.

The difficulty is in estimating variance of sample-calculated Laeken indicators. Standard statistical assumptions of independent and identically distributed observations are not appropriate. Complex survey design violates these assumptions. In addition the non-linear feature of estimators makes the variance calculations difficult. Furthermore, the existing in the sampling literature general variance formulae are developed for without-replacement (WOR) sampling designs. The EU-SILC in Estonia is actually with-replacement (WR) unequal probability design for households (household may enter through all of its members), and therefore the available basic results need to be redeveloped in this Report.

Many of the Laeken indicators can be seen as certain domain proportions estimated by the ratio type estimator. The difficult point is that the domain indicator in sample is constructed with random threshold. For example, if the domain is people below the median of some variable, and median is estimated from sample, then the estimated domain proportion may display additional variation (or its variation may be smaller), compared to the case with fixed threshold. For this case the variance estimator is not readily available in the literature. Study of this case is one of the aims of the project.

If the threshold is fixed then the linearization-based variance estimator is appropriate for the ratio type estimator. The general formulae for WOR-designs are available in Särndal et al. (1992, p. 176–181). With design-vector approach (e.g. Traat et al., 2004) we present these formulae more generally, valid both for WOR- and WR-designs. From these, more general, formulae we develop variance formulae for the EU-SILC design in Estonia. Under EU-SILC, sampling of households (hh) is probabililistically (closely) described by the multivariate hypergeometric distribution. Appropriate sampling design of hh's is called hypergeometric design (see e.g. Traat, et al., 2004; Ilves, 2005). The derived variance formulae for the fixed threshold case are slightly modified – the sample-calculated threshold is inserted. The performance of these formulae is studied in the simulation experiment.

The variance formulae of the median and other quantiles under WOR-designs can be also found in Särndal et al. (1992, p.197–205). They use inverse distribution function method. Since distribution function at certain point is in fact a domain proportion, estimated by a ratio, we can use here already developed by us framework for this case. Inversion of distribution function takes place analogically as in Särndal et al. (1992). We develop variance formulae for quantiles (including median and poverty threshold) for EU-SILC design.

The Gini coefficient can be considered as a function of the distribution function, however this function is quite complex. In this Report the variability of the estimated Gini coefficient is studied under EU-SILC design. The Jackknife variance estimator extended from the one known for sample sums is considered. Its performance is analysed in a simulation study.

In this Report we first present the estimation formulae for Laeken indicators and then concentrate to their variance estimation. We develop the variance formulae of Laeken indicators, such as median, poverty threshold, at-risk-of-poverty rate and related quantities, for EU-SILC design of Estonia. The performance of variance estimators is studied in a simulation experiment. The real Estonian data are used to mimic the population. The EU-SILC design of Estonia is applied for repeated sampling.

2. SAMPLING DESIGN OF THE EU-SILC IN ESTONIA

The primary target of the study is to derive variance formulae of Laeken indicators for the EU-SILC design of Estonia.

EU-SILC (EU Survey on Income and Living Conditions) is a sample survey expected to obtain comparative statistics on income distribution, living conditions and social exclusion at European level. The survey data is based on nationally representative probability samples to permit detailed analysis by population subgroups.

The member states have some flexibility in the design of EU-SILC. Data is required in both cross-sectional (pertaining to a given time in a certain time period) and longitudinal (pertaining to individual-level changes over time) dimensions. Therefore certain households will be surveyed on an annual basis.

Required sampling precision is specified by Eurostat and it is an essential component of the data quality standards in EU-SILC. Draft EU-SILC Regulation stipulates the minimum effective sample size in country level both for cross-sectional and longitudinal parts of the surveys. The minimum sample size requirements are specified taking into account the diversity of data sources and survey designs. The actual sample sizes will have to be larger to the extent that the design effects exceed 1.0 and to be able to compensate for all kinds of non-response. The actual sample sizes are to be calculated by the national statistical offices.

The minimum effective sample sizes fixed for Estonia are:

- 3,500 households and 7,750 persons aged 16 and over in the cross-sectional survey;
- 2,750 households and 5,750 persons aged 16 and over in the longitudinal survey.

In this research we focus on the cross-sectional part of the survey.

The survey design is stratified unequal probability sampling of households. The design is basically the same that is used for the Estonian Household Budget Survey (see Statistical Office of Estonia, 2003). Sampling is carried through among the records of population register, whereas the sampling frame consists of people 14 years old and older (14+).

Strata are formed geographically by grouping Estonian counties (and the capital city Tallinn) into three strata by the population size. Hiiu County forms a separate stratum as the smallest county with the population size times smaller of the next smallest. The rest of the regions are divided into two strata — big counties and small counties. Within each stratum systematic sampling procedure of persons is used with different sampling fractions in the defined strata. Each selected person brings its household (hh) into the sample. All members 16+ of that hh are questioned. That way sufficient data can be collected to produce good estimates for various levels of the population with efforts to get comparable estimates also on the county level.

2.1. Probabilistic description of the sampling design and sampling weights

We use sampling vector framework (Traat et al., 2004). We assume simple random sampling without replacement (SI) of persons in strata, instead of systematic sampling. This is often done while working out variance formulae and is justified by the knowledge that a systematic sample taken from the randomly ordered population is equivalent to a SI sample. It appears that the distribution of the sampling vector (sampling design) in a stratum is a classical well-known distribution in this case. Namely, the sampling design for households, in case they are selected by SI-sampling of persons from the population register, is the multivariate hypergeometric distribution (Traat, et al., 2001, Traat et al., 2004; Ilves, 2005). We call it hypergeometric design. Probabilistically, for small sampling fractions the hypergeometric design is close to the multinomial design, in the literature usually referred to as with-replacement design with unequal probabilities. Multivariate hypergeometric and multinomial distributions are well described in Johnson et al. (1997).

In practice, the true EU-SILC design slightly differs from the hypergeometric design. First, as already mentioned, instead of SI-sampling, persons are selected by systematic sampling. Second, under hypergeometric design the repeatedly (through its different members) selected hh has to be included into sample repeatedly (we assume this when working out variance formulae). In practice this hh is included only once. In reality, the repeated selection of an hh is very rare if the sampling fraction is small.

Let I_{hi} be the sampling indicator of the hh i (shows how many times the hh is sampled) in stratum h. The expected sampling count of that hh is

$$E(I_{hi}) = np_i, \quad p_i = m_{14hi} / M_{14h}, \tag{2.1}$$

where n is sample size in households, m_{14hi} is the number of 14+ people in hh i of stratum h, and $M_{14h} = \sum m_{14hi}$ is the total number of people in the frame (population register with 14+ persons). Note that here the index i refers to the hh, and summation, if not specified, runs through the entire population of hh's.

The expected sampling counts are proportional to the 14+ size of the households. The hh's of big size are more frequently sampled causing over-representation of big-size hh's. This needs down weighting by sampling weights:

$$w_{hi} = k_{hi} / E(I_{hi}), \quad h = 1,2,3, \tag{2.2}$$

where k_{hi} is an outcome of the sampling indicator I_{hi}, usually equal to 1. However, for the mathematical correctness we have to keep k_{hi}.

Remark 2.1. Usually in practice, the sample is very small compared to the population, implying that the repeated selection of hh's is very rare. The design is almost like a WOR design. For WOR designs the expected sampling count equals to the inclusion probability, $E(I_{hi}) = \pi_{hi}$.

Since each selected hh brings all its eligible members into sample then the sampling weights for all these members of hh i are equal and given in (2.2). In this report we do not consider the case where after nonresponse adjustments and calibration the personal weights in a household may differ.

For variance estimation of Laeken indicators also second order design characteristics are needed. They are specified later in this report.

3. LAEKEN INDICATORS AND THEIR ESTIMATORS

In December 2001 the Laeken European Council endorsed 18 statistical indicators to measure poverty and social exclusion in a comparable way in Member states of EU (see Eurostat [1]). They cover four important dimensions — financial poverty, employment, health and education). These indicators are called Laeken indicators, and are partly listed below:

I_1: At-risk-of-poverty rate;

I_{1a}: At-risk-of-poverty rate, by age and gender

I_{1b} : At-risk-of-poverty rate, by most frequent activity status and gender;

I_{1c} : At-risk-of-poverty rate, by household type;

I_{1d} : At-risk-of-poverty rate, by accommodation tenure status;

I_{1e} : At-risk-of-poverty threshold;

I_2 : Inequality of income distribution, S80/S20 income quintile share ratio;

…

I_{11} : Dispersion around at-risk-of-poverty threshold;

…

I_{14} : Inequality of income distribution Gini coefficient;

….

The list is not finished, additional indicators concerning different aspects of the society are worked out.

In this report we concentrate on at-risk-of-poverty rate I_1 (treatment of I_{1a} - I_{1d} is similar), on quantiles and their functions (median, I_{1e}, quintile share ratio I_2), Gini coefficient I_{14}.

3.1. Equivalized disposable income

The key notion in definition of Laeken indicators is Equivalized disposable income. The equivalized disposable income is defined as a new variable eqinc associated with a person (see Eurostat [1]). Let person i belong to the household (hh) k, then

$$eqinc_i = \frac{totinc_k}{eqsize_k}, \tag{3.1}$$

where $totinc_k$ is total income and $eqsize_k$ is equivalized size of hh k. The $eqsize_k$ is a sum of personal OECD weights in hh k: 1 for first adult, 0.5 for other adults (aged 14 or over), 0.3 for children (aged less than 14).

3.2. Estimators for indicators

In the following we denote y_i, x_i as study variables and w_i as a sampling weight (possibly adjusted). The index here refers to a person. The sampling weight is the same for the persons in the same household. Sample of persons is denoted by s.

The types of estimators of Laeken indicators considered in this report are.

Quantile estimators

Let y_i be sorted into ascending order. Then the estimated α-quantile of the variable y is

$$q_\alpha = \begin{cases} (y_j + y_{j+1})/2, & if \quad \sum_{i=1}^{j} w_i = \alpha \hat{M} \\ y_{j+1}, & if \quad \sum_{i=1}^{j} w_i < \alpha \hat{M} < \sum_{i=1}^{j+1} w_i \end{cases}, \tag{3.2}$$

where $\hat{M} = \sum_s w_i$ is the estimated population size (in persons).

The estimated median is received for $\alpha = 0.5$ and estimated quintiles for $\alpha = 0.2, 0.4, 0.6, 0.8$.

Several indicators are calculated as functions of quantile estimators. For example, the indicator at-risk-of-poverty-threshold (I_{1e}) is defined as 60% of median, so its estimator is

$$\hat{I}_{1e} = 0.6\, q_{0.5}. \tag{3.3}$$

The Income quintile share ratio (I_2) is estimated as

$$\hat{I}_2 = q_{0.8} / q_{0.2}. \tag{3.4}$$

The quantiles in the formulae (3.3) and (3.4) are based on the variable eqinc.

Estimators as ratios

Many Laeken indicators can be viewed as ratios and therefore estimated by

$$\hat{R} = \frac{\sum_s w_i y_i}{\sum_s w_i x_i}, \tag{3.5}$$

For Laeken indicators, y_i, x_i are usually $\{0,1\}$-variables indicating belonging of the individual to a certain group. For example, the estimator for the at-risk-poverty rate is received from (3.5) by setting $x_i \equiv 1$ and

$$y_i = \begin{cases} 1, & if \quad eqinc_i \le \hat{I}_{1e}, \\ 0, & otherwise. \end{cases} \tag{3.6}$$

In this case (3.5) takes the form:

$$\hat{I}_1 = \frac{\sum_s w_i y_i}{\sum_s w_i}. \tag{3.7}$$

Remark 3.1. In some applications (3.6) is defined with strict inequality. We prefer our definition to be consistent with distribution function definition. In this way $\hat{I}_1$ also estimates distribution function of the variable eqinc evaluated at $\hat{I}_{1e}$ (see section 4.3).

General formula (3.5) and its applications can be presented alternatively by taking into account the effect of $\{0,1\}$-variables. We end up with sums of weights over sets of people. For example at-risk-poverty rate takes the form:

$$\hat{I}_1 = \frac{\sum_{s1} w_i}{\sum_s w_i}, \tag{3.8}$$

where s1 denotes sampled people with $eqinc \le \hat{I}_{1e}$.

Analogously, for at-risk-poverty-rate by age and gender we have the basic formula (3.5) with

$$x_i = \begin{cases} 1, & if \quad i \in age/gender\ group, \\ 0, & otherwise, \end{cases}$$

$$y_i = \begin{cases} 1, & if \quad (i \in age/gender\ group)\ and\ (eqinc_i \le \hat{I}_{1e}), \\ 0, & otherwise. \end{cases}$$

The complication with ratio type estimator (3.5) appears in variance estimation if the domain variable y_i is defined with the random threshold, like with $\hat{I}_{1e}$ in (3.6). The latter is an additional source of variation whose effect is not clear. This case finds special attention in this paper.

Gini coefficient

The Gini coefficient is a statistical indicator used to measure the level of inequalities in an income distribution.

For better understanding of the meaning of Gini let us introduce the meaning of Lorenz curve (for longer introduction see Eurostat [2]). Let y_i (equivalized total net income, eqinc) be sorted in ascending order, $y_1 \le y_2 \le \cdots \le y_n$. Denote the population size M and $Y_i = \sum_{i=1}^{j} y_j$. Persons with unknown eqinc are excluded from calculations unless values are imputed.

For convenience, let also $Y_0=0$. Y_i represents the total income earned by the first i members of the income-sorted population.

The Lorenz curve is defined like the graph of Y_i versus i, for i from 0 to M, with the points (i, Y_i) joined by a straight line.

The definition of the Lorenz curve can be extended, without any modification, to distributions having negative incomes.

If there would be no financial inequality, i.e everyone had the same level of income, then X% of the population would earn also X% of the total income — in other words, the Lorenz curve would fall along the diagonal line shown in the above graph. This line is called the line of perfect equality. The degree of inequality in a population is shown by how far the Lorenz curve drops below the line of perfect equality.

The distance between the Lorenz curve and the line of perfect equality changes along the length of the curve, so what is needed is a single number, which summarizes the behaviour of the entire curve. One way to get such a number is to measure the area between the line of perfect equality and the Lorenz curve: the larger the area, the greater the inequality.

Gini coefficient is defined as the fraction of the total area under the line of perfect equality. In the numerator of this fraction is the area between the line of perfect equality and Lorenz curve. Gini therefore ranges from zero to one. For complete inequality, in which only one person has all the total income (if that were possible) the Lorenz curve would coincide with the straight lines at the lower and right boundaries of the curve, so the Gini coefficient would be one.

The inequality in two populations can be compared by comparing their Gini coefficients, the population with the higher value having greater inequality. Below is given an illustrative table of Gini indexes in different countries (United Nations 2004):

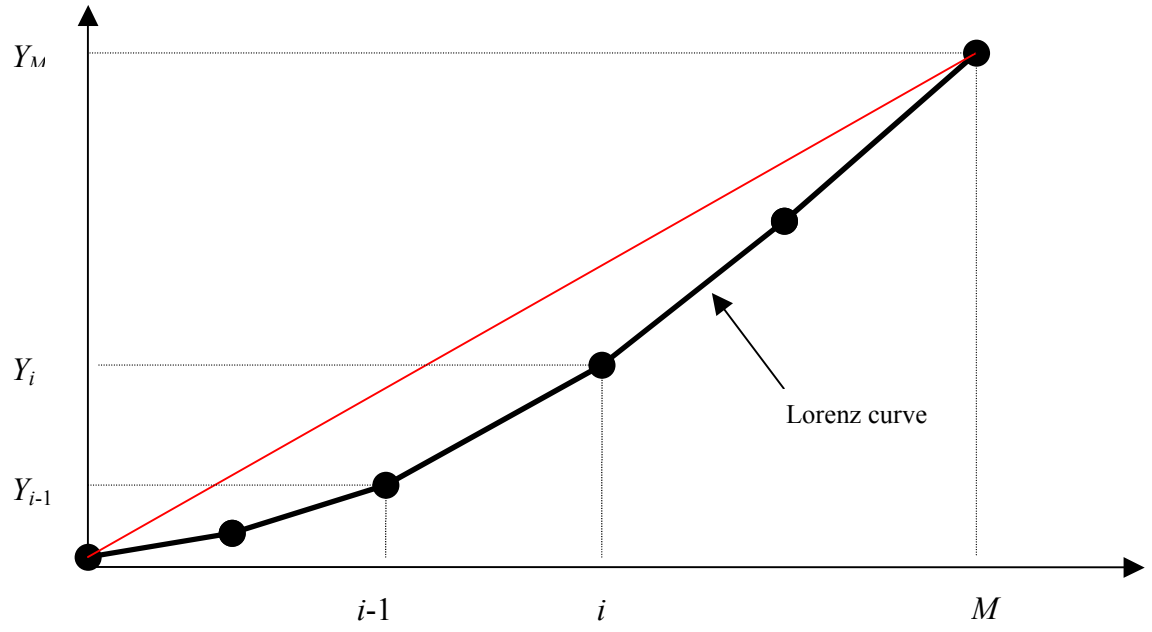
Table 3.1. **Gini coefficients in selected countries**, 2004

Country	Year	Gini index
Hungary	1999	0.244
Sweden	2000	0.250
Finland	2000	0.269
Germany	2000	0.283
Croatia	2001	0.290
Lithuania	2000	0.319
India	2000	0.325
UK	1999	0.360
Italy	2000	0.360
Estonia	2000	0.372
Turkey	2000	0.400
USA	2000	0.408
Russia	2000	0.456
Mexico	2000	0.546
Chile	2000	0.571

When Gini coefficient is based on the Lorenz curve of income distribution, it can be interpreted as the expected income gap between two individuals randomly selected from the population (Sen 1973).

Illustrated by the graph below the Gini index is equal to $M\,R/(M-1)$ where R is the ratio of the area enclosed by the two Lorenz curves by the area below the 45 degree line. The bigger the Gini index, the bigger is the inequality of incomes in the population.

Figure 3.1. **Lorenz curve**



Computationally the Gini coefficient G is estimated as (see Eurostat [2])

$$\hat{G} = \frac{2 \cdot \sum_s w_i \cdot y_i \cdot \hat{M}_i - \sum_s w_i^2 \cdot y_i}{\hat{M} \cdot \sum_s w_i \cdot y_i} - 1, \tag{3.9}$$

where $\hat{M} = \sum_s w_i$ is the estimated population size (in persons) and $\hat{M}_i = \sum_{j=1}^{i} w_j$ is the estimated population size with eqinc smaller or equal to that of person i.

4. VARIANCE OF LAEKEN INDICATORS

There are several moments, which need special attention when developing variance formulae for Laeken indicators:

- sampling unit is household but the estimators are formed with person-level data;
- sampling design is complex — unequal probabilities for households and persons;
- estimators are non-linear;
- domain variable in estimators has a random threshold
- calibration and weight adjustments may have introduced unequal weights for persons of the same household.

4.1. If sampling unit is household

The Laeken indicators as defined in Section 3 are calculated with person-level data. Correspondingly, the study variable y_i is associated with person i, and s denotes sample of persons. Also data file is organised by persons — each row corresponding to a certain person, including all of its data. However, persons are included into sample through their households. The sampling unit is household. Therefore variance formulae have to be developed in terms of hh's.

Let us denote s_{hh} — sample of hh's. Assuming that the weight is the same for each hh member we can write the person-level weighted sum as an hh-level weighted sum:

$$\hat{t}_y = \sum_s w_i y_i = \sum_{s_{hh}} w_i y'_i ,$$

where $y'_i = y_{i1} + y_{i2} + \dots y_{im_i}$ is the aggregated value of all persons in hh i: y_{ij} is the value of person j in hh i, m_i is the number of questioned persons in hh i. In a special case y_{ij} being a domain indicator, y'_i is still the sum, now meaning the number of persons in hh i belonging to that domain. For example, the estimated population size $\hat{M} = \sum_s w_i$ can be expressed as $\hat{M} = \sum_{s_{hh}} w_i m_i$, where m_i is the no. of people in the hh i. As a conclusion, there is no difficulty to express the estimators involving person-level sums in terms of hh-level sums.

In the subsequent variance formulae we skip the prime and let y_i denote both the person-level and hh-level value. Its meaning will be clear from the context, e.g. often from the summation set, s or s_{hh}.

When deriving variance estimators for quantiles the basic step is to construct variance estimators for distribution function at certain locations. Distribution function at a location is estimated by a ratio where both in the numerator and denominator are weighted sums. As became clear from the above it is not difficult to express these sums in terms of hh data.

4.2. Design characteristics while deriving variance estimators

For variance estimation also second order design characteristics are needed. For the notational simplicity we assume only single stratum and further skip the stratum index.

Let I_i be the sampling indicator of the hh i (shows how many times the hh is sampled). Then under hypergeometric design we have (properties of the multivariate hypergeometric distribution in Johnson et al., 1997):

$$E(I_i) = np_i, \quad V(I_i) = c\,n\,p_i q_i, \quad p_i = m_{14i}/M_{14}, \quad q_i = 1 - p_i, \quad c = (M_{14} - n)/(M_{14} - 1),$$

$$\Delta_{ij} = Cov(I_i, I_j) = -c\,n\,p_i p_j, \quad i \neq j,$$

$$E(I_i)^2 = n\,p_i(cq_i + np_i), \quad E(I_i I_j) = n\,p_i p_j (n - c), \quad i \neq j,$$

$$\breve{\Delta}_{ij} = \Delta_{ij} / E(I_i I_j) = -\frac{c}{(n-c)} = -\frac{M_{14} - n}{M_{14}(n-1)}, \quad i \neq j,$$

$$n - \text{sample size in hh's},$$

$$m_{14i} - \text{no. of } (14+) \text{ persons in hh},$$

$$M_{14} = \sum m_{14i} - \text{no. of people in the frame (population register with } (14+) \text{ people)},$$

4.3. Variance of ratio type estimators

The ratio type estimator in terms of hh-level data y_i, x_i is:

$$\hat{R} = \frac{\sum_{s_{hh}} w_i y_i}{\sum_{s_{hh}} w_i x_i} \underset{\text{denote}}{=} \frac{\hat{t}_y}{\hat{t}_x}, \quad (4.1)$$

If the value of y_i is fixed for the hh i, i.e. does not depend on the sample (what is not the case for random threshold), the variance estimator of (4.1) is well-known. It is developed from the Taylor series of (4.1):

$$\hat{R} \approx R + (\hat{t}_y - R\hat{t}_x)/t_x \underset{\text{alternatively}}{=} R + \sum_{s_{hh}} w_i (y_i - Rx_i)/t_x, \quad (4.2)$$

where t_y and t_x are the population sums and $R = t_y / t_x$ is the population ratio. The sum in (4.2), if R and t_x are fixed, is an ordinary design-weighted estimator and its Horvitz-Thompson type variance estimator has the form (e.g. Särndal et al. 1992, p. 176-181)

$$\hat{V}(\hat{R}) = \hat{t}_x^{-2} \sum\sum_{s_{hh}} \breve{\Delta}_{ij} w_i w_j (y_i - \hat{R}x_i)(y_j - \hat{R}x_j), \quad (4.3)$$

where

$$\breve{\Delta}_{ij} = \frac{E(I_i I_j) - E(I_i)E(I_j)}{E(I_i I_j)} \quad (4.4)$$

is design characteristic with I_i as sampling indicator of hh i. The formula (4.4) is usually given with inclusion probabilities $\breve{\Delta}_{ij} = (\pi_{ij} - \pi_i \pi_j)/\pi_{ij}$. Its form with expectations in it is not so known. It is a generalization, allowing to consider WR- and WOR designs simultaneously (e.g. Traat, Meister, Sõstra, 2001). The EU-SILC design has a fixed size in hh's. For fixed size designs there is an alternative variance estimator, the SYG-estimator (Sen-Yates-Grundy), which is more stable (Särndal et al., 1992, p. 45). In our more general setting, and expressed for a ratio, it has the following form:

$$\hat{V}(\hat{R}) = -\frac{1}{2\hat{t}_x^2} \sum\sum_{s_{hh}} \breve{\Delta}_{ij} \left(\frac{y_i - \hat{R}x_i}{E(I_i)} - \frac{y_j - \hat{R}x_j}{E(I_j)}\right)^2. \quad (4.5)$$

From this general form we have derived (Appendix 1) a special form for the hypergeometric design of households:

$$\hat{V}(\hat{R}) = \frac{M_{14} - n}{M_{14}} \frac{1}{n(n-1)} \frac{1}{\hat{t}_x^2} \sum \left(\frac{y_i - \hat{R}x_i}{p_i}\right)^2 I_i, \quad (4.6)$$

where M_{14} is number of people in the frame, n is number of sampled hh's, p_i is selection probability of the hh i (see Section 4.2), the sum goes over the hh population, but due to the indicator I_i essentially only over sampled hh's.

Remark 4.1. y_i is not always fixed value of the unit i. In our case, it depends on the sample (it is 0 or 1 depending on the sample-calculated threshold). In this case y_i is replaced by the sample-defined $\tilde{y}_i$.

In the special case of at-risk-of-poverty rate we have $\hat{R} = \hat{I}_1$, $x_i = m_i$ (no. of people in hh i), $\hat{t}_x = \sum_{s_{hh}} w_i m_i = \hat{M}$ (estimated total of people), y_i the no. of people under poverty threshold. Since study variable is eqinc which is the same for each hh member then $y_i = m_i \tilde{y}_i$, where

$$\tilde{y}_i = \begin{cases} 1, & \text{if} \quad \textit{eqinc} \text{ of a person in hh } i \text{ is} \le \hat{I}_{1e}, \\ 0, & \textit{otherwise}. \end{cases}$$

With these replacements the variance formula (4.6) takes the form:

$$\hat{V}(\hat{I}_1) = \frac{M_{14} - n}{M_{14}} \frac{1}{n(n-1)} \frac{1}{\hat{M}^2} \sum \left(\frac{m_i}{p_i}\right)^2 (\tilde{y}_i - \hat{I}_1)^2 I_i. \quad (4.6a)$$

The variance formulae for other ratio type estimators like e.g. $\hat{I}_{1a}, \ldots, \hat{I}_{1d}$ come analogously with suitable replacements from (4.6).

4.4. Variance of quantile estimators and of their functions

In fact, the quantile estimators in (3.2) are received by inverting the estimated distribution function (for the median, Särndal, et al., 1992, p.197-205). The true finite population distribution function of y_i at y is a ratio:

$$F(y) = \frac{\#\{y_i \le y\}}{M}.$$

This can be estimated by ratio estimator whose variance is known. With this information the confidence limits for F(y) are constructed. Inverting these limits by the function F^{-1} we get confidence limits for the quantiles — for median if F(y)=0.5, for quintiles if F(y)=0.2, 0.4, 0.6, 0.8.

Let Q_α be the true α-quantile of the variable y_i (i for person), meaning $F(Q_\alpha) = \alpha$. We can estimate $F(Q_\alpha)$ by an estimator analogous to (3.5):

$$\hat{F}(Q_\alpha) = \frac{\sum_s w_i z_i}{\sum_s w_i}, \tag{4.7}$$

where

$$z_i = \begin{cases} 1, & if \quad y_i \le Q_\alpha, \\ 0, & otherwise. \end{cases} \tag{4.8}$$

By equating $\hat{F}(Q_\alpha) = \alpha$ and inverting it we get the estimator of the true Q_α:

$$\hat{F}^{-1}(\alpha) = \hat{Q}_\alpha = q_\alpha,$$

where q_α is calculated in (3.2).

For variance estimation we have the same complication as before — the sampling unit is hh but the distribution function under interest is built on the person-level data. Again, we can solve it by expressing the basic ratio (4.7) in terms of hh's:

$$\hat{F}(Q_\alpha) = \frac{\sum_{s_{hh}} w_i m_i z_i}{\sum_{s_{hh}} w_i m_i}, \tag{4.9}$$

where i now refers to hh, m_i is no. of eligible members in hh i and

$$z_i = \begin{cases} 1, & if \quad y_{ij} \le Q_\alpha, \ \forall j, \quad (j \text{ refers to the member of hh } i), \\ 0, & otherwise. \end{cases} \tag{4.10}$$

In our case y_i is eqinc, which has the same value for each hh member. Variance estimator for $\hat{F}(Q_\alpha)$ in (4.9) is already given in general formula (4.5). We only have to replace y_i by $m_i z_i$ and x_i by m_i, imposing $\hat{t}_x = \hat{M}$ $(= \sum w_i m_i)$. Using also $\hat{F}(Q_\alpha) = \hat{R} = \alpha$ we get:

$$\hat{V}(\hat{F}(Q_\alpha)) = -\frac{1}{2\hat{M}^2} \sum\sum_{s_{hh}} \breve{\Delta}_{ij} \left(m_i \frac{z_i - \alpha}{E(I_i)} - m_j \frac{z_j - \alpha}{E(I_j)}\right)^2. \tag{4.11}$$

For the hypergeometric design with $E(I_i) = np_i$ this formula takes the analogous form to (4.6):

$$\hat{V}(\hat{F}(Q_\alpha)) = \frac{M_{14} - n}{M_{14}} \frac{1}{n(n-1)} \frac{1}{\hat{M}^2} \sum \left(\frac{m_i}{p_i}\right)^2 (z_i - \alpha)^2 I_i. \tag{4.12}$$

Finally, note that z_i in (4.10) is not calculable, since Q_α is not known. We use $\tilde{z}_i$ defined with q_α:

$$\tilde{z}_i = \begin{cases} 1, & if \quad y_{ij} \le q_\alpha, \ \forall j, \quad (j \text{ refers to the member of hh } i), \\ 0, & otherwise. \end{cases} \tag{4.13}$$

Since in practice repeated selection of hh's is very rare, the I_i can be taken as Bernoulli variable, and therefore the form of variance estimator for practice is:

$$\hat{V} = \hat{V}(\hat{F}(Q_\alpha)) = \frac{M_{14} - n}{M_{14}} \frac{1}{n(n-1)} \frac{1}{\hat{M}^2} \sum_{s_{hh}} (\frac{m_i}{p_i})^2 (\tilde{z}_i - \alpha)^2 . \quad (4.14)$$

For mathematical correctness, especially, if the repeated selection is more frequent, the repetition count I_i should be kept in mind.

Now, if $\hat{F}(Q_\alpha)$ is approximately normally distributed (which is the case for big samples) we can say that (c_1, c_2) is an approximate 95% confidence interval for $F(Q_\alpha)$, where

$$c_1 = \alpha - 1.96\sqrt{\hat{V}} \ , \ c_2 = \alpha + 1.96\sqrt{\hat{V}} \ . \quad (4.15)$$

Inverting the points c_1, c_2 with $\hat{F}^{-1}$ which means that we calculate q_{c_1}, q_{c_2} from (3.2), we get that (q_{c_1}, q_{c_2}) is the approximately 95% confidence interval for Q_α. From the last interval one can also estimate the variance of quantile estimator (assuming normality):

$$\hat{V}(q_\alpha) = [(q_{c1} - q_{c2})/(2 \cdot 1.96)]^2 \quad (4.16)$$

Remark 4.3. We made several approximations when deriving $\hat{V}$. One of them was replacement of z_i by $\tilde{z}_i$ with random threshold q_α. Therefore, as noted also by Särndal et al. (1992, p. 203), the method should be used with caution. We check its performance in our simulation study.

To develop variance of functions of quantiles (like quintile share ratio) one has to develop this function into Taylor series and find variance of it using knowledge on the variances of the components of Taylor series.

The most simple function of quantiles is the poverty threshold $\hat{I}_{1e} = 0.6 \, q_{0.5}$. Its variance in a straightforward way is

$$\hat{V}(\hat{I}_{1e}) = 0.6^2 \, \hat{V}(q_{0.5}) , \quad (4.17)$$

where $\hat{V}(q_{0.5})$ is calculated from (4.16) with $\alpha = 0.5$.

Remark 4.4. The variance $\hat{V}(\hat{I}_{1e})$ can be calculated alternatively with inversion of distribution function like it was done for quantiles. For this, one has to recognize that $\hat{I}_{1e}$ is a quantile of the variable eqinc. At the location $\hat{I}_{1e}$ the value of the distribution function is estimated by $\hat{I}_1$. Since variance of the latter is known, the inversion method can be applied. However, in our simulation experiments the simplest estimator (4.17) was more stable.

In case one is interested in confidence limits of I_{1e}, the normal-based symmetric limits constructed with (4.17) around $\hat{I}_{1e}$ are not suggested. The closest coverage rate to the nominal is given by the limits $q_1 = 0.6 q_{c_1}$, $q_2 = 0.6 q_{c_2}$ where q_{c_1}, q_{c_2} are confidence limits of the median (see section 5.3).

The variance formula for the income quintile share ratio $\hat{I}_2$ is not derived in this report but its variability is studied in the simulation experiment.

4.5. Variance of Gini coefficient

The variance properties of Gini coefficient $\hat{I}_{14}$ (here denoted by G) are not known for small samples. Also the large sample approximations to the variance of G are quite poor. Most of the formulations are mathematically complex, or they require a considerable amount of numerical computation.

Still different studies in the field are in progress. There are linearization methods. A general linearization-based variance estimator is given in Nygård and Sandström (1985). It has to hold for any (WOR) sampling design. But the expression is so complex, involving inclusion probabilities beyond the second order, so that it is hardly possible to develop it for some special complex sampling design. There are other analytic methods trying to consider linear term together with an approximated nonlinear term, but these methods are restricted to the simple designs. For example, Bloznelis (2004) has been working on estimating the variance of Gini coefficient under stratified simple random sampling design. Another branch of the methods is formed by resampling methods. Some authors have proposed the jackknife re-sampling technique to approximate a standard error for the Gini coefficient (see Giles, 2002). It has been found that the jackknife method is working satisfactorily for variance estimation in the case of Gini coefficient and it produces trustworthy confidence intervals for it, but again, not for all sampling designs. As stated by Yitzhaki (1999), for nonstratified samples, the calculation of the jackknife variance estimator requires only two runs over the data, and therefore enables users to calculate standard errors of the of the Gini estimates using a standard personal computer.

Some authors have discussed calculating variance estimates and confidence intervals of Gini via bootstrap re-sampling methods (see Dixon, 1993). There are many resampling possibilities in sample surveys. One can vary the resampling design in many possible ways. These issues are considered in Ollila (2003), and found that there are no unique methods and suggestions working well with any sampling design and any complex estimator.

However, the resampling methods are appealing due to their applicational simplicity. In this work we concentrate on the Jacknife method. We use the results derived in a handwritten manuscript Traat (2003). In this manuscript the statistic considered was the design-weighted sample sum. The two-phase sampling framework was considered with design-vectors approach. In the first phase the multinomial design was assumed, which is probabilistically very close to our hypergeometric design of households. In the second phase the SI-sampling of already selected hh's (multiples included) was assumed. In this scheme Jackknife is SI-sampling of size n-1. The resulting formula was

$$V_a(\hat{t}_a) = (n-1)V(\hat{t} \mid I_a), \tag{4.18}$$

where a refers to the first phase design, given I_a means, given the first phase sample, $\hat{t}_a = \sum (I_{ai} y_i)/np_i$ is the first-phase unbiased estimator and $\hat{t} = \sum (I_i \mid I_{ai}) y_i /(n-1)p_i$ is unbiased to $\hat{t}_a$ with respect to the second-phase design. We are interested in estimating $V_a(\hat{t}_a)$. The important thing here is the fact that $V(\hat{t} \mid I_a)$ can be estimated as variance of $\hat{t}$ over all Jacknife resamples. The result (4.18) which is valid under EU-SILC design for sample sums, is here extended to the Gini coefficient. We try to estimate variance of Gini coefficient under EU-SILC design by finding its Jacknife variance and multiplying it with n-1 (n sample size of hh's). This idea is studied in the simulation experiment.

5. SIMULATION STUDY

5.1. Population

In order to perform the simulation study the population is generated using the data from the Estonian Household Budget Survey 2003. In further stage the results can be redone on the data from the recently finalized data from the Estonian EU-SILC 2004 survey.

The database of the HBS 2003 involves data for 4615 households and their 12532 members. The survey uses a stratified systematic sampling design with three strata of counties formed on the population size (see Statistical Office of Estonia, 2003). Similarly with the EU-SILC survey the households are selected from the population register by a systematic sampling procedure with different inclusion probabilities for strata. The inclusion probabilities for households (and persons) are proportional to size.

In order to remove the effect of stratification some households from strata with bigger inclusion probabilities were removed. The sampling rate was calculated directly from the initial inclusion probabilities that were used in the selection procedure. The results are listed in the following table.

Table 5.1. **Sampling rates for strata to form the simulation population.**

Stratum	Initial inclusion probability in HBS 2003 (%)	Sampling rate for simulation population
Big counties	0.0357	1
Small counties	0.0735	0.485714286
Hiiu county	0.2095	0.170373

Similar step was taken to remove the effect of household size a portion of bigger households were removed by the following scheme:

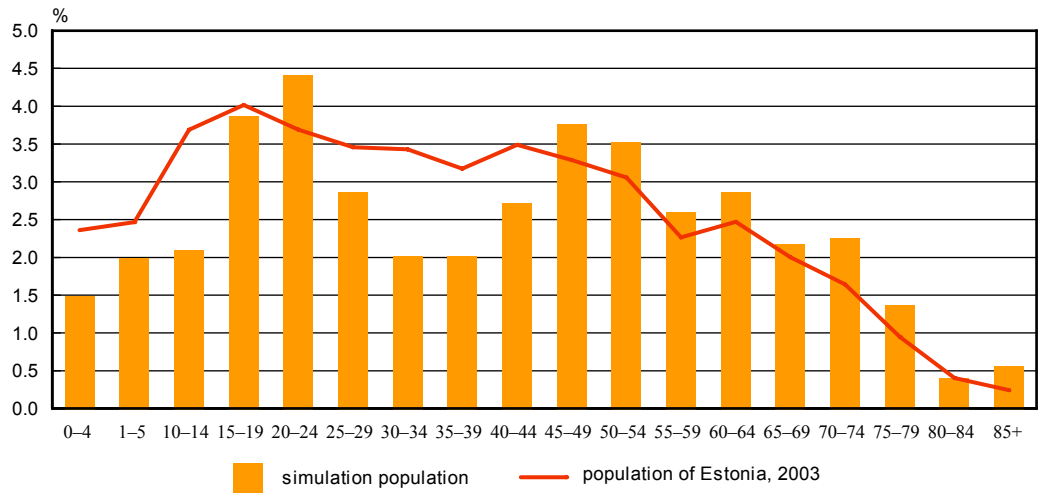
HH sized 2 persons → 1/2 of the sample removed;
HH sized 3 persons → 2/3 of the sample removed;
HH sized 4 persons → 3/4 of the sample removed;
HH sized bigger than 4 → 4/5 of the sample removed.

With these manipulations we hoped to get a quite representative database for Estonia. The database includes information on 3583 persons. The SAS program to perform the described selection is given in Appendix 2.

However, all these people did not have income variables since their households had participated only in the first interview where they gave general hh data, and had not given diary-based income data. We excluded these hh's and corresponding people from our population (programs in Appendix 2).

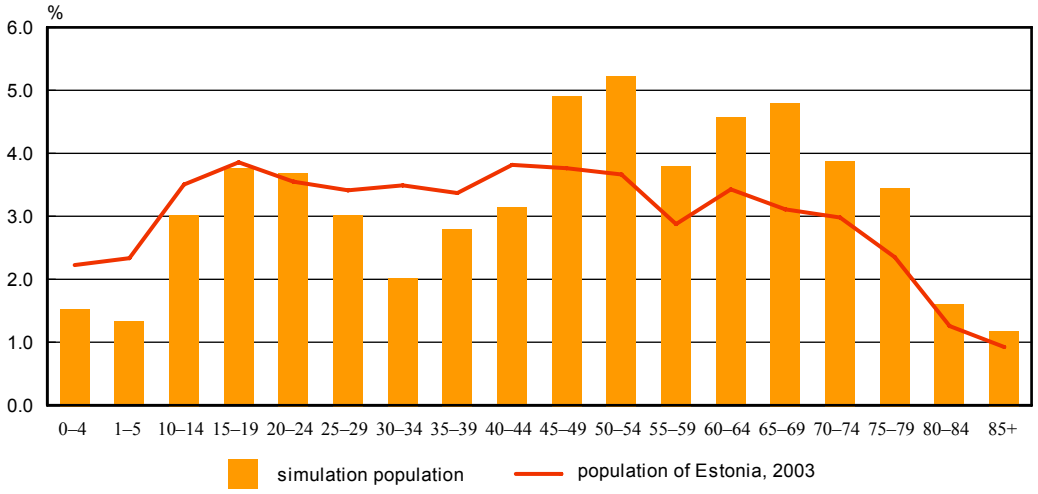
On the graphs below our simulation population is compared to the true Estonian population with respect to the sex-age distribution.

Figure 5.1. **Comparison of age distribution of men**



Graphs 5.1 and 5.2 are just for illustrative purposes. Our aim is to study variance estimators in a simulation experiment where it is not necessary that all the distribution in the simulation population exactly equal to the distributions in the true Estonian population.

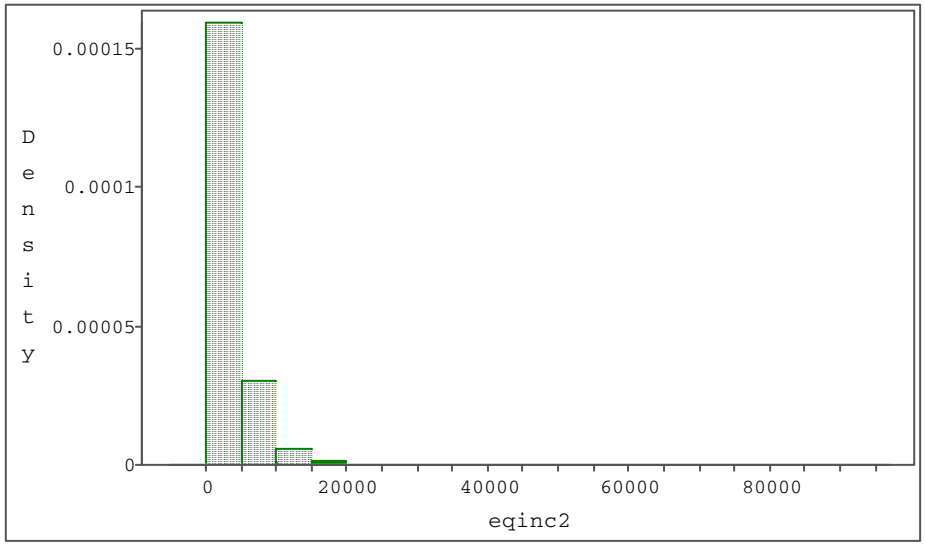
Figure 5.2. **Comparison of age distribution of women**



Basic characteristic of our population, for which simulation was targeted, where:

Total no. of people 2595;
Total no. of hh's 1263;
Median of eqinc 3000;
Poverty threshold $I_{1e} = 1800$;
At-risk-of-poverty rate $I_1 = 15.34\%$;
Income quintile share ratio $I_2 = 2.58$;
Gini coefficient $I_{14} = 0.370$.
The population distribution of the basic study variable eqinc is very asymmetric as seen below.

Figure 5.3. **Distribution of equivalized income in the population**



The Sampling Frame does not include all people but people 14+. All people in it have their hh identifiers. The programmes for the creation of Frame are in Appendix 2. The crucial characteristic of the Frame is:

Total no. of Frame people $M_{14} = 2332$.

5.2. Sampling and calculation of basic quantities

We have used two different environments for our simulations, the IML and the macro-environments, this is due to the availability of SAS-modules.

The SAS IML environment was used when studying $\hat{I}_1, \hat{I}_{1e}, \hat{I}_2$ and the median. The hypergeometric sampling was carried out in the following way. SI-sampling of persons was performed in the Frame. Their hh's were included into sample. Sample size was fixed to 100 hh. Data of all the members of these hh's were used. The basic study variable was equivalized income.

In each sample the median, the poverty threshold $\hat{I}_{1e}$, the at-risk-of-poverty rate $\hat{I}_1$ and and the income quintile share ratio $\hat{I}_2$ were calculated, as given in the formulae (3.2)-(3.4) and (3.7). Also modified $\hat{I}_1$, with fixed poverty threshold, was calculated. This was done in 4000 repetitions. Average and variance over all these 4000 outcomes characterizes the sampling bias and the true sampling variance of each of these quantities.

The performance of the derived variance formulae was checked in the same simulation experiment. In each of the samples $\hat{V}(med)$ was calculated with (4.16) in which the basic quantity, variance of the distribution function $\hat{V}(\hat{F})$, was calculated with (4.14), $\hat{V}(\hat{I}_{1e})$ was calculated with (4.17) and simultaneously with an alternative method described in Remark 4.4, and finally $\hat{V}(\hat{I}_1)$ was calculated with (4.6a).

For simulation study of Gini coefficient SAS Macro Language facilities were used. Sampling was performed by SAS procedure Surveyselect.

To describe the distribution of Gini coefficient a simulation was made, in which 1000 samples were drawn from the population according to the Estonian EU-SILC survey design. Therefore 100 persons were drawn from the Frame by SI-sampling and their hh's were included into the sample. Data of all the members of these hh's were used, whereas equivalized income was used to calculate the Gini coefficient. The results of this activity give us the true sampling distribution of the Gini coefficient.

We study the Jackknife method for variance estimation of Gini. First one sample of size 100 was selected from the Frame and the households of those persons were included in the sample (EU-SILC design). From this initial sample new samples were generated with exclusion of a household at a time. This way 100 samples of size 99 hh's were generated. Next, Gini index was calculated in each of the replicated samples and its distribution studied. Variance of Gini's was found over all the replicates. Its value was multiplied by 99 to get an estimate of the Gini variance under EU-SILC design, see Section 4.5.

Not to stay with only one resampling situation we did the procedure through with many initial samples (taken by the EU-SILC design). The Gini distribution was studied each time (as in single case) and then averaged over the 100 different cases. As a result for both Gini index and its standard deviation empirical point estimates and standard deviation were calculated, averaged over 100 samples. The results are analyzed in Section 5.3 and given in the Appendix 4.

5.3. Analysis of simulation results

The simulation results are given in Appendix 4. Here we bring some summarizing tables and comments.

Table 5.2. **Estimated median and related quantities**

	med	$\sqrt{\hat{V}(med)}$	$\sqrt{\hat{V}(\hat{F})}$
Mean	3036	216.6	0.05
Std Dev	216.5	52.3	0.0006

We see that median as calculated by (3.2) on average slightly overestimates the true median 3000. Overestimation is small: $36/3000 \approx 1\%$. The true sampling variability of the median is not big, $c.v. = 216.5/3036 \approx 7\%$. The second column says that the variance formula worked out by us performs very well, it produces almost unbiased variance estimator (compare 216.6 and 216.5), and the estimator is also quite stable (with standard deviation 52.3). The third column characterizes variability of the estimated distribution function at the estimated median. This is basic component when finding confidence intervals and variance of the median with inversion method. Stability of $\hat{F}$ guarantees good performance of the method. It appeared that the confidence intervals (4.15) of the median worked very well. The coverage rate was 95.2% instead of 95%.

The poverty threshold $\hat{I}_{1e}$ is median-based quantity and its performance is much defined by the median.

Table 5.3. **Poverty threshold $\hat{I}_{1e}$ and related quantities**

	$\hat{I}_{1e}$	$\sqrt{\hat{V}(\hat{I}_{1e})}$ formula (4.17)	$\sqrt{\hat{V}(\hat{I}_{1e})}$ inv. distr. method Remark 4.4
Mean	1821.8	130.0	155.1
Std Dev	129.9	31.4	44.5

Since the sample median overestimated the true one, it is reasonable to expect that the related quantity, poverty threshold, also overestimates the true value 1800. This is convinced by the Table 5.3. The true standard deviation of $\hat{I}_{1e}$ (129.9) is very well estimated by our formula (4.17). The estimator is unbiased and quite stable. The inverse distribution method for variance estimation, as described in Remark 4.4, does not work well here. The estimator has considerable bias and much bigger variability compared to the estimator in (4.17). The coverage rate of the symmetric confidence intervals constructed with (4.17) was 93.2%. The coverage rate of the confidence intervals derived from the ones for the median, was 95.2% (as for the median), thus much more exact.

Table 5.4. **At-risk-of-poverty rate $\hat{I}_1$ and related quantities**

	$\hat{I}_1$	$\hat{I}_1$ with fixed threshold	$\sqrt{\hat{V}(\hat{I}_1)}$ formula (4.6a)
Mean	0.157	0.153	0.037
Std Dev	0.039	0.034	0.0046

The $\hat{I}_1$ in the first column was calculated wit (3.6)-(3.7), i.e. with sample estimated poverty threshold. In the second column, the threshold was fixed to the true one, 1800. The true value is $I_1 = 0.1534$. The aim was to find out the increased (or decreased) variability due to the random threshold. The increase in standard deviation is around 15% (5/34). Nevertheless our variance formula seems to catch this increased variability (compare 0.037 and 0.039). The variance estimator is quite stable. The symmetric confidence intervals are conservative, the coverage rate was 97.5%.

The quintile share ratio was studied only in the simulation experiment. Its variance estimator was not derived.

Table 5.5. **Quintile share ratio $\hat{I}_2$**

	$\hat{I}_2$
Mean	2.61
Std Dev	0.29

The true value 2.58 is estimated almost unbiasedly (mean 2.61). Variability of this quantity is not big, $c.v.(\hat{I}_2) = 0.29 / 2.61 \approx 11\ \%$.

The sampling distribution of the Gini coefficient is summarized in the table below.

Table 5.6. **Gini coefficient $\hat{I}_{14}$**

	$\hat{I}_{14}$
Mean	0.363
Std Dev	0.041

In the next table the mean and standard deviation of Gini in the Jackknife resamples is given. The 3 initial samples with EU-SILC design are taken as basis.

Table 5.7. **Gini Coefficient in Jackknife samples**

	1st sample	2nd sample	3rd sample
Mean	0.454	0.368	0.350
Std Dev	0.0058	0.0025	0.0024
10 Std Dev	0.058	0.025	0.024

The true population Gini is 0.37. We see that under EU-SILC we get very slight underestimation of Gini. In Jackknife resampling the mean varies around the true value. More important, the direct Jackknife standard deviation is too small for describing true Gini standard deviation, but multiplying with $\sqrt{n-1}$ brings the numbers into right scale, see formula (4.18).

In the next table instead of 3 initial samples, 100 samples are observed and the means and standard deviations of each sample are averaged.

Table 5.8. **Gini coefficient over 100 initial samples**

	$\hat{I}_{14}$
Mean	0.365
Std Dev	0.0041
10 Std Dev	0.041

We see that on the average Jackknife works very well. Of course, in practice with only one sample at hand one has to be ready that Jackknife standard deviation is 2 times smaller than the true one. But anyway it can be used to get some hints about the variability of Gini coefficient.

6. SUMMARY

In this Report we concentrated on the EU-SILC design in Estonia and derived variance estimators of many Laeken indicators ($\hat{I}_1$, $\hat{I}_{1e}$ and median) under this design. With straightforward replacements these variance formulae are readily applicable to other related indicators like $\hat{I}_{1a},\ldots,\hat{I}_{1d}$ e.g. Special concern was on the fact that the variance formulae included sample dependent variables, i.e. y_i was not fixed to the unit i, instead was dependent on the sample-calculated threshold. The derived formulae were checked in the simulation experiment which mimicked the true sampling situation in Estonia. As a result we can state that all the derived variance estimators performed very well, they were almost unbiased and quite stable. The random threshold, used for defining some Laeken indicators (e.g. $\hat{I}_1$), slightly increased the variability, nevertheless this was well caught by the variance estimators. The method to plug in the sample-dependent y_i-variable in the variance estimators worked very well with our design.

For variance of Gini coefficient we applied the formula derived for Jackknife variance of sample sums in the sampling situation similar to EU-SILC. It appeared in the simulation study that the formula works considerably well.

During the work some additional issues emerged which might be studied under similar arrangements with ICON institute and statistical offices. We can mention here the nonresponse and corresponding adjustment of variance formulae; unequal weights for individuals in the same hh, the method developed here for variance estimation needs to be reconsidered then; the variance of complex functions (like income quintile share ratio).

REFERENCES

Bloznelis M. (2004) Normal approximation for stratified samples. In: Workshop on Survey Sampling Theory and Methodology, Tallinn, Statistical Office of Estonia, 18-23.

Giles D.E.A. (2002) Calculating a Standard Error for the Gini Coefficient: Some Further Results. Working Paper.

Ilves, M. (2005) Variance and its estimator for a practical self-weighting two-phase design. CD of the 55th Session of the International Statistical Institute (ISI), 5-12 April 2005, Sydney, Australia, ISBN: 1877040282, 14 pp.

Johnson, N.L., Kotz, S., Balakrishnan, N. (1997) Discrete Multivariate Distributions. New-York: Wiley.

Nygård, F., Sandström, A. (1985). Estimation of the Gini and the entropy inequality parameters in finite populations. Journal of Official Statistics 1, 399-412.

Ollila, P. (2003) A theoretical overview for variance estimation in sampling theory with some new techniques for complex estimators. Research Report 240 (Ph.D. thesis). Statistics Finland.

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

Sen A. (1973) On Economic Inequality. Oxford: Clarendon Press.

Statistical Office of Estonia (2003). Leibkonna eelarve uuring, 2002. Metoodiline ülevaade. Household Budget Survey, 2002. Methodological Report. Tallinn: Statistical Office of Estonia.

Traat, I., Bondesson, L. Meister, K. (2004) Sampling design and sample selection through distribution theory. Journal of Statistical Planning and Inference, vol. 123, 395-413.

Traat, I., Meister, K., Sõstra, K. (2001) Statistical inference in sampling theory. Theory of Stochastic Processes, vol. 7(23), 301-316.

Yitzhaki S. (1991) Calculating Jackknife Variance Estimators for Parameters of the Gini Method. Journal of Business & Economic Statistics, vol 9 , 235-239.

Eurostat and United Nations documents

[1] EU-SILC 131/04: Common Cross-sectional EU Indicators Based on EU-SILC; The Gender Pay Gap. Working Group on Statistics on Income and Living Conditions (EU-SILC), 29-30 March 2004, Eurostat, Luxembourg.

[2] EU-SILC 131-A/04: Theoretical Study of the Gini Index. June 2004, Eurostat, Luxembourg.

[3] DOC. E2/IPSE/2003: The 'Laeken' Indicators : Detailed Calculation Methodology. Working Group on Statistics on Income, Poverty and Social Exclusion. 28-29 April 2003, Eurostat, Luxembourg.

[4] United Nations (2004) Gini Index calculated for all countries. United Nations Human Development Report 2004, p50-53, available at:

http://hdr.undp.org/reports/global/2004/pdf/hdr04_HDI.pdf

APPENDICES

Appendix 1. Derivations

A1.1 SYG variance estimator under hypergeometric design

Let us consider an unbiased estimator of the population total $t = \sum y_i$ (if the summation set is not specified then the summation is over population):

$$\hat{t}_y = \sum I_i y_i / E(I_i) = \sum I_i \breve{y}_i ,$$

where $\breve{y}_i = y_i / E(I_i)$. Its unbiased SYG variance estimator has the form:

$$\hat{V}(\hat{t}_y) = -\frac{1}{2}\sum\sum_{s_{hh}} \breve{\Delta}_{ij}(\breve{y}_i - \breve{y}_j)^2 , \qquad (A1.1)$$

In the following we derive from (A1.1) the SYG estimator under hypergeometric design. Inserting sampling indicators in (A1.1) we have the double sum over the entire population, then developing the square we have:

$$\hat{V}(\hat{t}_y) = -\frac{1}{2}\sum\sum_{i\neq j} \breve{\Delta}_{ij}(\breve{y}_i^2 - 2\breve{y}_i\breve{y}_j + \breve{y}_j^2) I_i I_j . \qquad (A1.2)$$

For our hypergeometric design $\breve{\Delta}_{ij} = -(M_{14} - n)/[M_{14}(n-1)]$, and therefore the terms in (A1.1) take the form:

$$-\frac{1}{2}\sum\sum_{i\neq j} \breve{\Delta}_{ij}\breve{y}_i^2 I_i I_j = -\frac{1}{2}\sum_i I_i \breve{y}_i^2 \sum_{j\neq i} \breve{\Delta}_{ij} I_j = \frac{M_{14} - n}{2M_{14}(n-1)} \sum_i I_i \breve{y}_i^2 (n - I_i) ;$$

$$\sum\sum_{i\neq j} \breve{\Delta}_{ij}\breve{y}_i\breve{y}_j I_i I_j = -\frac{M_{14} - n}{M_{14}(n-1)} \sum\sum_{i\neq j} \breve{y}_i\breve{y}_j I_i I_j = -\frac{M_{14} - n}{M_{14}(n-1)} (\hat{t}_y^2 - \sum_i I_i^2 \breve{y}_i^2) .$$

Inserting these terms into (A1.2) we have

$$\hat{V}(\hat{t}_y) = \frac{M_{14} - n}{M_{14}(n-1)} (n\sum I_i \breve{y}_i^2 - \hat{t}_y^2) . \qquad (A1.3)$$

The alternative form of (A1.3) is

$$\hat{V}(\hat{t}_y) = \frac{M_{14} - n}{M_{14}} \frac{1}{n(n-1)} \sum (\frac{y_i}{p_i} - \hat{t}_y)^2 I_i . \qquad (A1.4)$$

The form (A1.4), apart from the constant $(M_{14} - n)/M_{14}$, is known for unequal probability with-replacement designs (see Sarndal et al, 1992, p. 51-52). The constant comes from the hypergeometric feature of the sampling design.

In case of the ratio $\hat{R}$ we have its SYG variance estimator in (4.5). It can be analogously simplified under the hypergeometric design. The result can be directly written out from (A1.3) or (A1.4) when replacing y_i by $y_i - \hat{R}x_i$. Then $\hat{t}_y$ relaces by

$$\sum I_i (y_i - \hat{R}x_i)/E(I_i) = 0 ,$$

and the SYG variance estimator for the ratio has the form

$$\hat{V}(\hat{R}) = \frac{M_{14} - n}{M_{14}} \frac{1}{n(n-1)} \frac{1}{\hat{t}_x^2} \sum (\frac{y_i - \hat{R}x_i}{p_i})^2 I_i .$$

Appendix 2. Generation of the population for simulation study

```
*=================================================================
==  The program inputs the data from the Estonian HBS 2003 survey,  ==
==  adds new variables used in the simulation and selection         ==
==  procedures and generates the representative pseudo-population.  ==
==                                                                  ==
==  INPUT FILES: DBF files PEREPILT, PEREISIK, HINCOME              ==
==  RESULTING FILE: SAS-datafile POP                                ==
==  LOCATION: 'F:\ESA\EUSILC\LAEKEN'                                ==
==                                                                  ==
==  AUTHOR: Elsa Leiten                                             ==
==  February, 2005.                                                 ==
=================================================================
;
libname esa 'F:\ESA\EUSILC\LAEKEN';

*========= INPUT DATAFILES - PEREISIK, PEREPILT, HINCOME =========;

proc access dbms=dbf;
  create work.perepilt.access;
  path='F:\ESA\EUSILC\LAEKEN\perepilt';
  assign=y;
  create work.perepilt.view;
  select all;
run;
proc access dbms=dbf;
  create work.perei.access;
  path='F:\ESA\EUSILC\LAEKEN\pereisik';
  assign=y;
  create work.perei.view;
  select all;
run;
proc access dbms=dbf;
  create work.income.access;
  path='F:\ESA\EUSILC\LAEKEN\hincome';
  assign=y;
  create work.income.view;
  select all;
run;

*** HOUSEHOLD PICTURE ;
data perep;
  set perepilt;
  keep leibkond a3 sisskaal lktyyp HH_COMP1 HH_COMP2 HH_COMP3 HH_COMP4 HH_SOC
maakond;
run;

*** PERSONAL INFORMATION ;
data pereisik;
  set perei;
  keep leibkond liige b4 b5 b20 b24 b25;
run;

*** INCOME DATA ;
```

```
data hincome;
  set income;
  keep leibkond s03 s04;
run;

proc sort data=perep; by leibkond; run;
proc sort data=pereisik; by leibkond; run;
proc sort  data=hincome; by leibkond; run;

*============ NEW VARIABLES ===============;
*** STRATA ;
data perep;
  set perep;
  if maakond in (1,37,44,59,67,78) then kiht=1;
  if maakond in (49,51,57,65,70,74,82,84,86) then kiht=2;
  if maakond in (39) then kiht=3;
run;

*** HH information added to the personal information ;
proc sql;
create table pereisik as
  select a.*, b.a3
  from pereisik a
  left join perep b
  on a.leibkond=b.leibkond;
run;

*** Calculation of number of HH members aged 14 and over ;
data pereisik;
  set pereisik;
  vanus = floor((intck('month',b5,a3)-(day(a3) < day(b5))) / 12);
  if vanus<0 then vanus=0;
  if vanus<14 then suur=0; else suur=1;
  if vanus<14 then laps=1; else laps=0;
  drop a3 b5;
run;

*** Number of children, adults and HH size by households ;
proc sql;
  create table arvud as
  select leibkond, sum(laps) as lapsi, sum(suur) as suuri, count(*) as suurus
  from pereisik
  group by leibkond;
quit;

*** Equivalized household size;
data arvud;
  set arvud;
  eqsize=1+0.5*(suuri-1)+0.3*lapsi;
  if suurus<5 then sklass=suurus;
      else sklass=5;
run;

*** Equivalized income data to the common income data;
```

```
data sisse;
merge income arvud;
eqinc1=s04/eqsize;
eqinc2=s03/eqsize;
run;

*============= SAMPLING PROCEDURE ============;

*** Household data with HH size information;
proc sql;
  create table leibkond as
  select a.leibkond, a.kiht, b.suurus, b.sklass
  from perep a
  left join arvud b
on a.leibkond=b.leibkond;

proc sort data=leibkond; by kiht; run;

title 'Sampling on strata information;
proc surveyselect
      data=leibkond
      out=valim1
      method=srs
      samprate=(1, 0.485714, 0.170373)
      outsize
      stats
      seed=923000;
strata kiht;
run;

proc sort data=valim1; by sklass; run;

title 'Sampling on household size';
proc surveyselect
      data=valim1
      out=valim2
      method=srs
      samprate=(1, 0.5, 0.33333, 0.25, 0.2)
      outsize
      stats
      seed=9998989;
strata sklass;
run;

*** The sampled households and persons;
data fail;
merge pereisik perep sisse;
by leibkond;
drop lapsi suuri suur laps a3 s03 s04 eqsize;
run;

* == THE FINAL POPULATION WITH NEEDED VARIABLES FOR SIMULATION =====;
proc sql;
  create table esa.pop as
```

```
      select  LEIBKOND, LIIGE, B4, VANUS, B20, B24, B25, MAAKOND, SUURUS,
            LKTYYP, HH_COMP1, HH_COMP2, HH_COMP3,HH_COMP4, HH_SOC, EQINC1,
            EQINC2, SISSKAAL
  from fail
   where leibkond in
      (select leibkond from valim2);
```

/* Other programmes for managing population data (Elsa Leiten, Imbi Traat) */

```
proc sql;*aggregates individual data for hh data;
create table esa.leibkond as select
    leibkond, mean(maakond)as maakond, mean(kiht) as kiht,
      max(liige) as suurus, sum(eqinc2) as eqinc2,
      sum(sisskaal) as sisskaal
from esa.andmestik1
group by leibkond;

/*excluding 0-records by eqinc2*/
data esa.hhpopul;
      set esa.leibkond;
      if eqinc2 > 0 then output;
run;

/* creating personal eqinc connected to hh */
data esa.hhpopul;
      set esa.hhpopul;
      eqinc1k=eqinc2;
      eqinc2=eqinc2/suurus;
run;
/* Population of persons */
data esa.ykisik1;
      set esa.andmestik1;
      if eqinc2 > 0 then output;
run;

/* Some population calculations */
data esa.ykisik1;
      set esa.ykisik1;
      if eqinc2>1800 then pov=0; else pov=1; *at-risk-of-poverty;
run;

/* Creating Frame */
data esa.frame1;
      set esa.ykisik1;
      if vanus >13 then output;
run;
proc summary data=esa.frame1 nway;
      class leibkond;
      output out=esa.frameleib;
run;
data esa.frame2;
      merge esa.frame1 esa.frameleib;
      by leibkond;
run;

/*making size14 variable in hhpopul etc.*/
data hhpopul;
      merge esa.hhpopul esa.frameleib;
      by hhcode;
```

```
run;
data esa.hhpopul;
      set esa.hhpopul;
      p=size14/2332;
      w=1/(100*p);
run;

data hhpopul;
      set esa.hhpopul;
      if eqinc2>1800 then y=0;else y=1;  *poverty group indicator;
      pv=size*y;
run;

*adding household selection probabilities and weights to each hh member;
proc sql;
      create table esa.personpop as
      select a.*, b.p, b.w, 1 as w1
      from esa.personpop a left join esa.hhpopul b
      on a.hhcode=b.hhcode;
      quit;
```

Appendix 3. Simulation programmes

A3.1 Simulations for median, poverty threshold, at-risk-of-poverty rate and quintile share ratio (Imbi Traat)

```
/*Hypergeometric sampling of hh's through SI-sampling of persons.
Calculating sample-based Laeken indicators like ratios and quantiles and their
variance estimators. Simulation results are saved into SAS-files where they are
later analyzed in SAS Insight*/

proc iml;
start;
use esa.frame2;
read all var {hhcode} into hhc;
nn=nrow(hhc);
use esa.hhpopul;
read all var {hhcode,size,eqinc2,eqinchh,p,w} into Y;
nnhh=nrow(Y);
M14=2332;
n=100;
fpc=(M14-n)/M14;
YS=j(n,ncol(Y),0);
cover=0;*coverage of conf.limits;
coverI1e=0;
coveralt=0;
coverI1=0;
*print nn;
ll=4000;
med=j(ll,1,0);*space reservations for simulated values;
I1e=med;
I1=med;
I1fix=med;
ST1=med;*st.error of I1;
ST2=med;*st.error of estimated distr.f. around 1/2;
STmed=med;*st.error of median;
STI1e=med;*st.error of I1e;
STI1ealt=med;
do l=1 to ll;*simuleerimine;
call SI(nn,n,iv);
tiv=iv[+];*print tiv;
hh=hhc[loc(iv=1)];
/* selecting sampled rows from hh population */
do i=1 to n;
abi=hh[i];
j=loc(Y[,1]=abi);
YS[i,]=Y[j,];
end;

/* median calculation*/
b=YS;
r=rank(YS[,3]);
YS[r,]=b;*sorting;
*proov1=b[1:5,];
*print proov1;
*proov2=YS[1:5,];
*print proov2;
*print r;
Mhat=sum(YS[,2]#YS[,6]);
call quantile(YS,0.5,quant);
med[l]=quant;

*print Mhat med;
I1e[l]=0.6#med[l];
Y1=1*(YS[,3]<=I1e[l]);
```

```
I1[l]=sum(YS[,2]#YS[,6]#Y1)/Mhat;
Y1fix=1*(YS[,3]<=1800);
I1fix[l]=sum(YS[,2]#YS[,6]#Y1fix)/Mhat;
Z1=1*(YS[,3]<=med[l]);
ST1[l]=sqrt(fpc#sum((YS[,2]/YS[,5])##2#(Y1-I1[l])##2)/n/(n-1))/Mhat;
ST2[l]=sqrt(fpc#sum((YS[,2]/YS[,5])##2#(Z1-0.5)##2)/n/(n-1))/Mhat;
/*Confidence limits for median with inversion method*/
c2l=0.5-1.96#ST2[l];
c2u=0.5+1.96#ST2[l];
call quantile(YS,c2l,quant);
medl=quant;
call quantile(YS,c2u,quant);
medu=quant;
if 3000>medl & 3000<medu then cover=cover+1;
STmed[l]=(medu-medl)/2/1.96;
/*Confidence limits for I1e with inversion method*/
cl=I1[l]-1.96#ST1[l];
cu=I1[l]+1.96#ST1[l];
call quantile(YS,cl,quant);
I1el=quant;
call quantile(YS,cu,quant);
I1eu=quant;
if 1800>I1el & 1800<I1eu then coverI1e=coverI1e+1;
STI1e[l]=(I1eu-I1el)/2/1.96;
STI1ealt[l]=0.6#STmed[l];
if I1e[l]-1.96#STI1ealt[l]<1800 & I1e[l]+1.96#STI1ealt[l]>1800 then
coveralt=coveralt+1;
if I1[l]-1.96#ST1[l]<0.1534 & I1[l}+1.96#ST1[l]>0.1534 then coverI1=CoverI1+1;
end;
*print med;
cover=cover/ll;
coverI1e=coverI1e/ll;
coveralt=coveralt/ll;
coverI1=coverI1/ll;
print cover coverI1e coveralt coverI1;
result=med||I1e||I1||I1fix||STmed||ST2||STI1e||ST1||STI1ealt;
*print result;
create sasresult from result
      [colname={'med' 'I1e' 'I1' 'I1fix' 'STmed' 'STfunc' 'STI1e' 'ST1'
'STI1ealt'}];
append from result;
finish;
start SI(nn,n,iv);
iv=j(nn,1,0);
n1=nn+1;
sumk=0;
do i=1 to nn;
pr=(n-sumk)/(n1-i);
eps=uniform(0);
if eps<pr then do; iv[i]=1; sumk=sumk+1;end;
end;
finish;
start quantile(YS,alpha,quant); *quantile calculation;
n=nrow(YS);
M1=YS[+,2];*no of people in sample;
YS1=j(M1,2,0);*for expanded weights and y;
k=0;
do i=1 to n;
do j=1 to YS[i,2];*expanding hh size times;
k=k+1;
YS1[k,1]=YS[i,6];*expanding weights;
YS1[k,2]=YS[i,3];*expanding y;
end;
end;
```

```
s=0;
i=1;
Mhat=sum(YS[,2]#YS[,6]);
abi=alpha#Mhat;
do while (s<abi);s=s+YS1[i,1];i=i+1;end;
if (s=abi) then quant=(YS1[i-1,2]+YS1[i,2])/2;else quant=YS1[i,2];
finish;
/*
start quintile(YS,alpha,quant); *alternative method for quantile calculation;
n=nrow(YS);
s=0;
i=1;
Mhat=sum(YS[,2]#YS[,6]);
abi=alpha#Mhat;
do while (s<abi);s=s+YS[i,2]#YS[i,6];i=i+1;end;
if (s=abi) then quant=(YS[i-1,3]+YS[i,3])/2;else quant=YS[i-1,3];
finish;
*/
run;quit;
```

A3.2 Simulations for Gini

```
*Macros GINI, SIMU, JACK, JACK_REP;

* ------------------------------------------------------------------
Macro GINI calculates a single Gini coefficient based on variable y in dataset
datain.
By default calculations are made on the previously generated dataset.
   input:   y        variable, on which Gini is calculate
            weight   weight variable (by default=1)
            datain   dataset containing these variables
                     (default = _last_)
            K        row number (by default=1),
                     i.e the number of simulation
   output:    the values of the gini coefficient G, the mean value
              of y  - Y, the weighted and unweighted number of cases
              (N, WN) on which it is based are put in the output dataset SUMS.
   author:    Elsa Leiten
  ------------------------------------------------------------------;
%macro gini(y=,weight=1,datain=_last_, k=1);

proc sort data=&datain;
by &y;

/* cumulative sums, their combinations from the formula of Gini */
data uus1;
set &datain;
      cweight+&weight;
      var0 = &weight;
      var1 = &weight*&y;
      var2 = &weight*&weight*&y;
      var3 = cweight*&weight*&y;
run;

/* sums of cumulative variables */
proc means data=uus1 sum mean n noprint;
      output out=sums
            sum(var0 var1 var2 var3)=sum0 sum1 sum2 sum3
            mean(&y)=Y;
run;

/*calculation of Gini based on calculated sums */
data sums;
```

```
set sums;
      nr=&k;
      G = (2*sum3 - sum2) / (sum0 * sum1) - 1;
      N=_freq_;
      WN=sum0;
format      N 4.0
            WN 8.0
            Y 4.0;
keep nr G Y N WN;
run;
%mend gini;

* ------------------------------------------------------------------
Macro SIMU draws samples from the populations i number of times
(by default 1000 times), according to selected method (default=SRS). Sample size is
n (by default 100).
Gini calculated each time by the macro GINI.
      input:  i      number of simulations (default number is 1000)
              m      sample size (by default n=100)
              kaal   weight variable (by default w)
             frame   dataset containing the frame
                     (by default frame of persons aged 4+)
             datain  dataset containing the data on personal level
                     (by default esa.personpop)
             met     selection method (default=SRS)
             pps     indicator of usage of pps sampling (default=0)
             suurus  size variable (by default=size14)
      output:     For each simulation:
                  the values of the gini coefficient G, the mean
                  value of y (Y) and the weighted and unweighted
                  number of cases (N, WN) into the output dataset
                  RESULTS.
      author:    Elsa Leiten
  ------------------------------------------------------------------;

%macro simu (i=1000, m=100,frame=esa.frame2, kaal=w,hhdata=personpop,
met=srs,pps=0,suurus=size14);

*simulation loop;
%DO k=1 %to &i;

      /*sample selection, without size variable*/
      %if &pps=0 %then %do;
      proc surveyselect noprint
            data=&frame
            out=sample
            method=&met
            sampsize=&m;
      run;
      %end;

      *proportional to size sampling;
      %if &pps=1 %then %do;
      proc surveyselect noprint
            data=&frame
            out=sample
            method=&met
            sampsize=&m;
            size &suurus;
      run;
      %end;

      *the data of the selected households;
      proc sql;
```

```
            create table hhsample as
            select * from &hhdata
            where hhcode in
                  (select hhcode from sample);

      *calculating Gini;
      %gini(y=eqinc2,weight=&kaal,datain=hhsample, k=&K);

      *adding the results to the results dataset;
      %if &k=1 %then %do;
            data results;
            set sums;
            run;
      %end;
      %else %do;
            data results;
            set results sums;
            run;
      %end;
%END;
%mend simu;

* ------------------------------------------------------------------
Macro JACK selects samples from a sample of size n leaving out one household at a
time. Gini calculated each time by the macro GINI.
      input:   n        sample size (by default n=100)
               frame    name of the sampling frame (default=sample)
               hdata2   dataset with sample data on personal level
                        (default=hhsample)
      output: For each replication:
                  the values of the gini coefficient G, the mean
                  value of y (Y) and the weighted and unweighted
                  number of cases (N, WN) into the output dataset
                  RESULTS.
      author:    Elsa Leiten
  ------------------------------------------------------------------;

%macro jack(n=100,frame2=sample,hhdata2=hhsample);

/*new weights compensating on the one missing household */
data &hhdata2;
      set &hhdata2;
      w2=w*100/99;
run;

*simulation loop;
%DO k=1 %to &n;
      data sample2;
      set &frame2;
      where nr<>&k;
      run;

      *the data of the selected households;
      proc sql;
            create table hhsample2 as
            select * from &hhdata2
            where hhcode in
                  (select hhcode from sample2);

      *calculating Gini;
      %gini(y=eqinc2,weight=w2,datain=hhsample2,k=&K);

      *adding the results to the results dataset;
      %if &k=1 %then %do;
```

```
            data results;
            set sums;
            run;
      %end;
      %else %do;
            data results;
            set results sums;
            run;
      %end;
%END;
%mend jack;

* ------------------------------------------------------------------
Macro JACK_REP: Selecting Jackknife samples (on SRS design) from the
populations i number of times (by default 1000 times). Sample size is n (by default
100).
For each sample n replicates are formed with size n-1 leaving one household out each
time.
Gini and its Standard deviation calculated each time by the macro GINI.
Averages on different simulations are then calculated.

      input:   I      number of simulations (default number is 1000)
               n      sample size (by default n=100)
               frame0  dataset containing the frame
                      (default=esa.frame2)
               hhdata0 dataset containing the personal data
                       (default=esa.personpop)
      output: the estimates of the gini coefficient G, the mean of
              y (Y) and their variances are output into the output
              dataset RESULTS2 (calculated on 1000 samples and their
              replicates).
      author:    Elsa Leiten
  ------------------------------------------------------------------;

%macro jack_rep (i=1000, n=100,frame0=esa.frame2, hhdata0=esa.personpop);

*simulation loop;
%DO l=1 %to &i;

      *generating a sample of 100 households for Jackknife
       simulations;
      %simu(i=1,m=100,frame=&frame0,hhdata=&hhdata0,met=srs,kaal=w);
      * generating:
            sample = set of selected households
            hhsample = household members of selected households;

      data sample;
            set sample;
            nr+1;
      run;

*replications of the sample with one household missing every time;
      %jack(n=100,frame2=sample,hhdata2=hhsample);
      * generating:
      results = dataset containing 100 Ginis calculated on
                replications;

      *calculating estimates on Gini and its variance on these
      replications of one sample;
      proc means data=results noprint;
      var g y n;
      output out=jack
            mean(g y n)=G Y N
            std(g y n)=G_std Y_std N_std
```

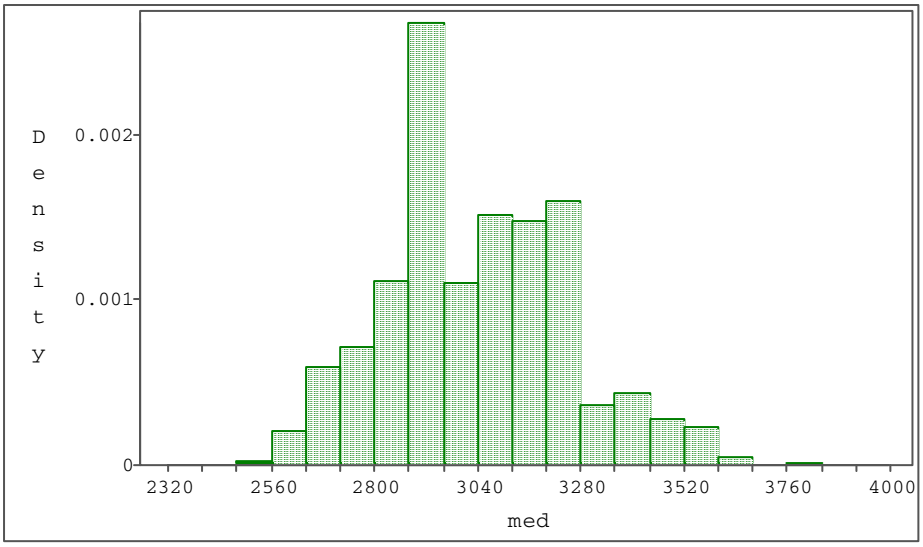
```
            N(g y n)=G_n Y_n N_n;
      run;

      data jack;
            set jack;
            valim=&l;
      run;

      *adding the results to the results dataset;
      %if &l=1 %then %do;
            data results2;
            set jack;
            run;
      %end;
      %else %do;
            data results2;
            set results2 jack;
            run;
      %end;
%END;
%mend jack_rep;


*============= TRUE VALUES IN POPULATION ================;

%gini(y=eqinc2,weight=w1,datain=esa.personpop);

proc print data=sums;
      title "--- Gini, Mean, Cases, Weighted Sum of Cases ---";
      title2 'True population values';
      var G Y N WN;
run;

*======== EMPIRICAL DISTRIBUTION OF GINI ================;

*selecting address persons;
%simu(i=1000,m=100,frame=esa.frame2,hhdata=personpop, kaal=w);

data esa.simu_gdistr;
      set results;
run;

*Table of distribution characteristics;
proc means data=esa.simu_gdistr;
      var g y n;
run;

title 'Empirical distribution of Gini';
proc univariate data=esa.simu_gdistr noprint;
      histogram g / cfill=green ;
run;

*========= JACKKNIFE  - repeated 1000 times ==============;

%jack_rep (i=100, n=100,frame0=esa.frame2,hhdata0=esa.personpop);
data esa.sample;
      set sample;
run;
data esa.hhsample;
      set hhsample;
run;

data esa.simu_jackknife;
      set results2;
```

```
run;

proc means data=esa.simu_jackknife;
      var g y g_std y_std;
run;

title 'Simulation of Gini index';
title2 'Jackknife';
proc univariate data=esa.simu_jackknife noprint;
      histogram g/ cfill=green ;
run;

*========= JACKKNIFE  - 1 sample ===============;

*selecting one sample;
%simu(i=1,m=100,frame=esa.frame2,hhdata=esa.personpop,met=srs,kaal=w);
data sample;
      set sample;
      nr+1;
run;

*replications;
proc means data=results;
      var g y n;
run;

title 'Jackknife (1 sample of size 100 with 100 replicates)';
proc univariate data=results noprint;
      histogram g / cfill=green ;
run;
```

Appendix 4. Simulation results

A4.1. Results for median

The sampling distribution of median is received by calculating it with (3.2) in 4000 rounds. Sample is taken as described in 5.2.

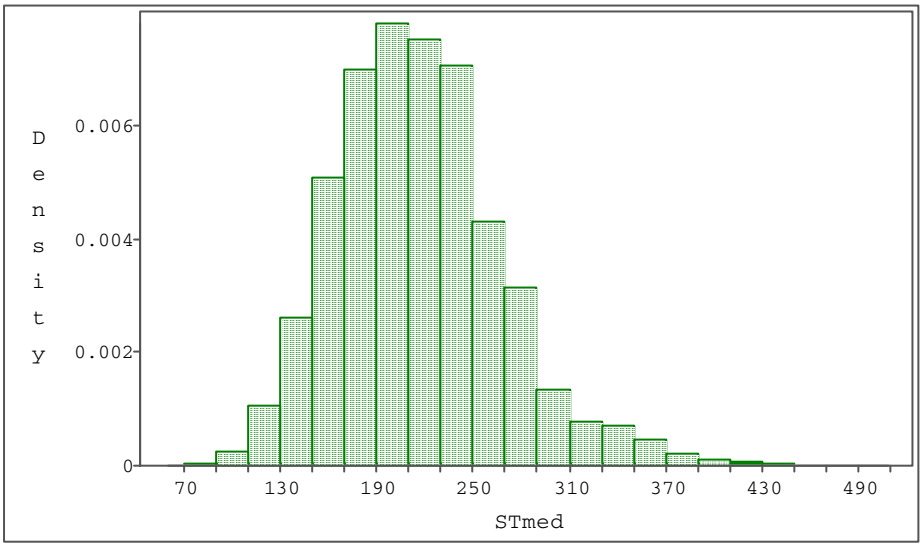


Moments			
N	4000.0000	Sum Wgts	4000.0000
Mean	3036.2988	Sum	12145195.2
Std Dev	216.4820	Variance	46864.4368
Skewness	0.3328	Kurtosis	0.0398
USS	3.706E+10	CSS	187410883
CV	7.1298	Std Mean	3.4229

Quantiles			
100% Max	3956.0000	99.0%	3572.5000
75% Q3	3186.4286	97.5%	3513.8900
50% Med	3021.7857	95.0%	3425.9400
25% Q1	2893.3509	90.0%	3313.1870
0% Min	2350.0000	10.0%	2757.5000
Range	1606.0000	5.0%	2682.0500
Q3-Q1	293.0777	2.5%	2659.0000
Mode	2947.6974	1.0%	2617.7333

STmed — standard deviation of the median

Variance of the median is estimated as described in Section 4.3. Standard deviation is square root of it. Sampling distribution of STmed demonstrates performance of the variance estimator derived by us.

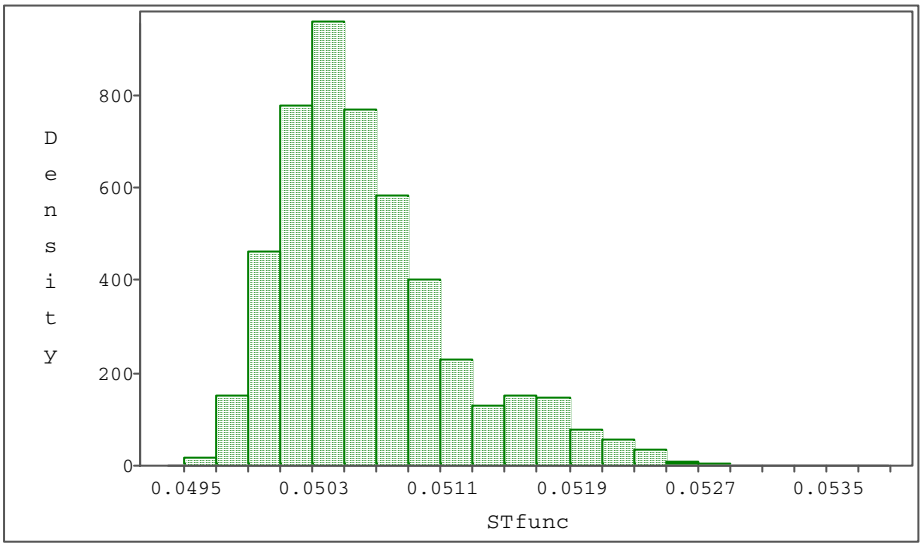


Moments			
N	4000.0000	Sum Wgts	4000.0000
Mean	216.6058	Sum	866423.186
Std Dev	52.3429	Variance	2739.7833
Skewness	0.6809	Kurtosis	1.0635
USS	198628678	CSS	10956393.4
CV	24.1651	Std Mean	0.8276

Quantiles			
100%	Max	502.7342	99.0%	369.5085
75%	Q3	246.4711	97.5%	340.8505
50%	Med	213.0869	95.0%	309.8036
25%	Q1	179.3804	90.0%	281.7177
0%	Min	73.9191	10.0%	153.6896
	Range	428.8151	5.0%	140.9143
	Q3-Q1	67.0907	2.5%	128.2822
	Mode	219.7279	1.0%	115.9862

STfunc — standard deviation of the distribution function value at the median

This is used for constructing a confidence interval around distribution function value ½. Inverting the endpoints of this interval, gives confidence limits for median. Small and stable value of STfunc is a crucial thing for narrow confidence limits of the median.

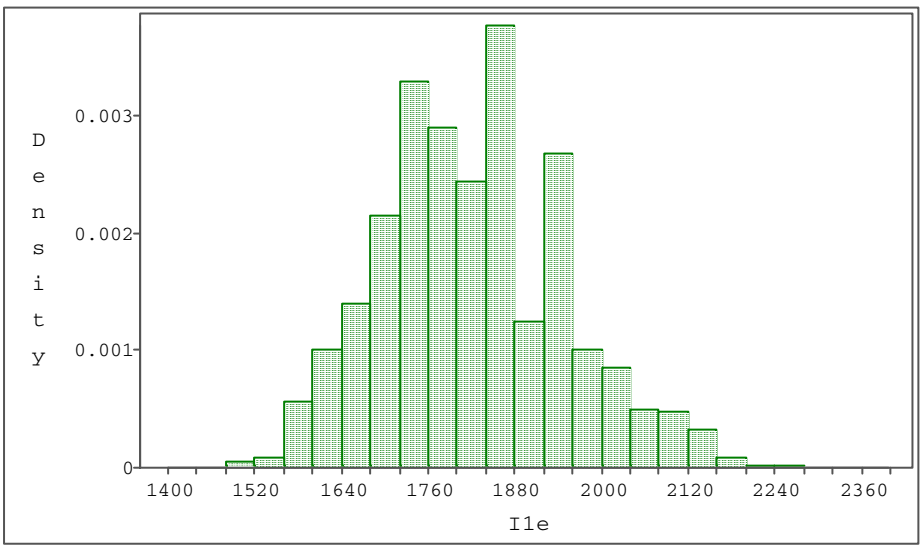


Moments			
N	4000.0000	Sum Wgts	4000.0000
Mean	0.0507	Sum	202.6019
Std Dev	0.0006	Variance	3.200E-07
Skewness	1.1593	Kurtosis	1.5892
USS	10.2632	CSS	0.0013
CV	1.1169	Std Mean	8.945E-06

Quantiles			
100% Max	0.0538	99.0%	0.0523
75% Q3	0.0509	97.5%	0.0521
50% Med	0.0505	95.0%	0.0518
25% Q1	0.0503	90.0%	0.0515
0% Min	0.0496	10.0%	0.0501
Range	0.0042	5.0%	0.0499
Q3-Q1	0.0006	2.5%	0.0499
Mode	0.0500	1.0%	0.0498

A4.2. Simulation results for poverty threshold I_{1e}

Poverty threshold $\hat{I}_{1e}$ is calculated as 0.6 of the estimated median. The sampling distribution of it is given below

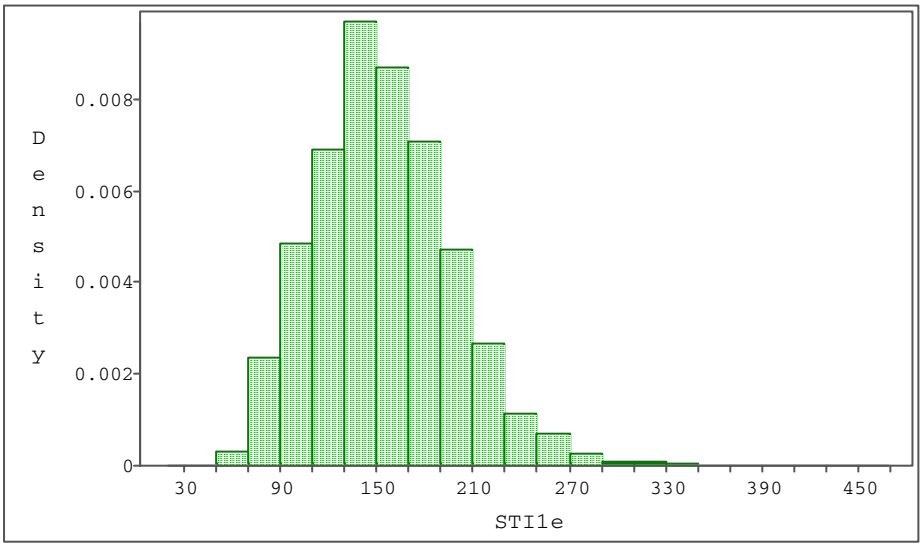


Moments			
N	4000.0000	Sum Wgts	4000.0000
Mean	1821.7793	Sum	7287117.11
Std Dev	129.8892	Variance	16871.1973
Skewness	0.3328	Kurtosis	0.0398
USS	1.334E+10	CSS	67467917.8
CV	7.1298	Std Mean	2.0537

Quantiles			
100% Max	2373.6000	99.0%	2143.5000
75% Q3	1911.8571	97.5%	2108.3340
50% Med	1813.0714	95.0%	2055.5640
25% Q1	1736.0105	90.0%	1987.9122
0% Min	1410.0000	10.0%	1654.5000
Range	963.6000	5.0%	1609.2300
Q3-Q1	175.8466	2.5%	1595.4000
Mode	1768.6184	1.0%	1570.6400

STI1e — standard deviation of $\hat{I}_{1e}$ with inversion method

This standard deviation is calculated with inversion method of the distribution function, see Remark 4.4. The method is biased and not so stable.

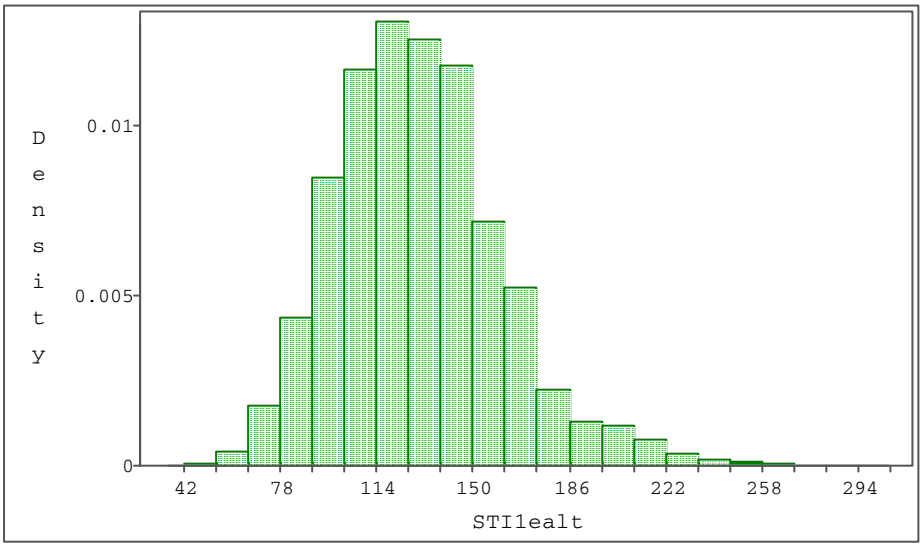


Moments			
N	4000.0000	Sum Wgts	4000.0000
Mean	155.1200	Sum	620480.060
Std Dev	44.5369	Variance	1983.5387
Skewness	0.7866	Kurtosis	2.2080
USS	104181047	CSS	7932171.23
CV	28.7113	Std Mean	0.7042

Quantiles			
100% Max	467.8567	99.0%	274.7307
75% Q3	181.6368	97.5%	251.5995
50% Med	151.6220	95.0%	229.9745
25% Q1	124.4534	90.0%	211.1745
0% Min	35.2041	10.0%	101.1905
Range	432.6526	5.0%	88.5459
Q3-Q1	57.1834	2.5%	79.9754
Mode	132.6464	1.0%	72.3302

STI1alt — standard deviation of $\hat{I}_{1e}$ with simple formula

Here Standard deviation of $\hat{I}_{1e}$ is calculated as square root of the formula (4.17). Corresponding sampling distribution is below.

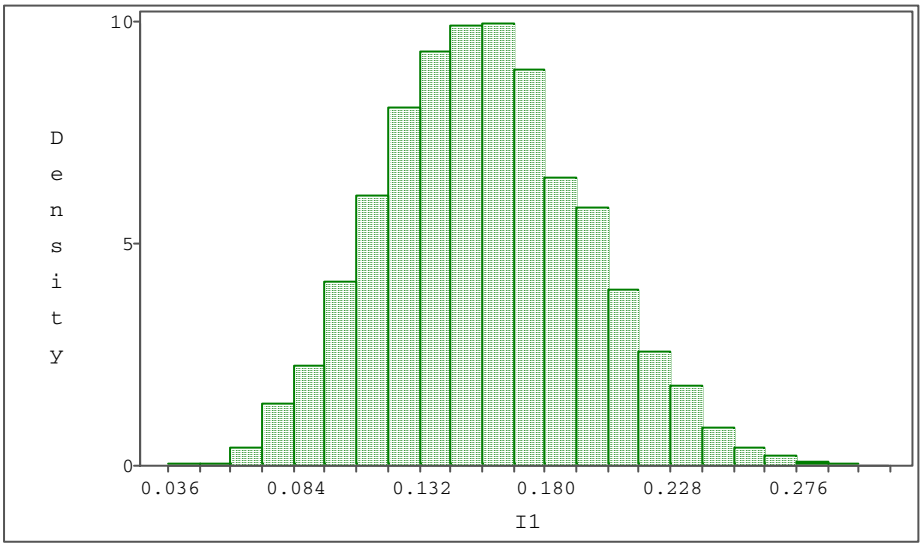


Moments			
N	4000.0000	Sum Wgts	4000.0000
Mean	129.9635	Sum	519853.912
Std Dev	31.4058	Variance	986.3220
Skewness	0.6809	Kurtosis	1.0635
USS	71506324.1	CSS	3944301.64
CV	24.1651	Std Mean	0.4966

Quantiles			
100%	Max	301.6405	99.0%	221.7051
75%	Q3	147.8827	97.5%	204.5103
50%	Med	127.8522	95.0%	185.8821
25%	Q1	107.6282	90.0%	169.0306
0%	Min	44.3515	10.0%	92.2138
	Range	257.2891	5.0%	84.5486
	Q3-Q1	40.2544	2.5%	76.9693
	Mode	131.8367	1.0%	69.5917

A4.3. Simulation results for at-risk-of-poverty rate I_1

Here the estimator $\hat{I}_1$ is calculated with sample-based poverty threshold, i.e. it is a proper Laeken indicator. Below we see its sampling distribution and true variability (which we want to estimate).

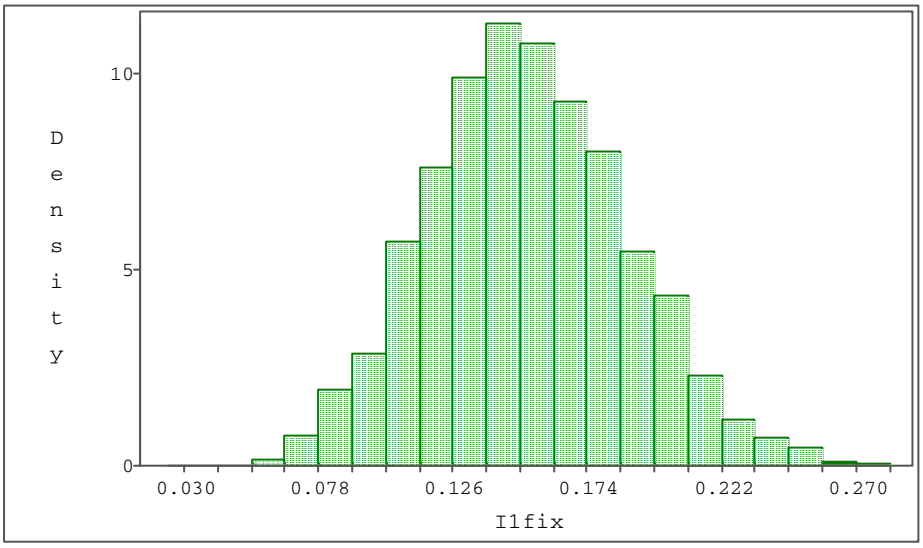


Moments			
N	4000.0000	Sum Wgts	4000.0000
Mean	0.1572	Sum	628.7142
Std Dev	0.0392	Variance	0.0015
Skewness	0.2285	Kurtosis	-0.0663
USS	104.9735	CSS	6.1531
CV	24.9562	Std Mean	0.0006

Quantiles			
100% Max	0.3036	99.0%	0.2530
75% Q3	0.1829	97.5%	0.2375
50% Med	0.1557	95.0%	0.2250
25% Q1	0.1298	90.0%	0.2085
0% Min	0.0407	10.0%	0.1077
Range	0.2629	5.0%	0.0956
Q3-Q1	0.0531	2.5%	0.0846
Mode	0.1220	1.0%	0.0745

I1fix — at-risk-of-poverty rate with fixed threshold

Here $\hat{I}_1$ is calculated with fixed poverty threshold 1800. The aim was to see how much the two distributions differ, how much the random threshold increases the variability compared to the fixed one. This estimator is theoretically known as unbiased.

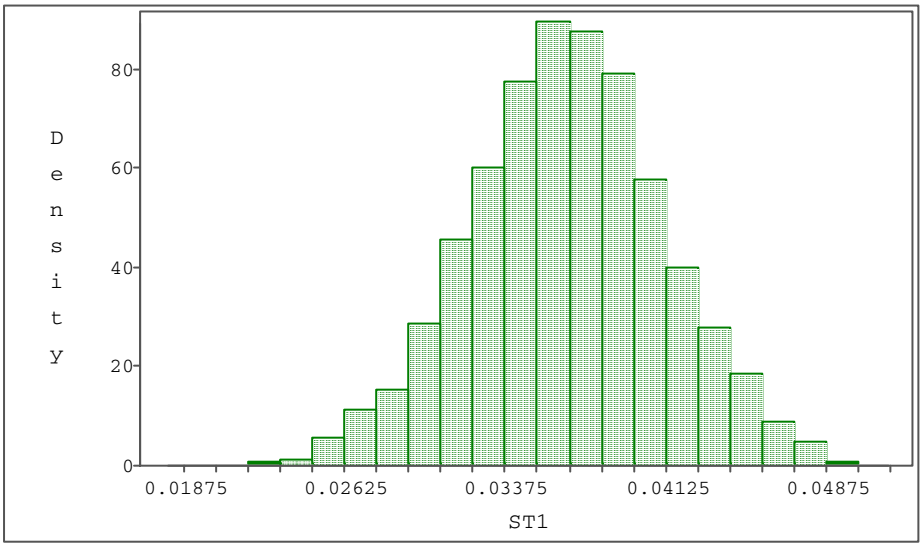


Moments			
N	4000.0000	Sum Wgts	4000.0000
Mean	0.1530	Sum	611.9123
Std Dev	0.0358	Variance	0.0013
Skewness	0.1932	Kurtosis	-0.0486
USS	98.7369	CSS	5.1278
CV	23.4078	Std Mean	0.0006

Quantiles			
100% Max	0.2813	99.0%	0.2428
75% Q3	0.1770	97.5%	0.2268
50% Med	0.1517	95.0%	0.2142
25% Q1	0.1282	90.0%	0.2004
0% Min	0.0362	10.0%	0.1078
Range	0.2452	5.0%	0.0958
Q3-Q1	0.0487	2.5%	0.0862
Mode	0.1577	1.0%	0.0757

ST1 — standard deviation of $\hat{I}_1$

Here the variance estimator (4.6a) derived by us was experimented in the simulation study. The sampling distribution of the square root of the estimator is below.



Moments			
N	4000.0000	Sum Wgts	4000.0000
Mean	0.0366	Sum	146.5177
Std Dev	0.0046	Variance	2.100E-05
Skewness	-0.0538	Kurtosis	0.0252
USS	5.4508	CSS	0.0840
CV	12.5097	Std Mean	7.245E-05

Quantiles			
100% Max	0.0514	99.0%	0.0471
75% Q3	0.0396	97.5%	0.0457
50% Med	0.0367	95.0%	0.0443
25% Q1	0.0336	90.0%	0.0426
0% Min	0.0195	10.0%	0.0309
Range	0.0319	5.0%	0.0291
Q3-Q1	0.0060	2.5%	0.0272
Mode	.	1.0%	0.0258

Coverage rates of the confidence intervals of Laeken indicators

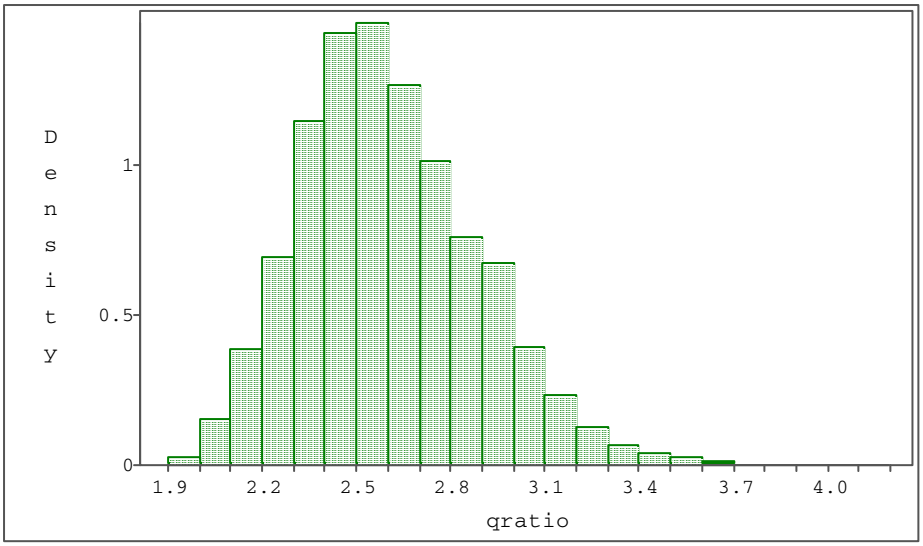
Med I_{1e} I_{1e} (alt.) I_1

0.952 0.906 0.932 0.975

Interval for the median was calculated as described in Section 4.3. Other intervals were symmetric normal-based intervals calculated with standard deviations studied above in this Appendix. Later the programme was changed to calculate median-based confidence intervals for I_{1e}. This method appeared to be most precise with coverage rate 0.952.

A4.4. Simulation results for quintile ratio

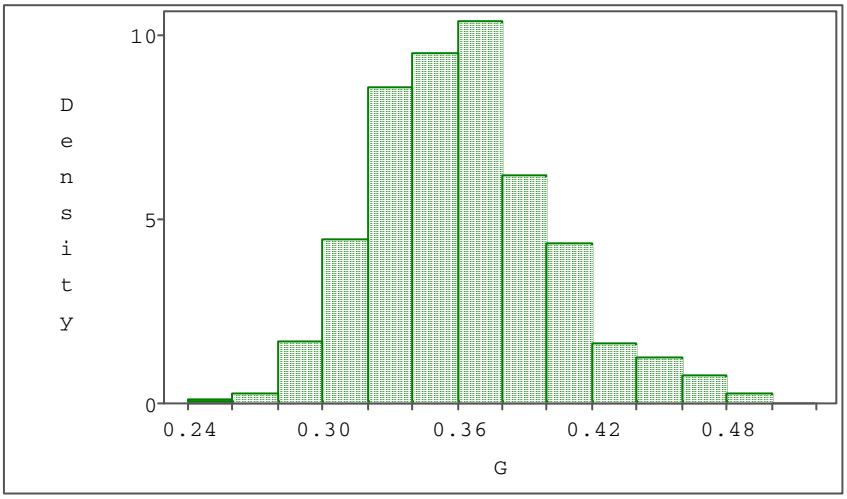
Sampling distribution of quintile share ration (3.4) is below.



Moments			
N	4000.0000	Sum Wgts	4000.0000
Mean	2.6088	Sum	10435.1742
Std Dev	0.2884	Variance	0.0832
Skewness	0.6038	Kurtosis	0.6249
USS	27555.8959	CSS	332.6806
CV	11.0560	Std Mean	0.0046

Quantiles			
100% Max	4.1525	99.0%	3.4258
75% Q3	2.7885	97.5%	3.2338
50% Med	2.5805	95.0%	3.1134
25% Q1	2.4055	90.0%	2.9905
0% Min	1.9027	10.0%	2.2660
Range	2.2498	5.0%	2.1875
Q3-Q1	0.3830	2.5%	2.1180
Mode	2.4923	1.0%	2.0637

A4.5 Simulations for Gini coefficient



Moments			
N	1000.0000	Sum Wgts	1000.0000
Mean	0.3630	Sum	363.0494
Std Dev	0.0408	Variance	0.0017
Skewness	0.4785	Kurtosis	0.4223
USS	133.4705	CSS	1.6656
CV	11.2471	Std Mean	0.0013

Quantiles			
100% Max	0.5075	99.0%	0.4765
75% Q3	0.3883	97.5%	0.4585
50% Med	0.3606	95.0%	0.4372
25% Q1	0.3345	90.0%	0.4155
0% Min	0.2454	10.0%	0.3145
Range	0.2621	5.0%	0.3042
Q3-Q1	0.0538	2.5%	0.2935
Mode	.	1.0%	0.2816

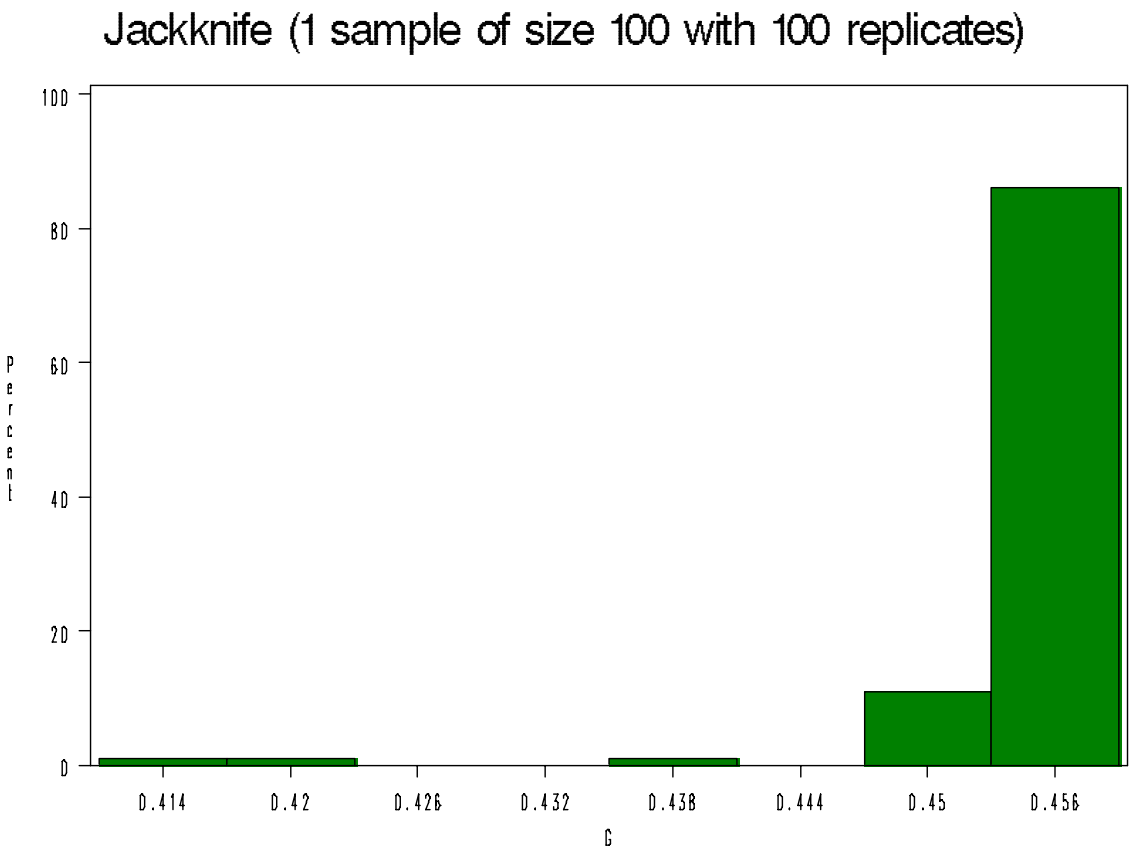
A4.6. Jackknife simulations based on 3 different samples from population

Jackknife (1st sample of size 100 with 100 replicates)

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
G	100	0.4539148	0.0058331	0.4131069	0.4576936
Y	100	4814.37	60.6720990	4496.57	4871.45
N	100	251.6300000	1.5676979	244.0000000	254.0000000

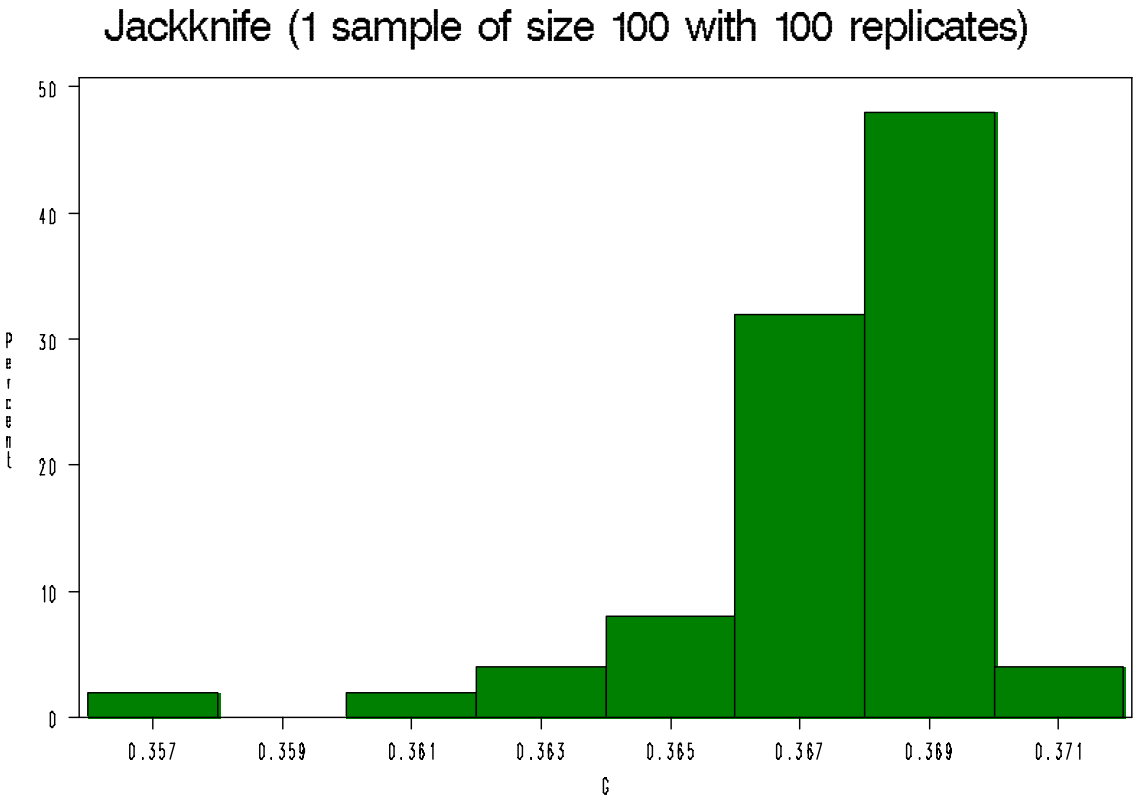
G – Gini index
Y – Equivalized income
N – sample size



Jackknife (2nd sample of size 100 with 100 replicates

The MEANS Procedure

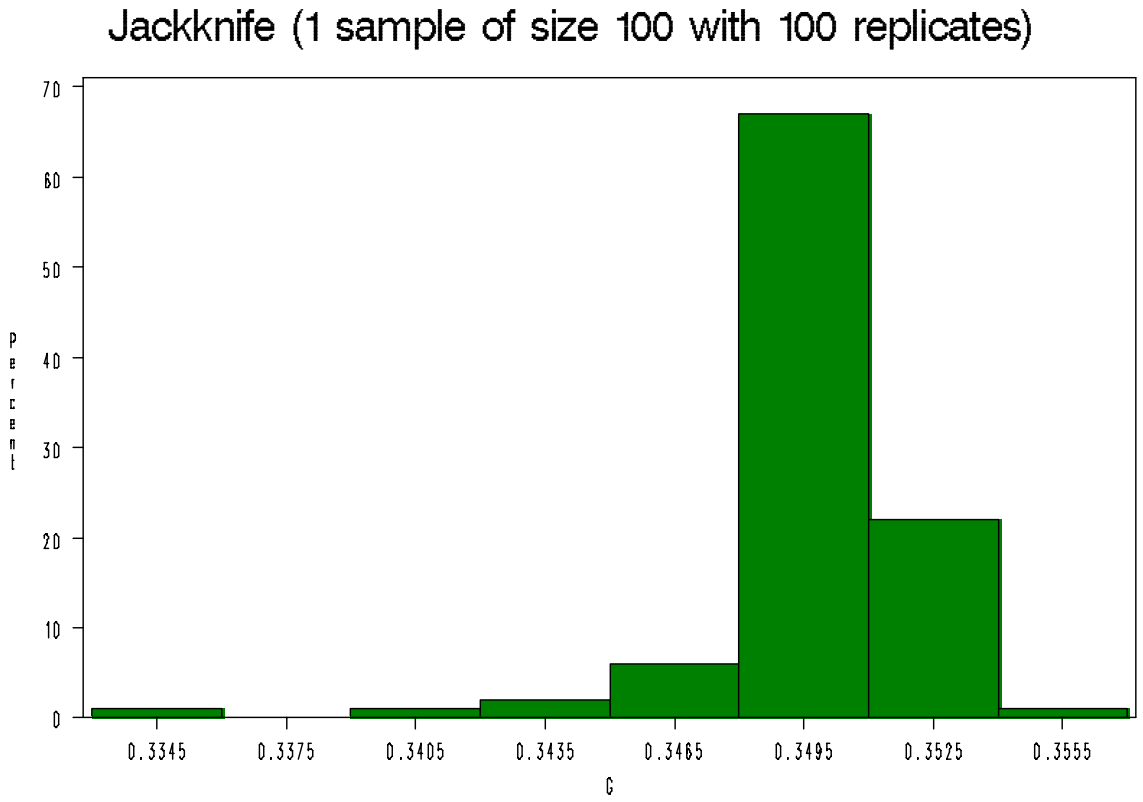
Variable	N	Mean	Std Dev	Minimum	Maximum
G	100	0.3676418	0.0025118	0.3569219	0.3710280
Y	100	4453.72	44.9802833	4154.57	4526.19
N	100	244.7400000	1.3752502	240.0000000	247.0000000



Jackknife (3rd sample of size 100 with 100 replicates

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
G	100	0.3495293	0.0023974	0.3341691	0.3548788
Y	100	4095.98	39.8288014	3762.03	4172.05
N	100	228.7600000	1.4221907	225.0000000	231.0000000


Jackknife (1 sample of size 100 with 100 replicates)
Percent
70
60
50
40
30
20
10
0
0.3345
0.3375
0.3405
0.3435
0.3465
0.3495
0.3525
0.3555
G

Appendix A4.7. The distribution of mean Gini

The distribution of mean Gini is calculated over 100 new initial samples, where the mean Gini itself is calculated over 100 Jackknife simulations in each of these initial samples.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
G	100	0.3652869	0.0460020	0.2767972	0.5005039
G_std	100	0.0040559	0.0019742	0.0018454	0.0118517