Variance Estimation for Indicators of Social Cohesion

Muennich, Ralf*
E-mail: Muennich@uni-trier.de

Zins, Stefan*
E-mail: zins@uni-trier.de

**University of Trier, Social and Economical Statistics Department*
Universitätsring 15
54296 Trier, Germany

Abstract

In March 2000 the Lisbon European Council has asked the Members States of the European Union to make steps towards to improve social cohesion and combating poverty by 2010. In order to adequately measure poverty and social cohesion, the Laeken indicators were agreed on as a set of indicators. Within the European Statistical System, additionally to the statistics accuracy measures have to be reported within quality reports. Accuracy measures, however, are mainly based on variance estimation methods.

The paper will address variance estimation methods for linear statistics, such as means and totals, as well as the non-linear Laeken indicators under a variety of sampling designs. The methodology comprises linearization and resampling methods. Special emphasis will be put on peculiarities in the data which in general lead to difficulties in the selection of methods. To empirically explore the quality of the proposed variance estimates, the results from a Monte-Carlo study will be presented which is based on a synthetic but close-to-reality dataset.

Keywords: variance estimation, non-linear statistics, design effects, small area estimation

Indroduction

The European Laeken indicators were constructed to measure poverty and social cohesion within the European Union. To ensure comparability of the Laeken indicators within Europe, the European Survey on Income and Living Conditions (EU-SILC) was launched. As a result from the open method of coordination, the SILC sampling design varies from country to country. This urges the needs of comparable accuracy measurement methods for the SILC sampling designs.

Variance estimation in survey sampling is essential for statistical inference. For indicators of poverty and social exclusion estimated from sample survey data, it gives much needed information on the accuracy of the estimators. Further, it enables the statistician to construct valid confidence intervals (CI) for the estimated indicators $\hat{\theta}$. The two main problems in this context are

(i) that due to complex survey designs (e.g. unequal probability sampling) it is not practicable to estimate $\mathrm{V}(\hat{\theta})$ directly,

(ii) calculating $\hat{\theta}$ involves the estimation of non-smooth statistics.

Hence, two different kinds of approximation methods have to be considered. First, the approximation of the variance of the statistic in question in general, regardless of the actual sampling designs, and second the approximation of the design variance of a particular statistic.

Variance estimation in complex sample surveys

First, we introduce a general framework which serves as the basis of our analysis. Let us consider the finite population $\mathcal{U}$ of N identifiable units, so that they can be represented by integers 1,2, ..., N, $\mathcal{U} = \{1, \ldots, N\}$. Now we want to draw a sample of n units from $\mathcal{U}$ by means of random sampling without replacement. The probability of inclusion for the i-th element in $\mathcal{U}$ is denoted by π_i, with $\pi_i = \sum_{s \ni i} p(s)$ and the second-order inclusion probability π_{ij}, with $\pi_{ij} = \sum_{s \ni \{i,j\}} p(s)$. We can now use the Horvitz-Thompson estimator

$$(1) \quad \hat{\tau} = \sum_{i \in s} y_i \cdot \frac{1}{\pi_i} \,,$$

to estimate the total $\tau = \sum_{i \in \mathcal{U}} y_i$. In case of fixed size designs Yates and Grundy (1953, p. 257) derived the variance of $\hat{\tau}$ from (1) as

$$(2) \quad \mathrm{V}(\hat{\tau}) = -\frac{1}{2} \sum_{i}^{N} \sum_{j \neq i}^{N} (\pi_{ij} - \pi_i \cdot \pi_j) \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = \sum_{i=1}^{N} \sum_{j<i}^{N} (\pi_i \cdot \pi_j - \pi_{ij}) \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

with the corresponding variance estimator

$$(3) \quad \hat{\mathrm{V}}(\hat{\tau}) = \sum_{i=1}^{n} \sum_{j<i}^{n} \frac{\pi_i \cdot \pi_j - \pi_{ij}}{\pi_{ij}} \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \,,$$

(cf. Cochran, 1977, p. 260f.). For (3) to be unbiased, a necessary and sufficient condition is that $\pi_{ij} > 0$, $\forall\ i,j \in \mathcal{U}$. For an overview of variance estimation for various sampling designs we refer to Wolter (2007, section 1.4).

To avoid the calculation of the double sum in (3), different approximations can be found in the literature that only use π_i but dispense with the π_{ij} (cf. Berger and Skinner, 2004). The use of such an approximation has not only computational advantages but it is also of great practical relevance because the calculation of the second-order inclusion probabilities can be avoided. A general approximation of the variance of (1) can be written as

$$(4) \quad \mathrm{V}_{\mathrm{approx}}(\hat{\tau}) = \sum_{i \in \mathcal{U}} b_i \cdot e_i^2 \,,$$

where

$$e_i = \frac{y_i}{\pi_i} - \beta = \frac{y_i}{\pi_i} - \frac{\sum_{j \in \mathcal{U}} b_j \cdot \frac{y_j}{\pi_j}}{\sum_{j \in \mathcal{U}} b_j} \,,$$

(cf. Deville and Tillé, 2005, p. 573). According to the choice of b_i there exist numerus variants of approximation (4). An analysis of the literature on different values of b_i can be found in Matei and Tillé (2005).

Variance estimation for non-linear estimators

Since the statistics involved are highly non-linear, standard variance estimation procedures cannot be applied directly. Thus, resampling methods or linearization techniques have to be used to estimate the variance.

Linearization methods approximate the non-linear estimator by a linear function. Afterwards standard variance formulae for the given design can be applied to the so-called linearized values. This indirect approach estimates the asymptotic variance of an estimator which results in biased but typically consistent variance estimators (see Wolter, 2007, chapter 6).

The bottom line of linearization is to construct a linearized variable $z_i \; \forall \; i \in s$ such that the variance of the total of this variable z is an approximation for the variance of the estimator of interest $\hat{\theta}$. Thus, its possible to write $\mathrm{V}\left(\sum_{i \in s} z_i w_i\right) \approx \mathrm{V}(\hat{\theta})$, where w_i is some survey weight associated with the i-th sampling element. In case $w_i = \pi_i^{-1}$ we can directly use estimator (3), or its approximation, to estimate the asymptotic variance of $\hat{\theta}$.

A general approach to linearization which allows also for discontinuous estimation function, as it is the case for estimators for poverty and income inequality, is the use of influence functions (see Hampel et al., 1986). The derivation of influence functions requires differentials of $\hat{\theta}$ in the sense of Gâteaux (cf. Shao, 2003, pp. 339). The values z_i are obtained by evaluating the influence function for all observations y_i in the sample. For the derivation of the influence functions of the indicators, we refer to the literature where the corresponding z_i values can be found. Further details can be found for the at-risk-of-poverty rate (ARPR) in Deville (1999) and Osier (2009), for the the quintile share ratio (QSR) in Hulliger and Münnich (2006), Osier (2009) and Langel and Tillé (2011), for the Gini coefficient (GINI) in Kovačević and Binder (1997), and for the relative median poverty gap (RMPG) in Osier (2009).

Another category of approaches to variance estimation are resampling methods. Their general characteristic is to draw (sub-)samples from a given population or the original sample and to estimate the population parameter of interest from each sample. The variance estimation based on this repeated estimates which aims at building the estimation distribution by the resamples. In this study, we use balanced repeated replication (BRR) and the bootstrap. The classical jackknife method suffers from the non-smooth statistics we use.

For details of the resampling methods, we refer to Shao et al. (1998), Davison and Sardy (2004), or Münnich (2008). Extensions for without replacement bootstrap can be drawn from Rao and Wu (1988) or Sitter (1992a).

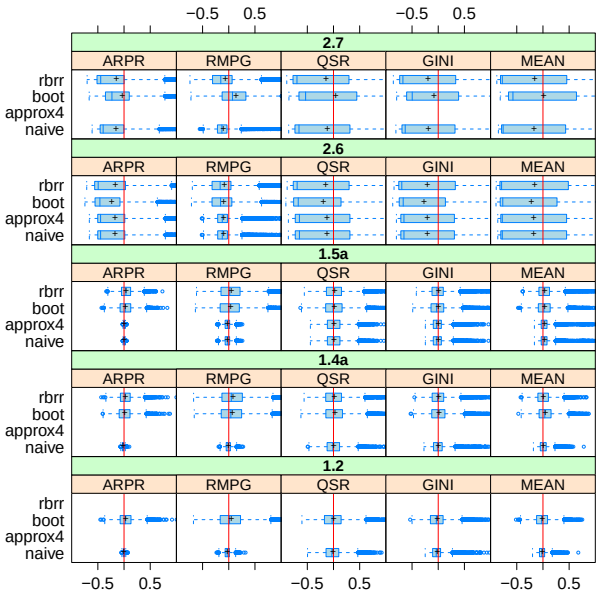

Figure 1: Relative Bias of Variance Estimates

The simulation study

The aim of the study is to evaluate the variance estimation methods under different sampling designs which are of particular interest for SILC. The study includes three one-stage designs and two

two-stage designs as follows:

1.2 Simple random sampling, sampling Units: HH

1.4a Stratified sampling, sampling Units: HH, Stratification: NUTS 2.

1.5a Stratified PPS-Sampling, Sampling Units: HH, Stratification: NUTS 2, Size Measure: Household Size.

2.6 Two-Stage Stratified PPS-Sampling, Sampling Units: Municipality (First Stage) and HH (Second Stage), Stratification (First Stage): NUTS 2 and Degree of Urbanisation, Size Measure: Household Size.

2.7 Two-Stage Stratified Sampling, Sampling Units: Municipality (First Stage) and HH (Second Stage), Stratification (First Stage): NUTS 2 and Degree of Urbanisation.

Within the study, 10,000 samples are drawn from the Amelia dataset (see Alfons et al., 2011b, chapter 4) from each sampling design in order to evaluate the sampling distributions of the variance estimators. The following five indicators are evaluated: The ARPR, the RMPG, the QSR, the GINI, and the income mean (MEAN). All indicators are estimated for the equivalised disposable income of each person in the sample. The definition of the equivalised disposable can be found in Eurostat (2009).

Figure 1 shows the distribution of the relative bias for each variance estimator and indicator as boxplots. Boxplots labeled *naive* relate to the standard estimators in case of equal probability sampling, the ones labeled as *approx4* refers the approximation estimator of kind (4), used for estimating the variance for designs 1.5a and 2.6 (see *estimator 4* in Matei and Tillé, 2005). The resampling methods are labeled with *boot* and *rbrr* for the Bootstrap and Balanced Repeated Replication, respectively.

For one-stage designs (1.2, 1.4a and 1.5a) all variance estimators are almost unbiased. In general, the direct estimators, *approx4* and *naive*, show a greater efficiency than the resampling methods. It is noteworthy that using the naive variance estimator for design 1.5a results also to a comparable small relative bias. Another observation is that the bootstrap and the BRR estimators give also reasonable results for design 1.5a although these methods do not explicitly account for unequal probability sampling.

Finally, it needs to be mentioned that for the two-stage designs all variance estimators, except the bootstrap for design 2.7, underestimated the variance. In general they have negative relative bias well above 10%. In case of the variance estimators based on linearization it is assumed that there might be a problem of convergence. Casually speaking for the linearization method to work we need that certain remainder term R_n to converge in probability to zero as the sample size increases to infinity (see section 2.1 in Münnich and Zins, 2011). A sufficient condition for this to hold is met in the case of iid observations y_i, which is not straight forward to prove for general survey designs (see Demnati and Rao, 2004). Serfling (1980) makes some suggestions on how to analyse R_n in practice.

Figure 2 displays the kernel density estimators for the distribution of the different variance estimates. The kernel density estimation of the different variance estimation methods in figure 2 shows that the distribution of the *naive* approach and the *approx4* method for the designs with only one stage are almost normally distributed for all indicators. In comparison to these methods the distribution of the resampling methods for the ARPR and the RMPG is very flat and indicates a high variance for this variance estimators. For the two-stage designs have the empirical distribution of most variance estimators a heavy positive skewness. This reflects also the problem with the variance estimation for these designs. Nevertheless, one shall not forget that the size of the clusters and the

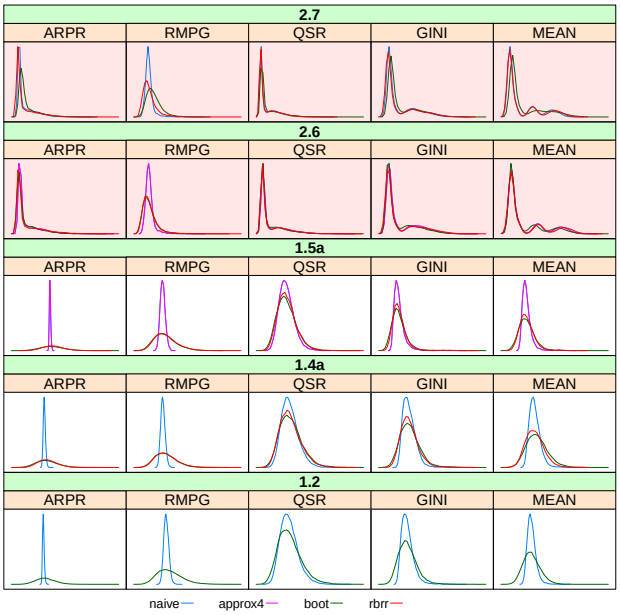

Figure 2: Density Estimation of Point Estimates

distributions therein is varying a lot such that the convergence is more problematic than in one-stage designs.

Summary and Outlook

The given example shows that variance estimation methods can appropriately be applied also in the non-linear case, which is evident for the use with modern indicators such as the Laeken indicators. Two or more stage designs still can give a lot of problems once the cluster sizes are varying a lot. This can have also a negative impact on design effect estimation, where model-based methods might seem preferable. A thorough overview of the methodology can be drawn from Bruch et al. (2011) and Münnich and Zins (2011). Further simulation results is given in Alfons et al. (2011a).

Acknowledgements

This research was part of the research project Advanced Methodology for European Laeken Indicators (AMELI; `http://ameli.surveystatistics.net`) funded by the European Commission within the 7th Framework Program.

References

Alfons, A., **Bruch, C.**, **Filzmoser, P.**, **Graf, E.**, **Graf, M.**, **Hulliger, B.**, **Kolb, J.-P.**, **Risto, L.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**a): *Report on the Simulation Results*. Technical report, AMELI deliverable D7.1, http://ameli.surveystatistics.net/.

Alfons, A., **Templ, M.**, **Filzmoser, P.**, **Kraft, S.**, **Hulliger, B.**, **Kolb, J.-P.** and **Münnich, R.** (**2011**b): *Synthetic Data Generation of SILC Data*. Technical report, AMELI deliverable D6.2.
URL `http://ameli.surveystatistics.net/`

Berger, Y. and **Skinner, C.** (**2004**): *Variance Estimation for Unequal Probability Designs*. Technical report, DACSEIS deliverable D6.1, http://www.dacseis.de.

Bruch, C., **Münnich, R.** and **Zins, S.** (**2011**): *Variance Estimation for Complex Surveys*. Technical report, AMELI deliverable D3.1, http://ameli.surveystatistics.net/.

Cochran, W. G. (**1977**): Sampling Techniques. New York: Wiley.

Davison, A. C. and **Sardy, S.** (**2004**): *Resampling Methods for Variance Estimation*. Technical report, DACSEIS deliverable D5.1, http://www.dacseis.de.

Demnati, A. and **Rao, J.** (**2004**): *Linearization Variance Estimators for Survey Data*. Survey Methodology, 30 (1), pp. 17 – 26.

Deville, J.-C. (**1999**): *Variance estimation for complex statistics and estimators: Linearization and residual techniques*. Survey Methodology, 25 (2), pp. 193– 203.

Deville, J.-C. and **Tillé, Y.** (**2005**): *Variance approximation under balanced sampling*. Journal of Statistical Planning and Inference, 18 (2), pp. 569 – 591, ISSN 0378-3758.

Eurostat (**2009**): *Algorithms to compute Overarching Indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC)*. Technical report, Eurostat Doc LC-ILC/11/08/EN – Rev. 2.

Hampel, F. R., **Ronchetti, E. M.**, **Rousseeuw, P. J.** and **Stahel, W. A.** (**1986**): Robust statistics. New York: Wiley, the approach based on influence functions.

Hulliger, B. and **Münnich, R.** (**2006**): *Variance Estimation for Complex Surveys in the Presence of Outliers*. Proceedings of the American Statistical Association, Survey Research Methods Section, pp. 3153–3161.

Kovačević, M. S. and **Binder, D. A.** (**1997**): *Variance Estimation for Measures of Income Inequality and Polarization - The Estimating Equations Approach*. Journal of Official Statistics, 13 (1), pp. 41 – 58.

Langel, M. and **Tillé, Y.** (**2011**): *Statistical inference for the quintile share ratio*. Journal of Statistical Planning and Inference, 141 (8), pp. 2976–2985.

Matei, A. and **Tillé, Y.** (**2005**): *Evaluation of Variance Approximations and Estimators in Maximum Entropy Sampling with Unequal Probability and Fixed Sample Size*. Journal of Official Statistics, 21 (4), pp. 543 – 570.

Münnich, R. (**2008**): *Varianzschätzung in komplexen Erhebungen*. Austrian Journal of Statistics, 37 (3&4), pp. 319 – 334.

Münnich, R. and **Zins, S.** (**2011**): *Variance Estimation for Indicators on Social Exclusion and Poverty*. Technical report, AMELI deliverable D3.2.
URL `http://ameli.surveystatistics.net/`

Osier, G. (**2009**): *Variance estimation for complex indicators of poverty and inequality using linearization techniques*. Survey Research Methods, 3 (3), pp. 167 – 195.

Rao, J. N. K. and **Wu, C. F. J.** (**1988**): *Resampling inference with complex survey data*. Journal of the American Statistical Association, 83 (401), pp. 231 – 241.

Serfling, R. J. (**1980**): Approximation theorems of mathematical statistics. New York: Wiley, wiley Series in Probability and Mathematical Statistics.

Shao, J. (**2003**): Mathematical Statistics. New York: Springer, second ed.

Shao, J., **Chen, Y.** and **Chen, Y.** (**1998**): *Balanced repeated replication for stratified multistage survey data under imputation*. Journal of the American Statistical Association, 93 (442), pp. 819 – 831.

Sitter, R. R. (**1992a**): *A resampling procedure for complex survey data*. Journal of the American Statistical Association, 87, pp. 755 – 765.

Wolter, K. M. (**2007**): Introduction to variance estimation. New York: Springer.

Yates, F. and **Grundy, P. M.** (**1953**): *Selection Without Replacement from Within Strata with Probability Proportional to Size*. Journal of the Royal Statistical Society. Series B (Methodological), 15 (2), pp. 253–261.