Measuring Poverty and Income Inequality Indicators and their Evolution over Time from Complex Survey Samples

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

Burgard, Jan Pablo*
E-mail: JPBurgard@uni-trier.de

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

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

Abstract

Estimating the change in poverty and income inequality indicators over time from repeated sample surveys is of increasing interest for evidence based policy making. Therefore, the topic of cross-sectional estimation is addressed first, where a focus is put on variance estimation for non-linear statistics and on the usage of model based estimators to improve the accuracy of point estimates. Secondly, variance estimation for change is considered, where change is measured through differences in cross-sectional estimates estimated from overlapping samples.

Keywords: Non-linear Statistics; Change over Time; Small Area Estimation; Variance Estimation

Indroduction

With *Europe 2020*, the EU has set up a new growth strategy for the forthcoming decade. To assess the issues of poverty and social exclusion, a set of indicators has been agreed on to be published each year by all member states. These indicators include the at-risk-of-poverty rate (ARPR) or the quintile share ratio (QSR). To measure these indicators on EU level in a comparable way, the European Statistics on Income and Living Conditions (EU-SILC) has been introduced. They include separate national sample surveys which collect micro data on an annual basis.

Since indicators are estimated from sample surveys´, variance estimation is needed in order to provide information on the sampling error. Because the statistics involved are highly non-linear, standard variance estimation procedures cannot be applied directly. Thus, resampling methods or linearization techniques can be used to assess the variance. Another aspect of sample surveys are their often limited sample sizes, especially with respect to particular domains or areas within the population. Estimation can be improved by model based (small area) estimators. These methods help to augment the information by borrowing strength from other domains or areas, and hence increase the effective sample size.

Besides interpreting indicator estimates for each year separately, their evolution over time is of major interest. Therefore, longitudinal measures have to be used. For instance, the difference between indicator values measured at two different years may be used as a measure of change. If a change is observed then the question arises whether it is statistically significant or not. For this reason variance estimation for measures of change is required. It is necessary to construct a statistical test, which can

be used as a basis for decision-making. If the samples are overlapping, then correlation through time between indicators has to be taken into account, as it is the case for the EU-SILC survey which has a rotational pattern (see Verma et al., 2007).

In this paper approaches to handle the above mentioned topics are presented exemplary for the ARPR. First, the issue of estimating the design variance of the ARPR is addressed. Second, a model based approaches to the estimation of the ARPR in small areas or domains is presented. Third, a method for variance estimation of change will be depicted. Finally, the paper concludes with a summary and an outlook.

Design Based Estimation of the APRR

The ARPR is defined as the share of persons in a population with an income below the at-risk-of-poverty threshold (ARPT). Within the EU the ARPT is set to 60% of the median equivalised disposable income (EDI) (for a definition of the EDI see Eurostat, 2009). Thus, the ARPR can be defined as

$$\text{ARPR} = F(0.6F^{-1}(0.5)) \ ,$$

where F is the distribution function of the EDI and F^{-1} is the inverse of F, i.e. $F^{-1}(0.5)$ is the median. If the ARPR is estimated based on a sample vector $\mathbf{y} = (y_1, \ldots, y_n)$, with y_i as EDI of the i-th element in a sample s of fixed size n, drawn from a finite population U of size N, then the following estimator may be used:

$$\widehat{\text{ARPR}} = \hat{F}(0.6\hat{F}^{-1}(0.5)) \ ,$$

where $\hat{F}(x) = \sum_{i \in s} w_i \mathbb{1}(y_i \leq x)(\sum_{i \in s} w_i)^{-1}$, $\hat{F}^{-1}(p) = \inf\left\{x \in \mathbb{R} : p \leq \hat{F}(x)\right\}$, and w_i is the survey weight of the i-th sampling unit.

Two widely used approaches to estimate the variance of $\widehat{\text{ARPR}}$ exist: resampling methods and linearization techniques. Resampling methods like the bootstrap, balanced repeated replication or jackknife routines select two or more (sub-) samples from a given population, or possibly from a sample, and compute a separate estimate of the population parameter of interest from each (sub-) sample. Variance estimation is done from the combination of all (sub-) samples. In general, there is no need to adapt resampling methods to a specific statistic, it might be difficult though to use them in the presence of complex survey designs, like sampling with unequal probabilities (Wolter, 2007; Bruch et al., 2011). Linearization techniques, quite to the contrary of the resampling, allow the utilization of standard variance estimation techniques. These are available for most survey designs used in practice (see e.g. Lohr, 1999). However, linearization requires that, at first, for each estimator a linear function approximating the estimation function is derived.

The main idea of linearization is to reduce the problem of estimating a non-linear statistic to that of a linear one. For statistics which can be expressed by functions that are continuously differentiable up to order two and are asymptotically normal, the Taylor method leads to proper results. This is, for instance, the case if the estimator can be displayed as a ratio of estimated totals or means. However, $\hat{F}$ is a discontinuous function which makes $\widehat{\text{ARPR}}$ not suitable for the Taylor approach. A solution to this problem is to use the concept of influence functions, which, up to now, is widely used in the field of robust statistics (see Hampel et al., 1986). The derivation of influence functions requires differentials of the estimator in the sense of Gâteaux (see Shao, 2003, pp. 339). Deville (1999) used the following form of an influence function of an indicator θ

$$IT\left(\theta(M), y\right) = \lim_{\epsilon \to \infty} \frac{T\left(M + \epsilon\delta_y\right) - T(M)}{\epsilon} \ ,$$

where δ_y is the unit mass at point $y \in \mathbb{R}$ with $M = \sum_{i \in U} \delta_{y_i}$ and $\theta(M)$ is defined as a functional with respect to the finite discrete measure M. If we estimate $\theta(M)$ by substituting M with a stochastic measure $\hat{M} = \sum_{i \in s} w_i$, it can be shown that $\mathrm{V}(\sum_{i \in s} z_i w_i)$ equals the asymptotic variance of $\theta(\hat{M})$, where z_i is the influence function $IT\,(\theta(M), y_i)$ (for the necessary assumptions see Deville, 1999 or Goga et al., 2009). Deville (1999) gave some practical rules to derive influence functions for varied functionals $\theta(M)$, which can be used to derive the influence function of $\widehat{\mathrm{ARPR}}$, which is given by

$$\begin{aligned} IT(\mathrm{ARPR}, y_i) =& \frac{1}{N}\left(\mathbb{1}[y_i \leq 0.6F^{-1}(0.5)] - \mathrm{ARPR}\right) \\ & - \frac{0.6F'[0.6F^{-1}(0.5)]}{F'[F^{-1}(0.5)]}\left(\frac{\mathbb{1}[y_i \leq F^{-1}(0.5)] - 0.5}{N}\right), \end{aligned}$$

where F' is the derivative of F. Because $IT(\mathrm{ARPR}, y_i)$ involves the unknown quantities ARPR, F^{-1}, and F', they need to be subsituted by estimates which gives $\hat{z}_i$, the sample estimate of z_i. For ARPR and F^{-1} estimators are given above and an estimator for $F'(x)$ can be obtained by a kernel density estimator (see e.g. Silverman, 1986). Finally, the asymptotic variance of $\widehat{\mathrm{ARPR}}$ can be estimated by $\widehat{\mathrm{V}}(\sum_{i \in s} \hat{z}_i w_i)$, where $\widehat{\mathrm{V}}$ is a variance estimator for an estimated total. In a pure design based framework we have $w_i = \pi_i^{-1}$, where π_i is the inclusion probability of the i-th element into sample s, and the estimator

$$(1) \quad \widehat{\mathrm{V}}\left(\sum_{i \in s} z_i w_i\right) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j>i}}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{z}_i}{\pi_i} - \frac{\hat{z}_j}{\pi_j}\right)^2$$

can be used, where π_{ij} is the probability of including both the i-th and the j-th element into sample s (see Cochran, 1977, p. 261).

Model Based Estimation of the ARPR

Sometimes it is useful to incorporate models into the estimation process. Especially when sampling fractions are small, direct estimators run into serious problems. Therefore, in small area statistics models are build over a wider population in order to borrow strength from other areas. For estimating the ARPR in a small area context two approaches are presented. Lehtonen and Veijanen (2009) and Lehtonen et al. (2011) propose to estimate the probability of a unit to fall under the poverty threshold. Assume $\widehat{\eta}_{id}$ is the estimated probability of unit i in area d to be at-risk-of-poverty under a certain model. Then an asymptotically design unbiased estimator for the ARPR is:

$$(2) \quad \widehat{\mathrm{ARPR}}_d^{\mathrm{GLMM}}(\widehat{\eta}_{id}) = N_d^{-1}\Big(\sum_{i \in U_d} \widehat{\eta}_{id} + \sum_{i \in s_d} w_{id}(\mathbb{1}(y_{id} < \mathrm{ARPT}) - \widehat{\eta}_{id})\Big)$$

The probability $\widehat{\eta}_{id}$ can be estimated, e.g., under a logistic model or a logistic mixed model (GLMM). To estimate the variance of (2) resampling methods can be employed (see e.g. Myrskylä, 2007 where several resampling methods are evaluated).

Molina and Rao (2010) proposed another approach. They model the income variable using a unit-level linear mixed model.

$$Y_{id} = X_{id}^T \beta + u_d + e_{id}, \quad v_d \sim N(0, \sigma_v^2), \quad e_{id} \sim_{iid} N(0, \sigma_e^2), \quad i = 1, \cdots, N_d, d = 1, \cdots, D\,.$$

From the fitted model a set of K prediction vectors $\widehat{y}_{id}^{(k)} = x_{id}\widehat{\beta} + \zeta_{id}$ is derived for the nonsampled population, with ζ_{id} beeing the sum of two random numbers $\zeta_d^{(u)} \sim N(0, \widehat{\sigma}_u^2(1 - \widehat{\gamma}_d))$ and $\zeta_{id}^{(e)} \sim N(0, \widehat{\sigma}_e^2)$. The ARPR is then estimated by

$$(3) \quad \widehat{\mathrm{ARPR}}_d^{\mathrm{MOLRAO}} = \frac{1}{N_d}\left\{\sum_{i \in s_d} \mathbb{1}(y_{id} < \mathrm{ARPT}) + \sum_{i \in U_d \setminus s_d} \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}(\widehat{y}_{id}^{(k)} < \mathrm{ARPT})\right\}.$$

Molina and Rao (2010) suggest to use a parametric bootstrap in order to obtain a MSE estimate for estimator (3).

Estimation of Measures of Change

Suppose we have a rotational sampling scheme which provides repeated samples from the same (stationary) population U over time. Further, let s_0 and s_1 denote two overlapping samples of equal size n, reported at time points $t = 0$ and $t = 1$, i.e. $s_0 \cap s_1 \neq \emptyset$. If one is interested in the development of indicator θ, estimated between $t = 0$ and $t = 1$, an intuitive estimator for change would simply be the difference $\widehat{\Delta} = \hat{\theta}_0 - \hat{\theta}_1$, where $\hat{\theta}_0$ is the estimate of θ based on sample s_0 and $\hat{\theta}_1$ the estimator of θ based on s_1. The variance of $\widehat{\Delta}$ is given by

$$\mathrm{V}\left(\widehat{\Delta}\right) = \mathrm{V}(\hat{\theta}_0) + \mathrm{V}(\hat{\theta}_1) - 2\,\mathrm{Cov}(\hat{\theta}_0, \hat{\theta}_1)\ .$$

If $\hat{\theta}$ is not suitable for linearization via Taylor series, Goga et al. (2009) show that $\mathrm{V}(\widehat{\Delta})$ can be approximated by using (partial) influence functions, hence the following approximation applies

$$\mathrm{V}\left(\widehat{\Delta}\right) \approx \mathrm{V}\left(\sum_{i\in s_0} w_{i_0} z_{i_0}\right) + \mathrm{V}\left(\sum_{i\in s_1} w_{i_1} z_{i_1}\right) - 2\,\mathrm{Cov}\left(\sum_{i\in s_0} w_{i_0} z_{i_0}, \sum_{i\in s_1} w_{i_1} z_{i_1}\right) \tag{4}$$

(see also Dell and d'Haultfoeuille, 2008). The first two terms in (4) are approximations for the variances of $\hat{\theta}_0$ and $\hat{\theta}_1$, where z_{i_0} and z_{i_1} are the influence functions of $\hat{\theta}_0$ and $\hat{\theta}_1$, respectively, which can be estimated by (1). The third term is the covariance between the estimated totals of z_{i_0} and z_{i_1}. Setting $w_{i_0} = \pi_{i_0}^{-1}$ and $w_{i_1} = \pi_{i_1}^{-1}$, where π_{i_0} and π_{i_1} are the inclusion probabilities of the i-th element into sample s_0 and s_1, respectively, the covariance can be estimated by

$$\begin{aligned}\widehat{\mathrm{Cov}}(\sum_{i\in s_0} w_{i_0}\hat{z}_{i_0}, \sum_{i\in s_1} w_{i_1}\hat{z}_{i_1}) &= \sum_{i\in s}\sum_{j\in s} \frac{\pi_{ij_{01}} - \pi_{i_0}\pi_{j_1}}{\pi_{ij_{01}}} \frac{\hat{z}_{i_0}}{\pi_{i_0}} \frac{\hat{z}_{j_1}}{\pi_{j_1}} \\ &= \sum_{i\in s} \left(\frac{\pi_{ii_{01}} - \pi_{i_0}\pi_{i_1}}{\pi_{ii_{01}}}\right) \hat{z}_{i_0}\hat{z}_{i_1} - \sum_{i\in s}\sum_{\substack{j\in s \\ j\neq i}} \left(\frac{\pi_{ij_{01}} - \pi_{i_0}\pi_{j_1}}{\pi_{ij_{01}}}\right) \hat{z}_{i_0}\hat{z}_{j_1}\ ,\end{aligned}$$

where $\pi_{ij_{01}}$ denotes the joint probability of including the i-th element in s_0 and the j-th element in s_1 and $\pi_{ii_{01}}$ is the probability of including the i-th element in both s_0 and s_1 (see Tam, 1984). How this longitudinal inclusion probabilites can be calculated depends on the rotational sampling scheme. Tam (1984) gave a framework to estimate the covariance from overlapping samples and applied it to some typical rotational sampling schemes. Qualité and Tillé (2008) also provide an overview on how to estimate the covariance between estimated totals for some specific cases of repeated samples. Berger (2004) presents an approach which allows for sampling designs with unequal selection probabilities.

Consider that $\hat{\Delta}$ is the difference between estimator of kind (2) or (3), variance estimation might not be possible by using the same variance formula as in (4). In contrast to classical estimators, where linerization is well studied, we have chosen for the model based case to follow a resampling approach instead. Therefore, we propose to employ the follwoing bootstrap method. Efron (1979) introduced the bootstrap for a iid sample vector $\mathbf{y}$. Here, the version of a Monte Carlo bootstrap is considered, where replicates y^* of y are generated by taking repeatedly random resamples of size n from $\hat{F}$. If we have samples on two occasions then we also need replicates $\mathbf{y}_0^*$ and $\mathbf{y}_1^*$ from the sample vectors $\mathbf{y}_0$ and $\mathbf{y}_1$, corresponding the samples s_0 and s_1, respectively. If, however, samples s_0 and s_1 are generated by a rotational sampling scheme, then we need to coordinate the generation of the replicates. For simplicity we assume that sets $s_{01} = s_0 \cap s_1$ and $s_{1\setminus 0} = s_1 \setminus s_0$ have fixed sizes n_{01} and $n_{1\setminus 0}$ and that s_0 and $s_{0\setminus 1}$ are selected by simple random samling. For elements in the overlapping part we take one

resample $s_{01}*$ of size n_{01} from s_{01}. From the non-overlapping parts we take a resample $s^*_{1\backslash 0}$ from $s_{1\backslash 0}$ of size $n_{1\backslash 0}$ and a resample $s^*_{0\backslash 1}$ from $s_{0\backslash 1} = s_0 \backslash s_1$ of size $n - n_{01}$. Thus, we have $s^*_0 = \{s^*_{01}, s^*_{0\backslash 1}\}$ and $s^*_1 = \{s^*_{01}, s^*_{1\backslash 0}\}$ boostrap replications for samples s_0 and s_1. The boostrap variance estimator for $\widehat{\Delta}$ would be

$$\mathrm{V}_{\mathrm{boot}}\left(\widehat{\Delta}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\Delta}^*_b - \frac{1}{B}\sum_{b=1}^{B}\widehat{\Delta}^*_b\right)^2 ,$$

where $\widehat{\Delta}^*_b$ is the estimate of Δ based on the b-th replicate of y_0 and y_1. This boostrap variance estimator is suitable if it can be assumed that the elements in the non-overlapping parts $s_{0\backslash 1}$ and $s_{1\backslash 0}$, are independend of each other (see Roberts et al., 2001).

Summary and Outlook

The methods presented within this paper can readily be extended to indicators of income inequality like the QSR or the Gini coefficient. For the estimation of the design variance of such indicator a thorough overview can be found in Münnich and Zins (2011). For small area estimates of these indicators we refer to Lehtonen et al. (2011).

In order to compare the performance of the different methods, empirical results from a large scaled design based Monte Carlo study will be presented. A special focus is put on the influence of sampling designs on the accuracy of the different estimators. Additionally, the importance of good auxiliary variables for the accuracy of small area estimators will be assessed.

Acknowledgements

This research was part of the research project Advanced Methodology for European Laeken Indicators (AMELI) funded by the European Commission within the 7th Framework Program.

References

Berger, Y. G. (**2004**): *Variance estimation for measures of change in probability sampling.* The Canadian Journal of Statistics. La Revue Canadienne de Statistique, 32 (4), pp. 451–467.

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.

Dell, F. and **d'Haultfoeuille, X.** (**2008**): *Measuring the Evolution of Complex Indicators: Theory and Application to the Poverty Rate in France.* Annals of Economics and Statistics, 90, pp. 259–290.

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

Efron, B. (**1979**): *Bootstrap methods: Another look at the jackknife.* The Annals of Statistics, 7 (1), pp. 1 – 26.

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.

Goga, C., **Deville, J.-C.** and **Ruiz-Gazen, A.** (**2009**): *Use of functionals in linearization and composite estimation with application to two-sample survey data.* Biometrika, 96 (3), pp. 691–709.

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.

Lehtonen, R. and **Veijanen, A.** (**2009**): *Design-based Methods of Estimation for Domains and Small Areas.* **Rao, C.** (editor) Handbook of Statistics - Sample Surveys: Inference and Analysis, *Handbook of Statistics*, vol. 29, Part 2, pp. 219 – 249, Elsevier.

Lehtonen, R., **Veijanen, A.**, **Myrskylä, M.** and **Valaste, M.** (**2011**): *Small Area Estimation of Indicators on Poverty and Social Exclusion.* Technical report, AMELI deliverable D2.2, http://ameli.surveystatistics.net/.

Lohr, S. (**1999**): Sampling: Design and Analysis. Pacific Grove: Duxbury Press.

Molina, I. and **Rao, J. N. K.** (**2010**): *Small area estimation of poverty indicators.* Canadian Journal of Statistics, 38 (3), pp. 369–385.

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/`

Myrskylä, M. (**2007**): Generalized Regression Estimation for Domain Class Frequencies. Research reports no. 247, University of Helsinki.

Qualité, L. and **Tillé, Y.** (**2008**): *Variance Estimation of changes in repeated surveys and its application to the Swiss survey of value added.* Survey Methodology, 34 (2), pp. 173 – 181.

Roberts, G., **Kovacevic, M.**, **Mantel, H.** and **Phillips, O.** (**2001**): *Cross-Sectional Inference Based on Longitudinal Surveys: Some Experiences with Statistics Canada Surveys.* FCSM Conference Papers.

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

Silverman, B. W. (**1986**): Density estimation for statistics and data analysis. London: Chapman & Hall.

Tam, S. M. (**1984**): *On covariances from overlapping samples.* The American Statistician, 38 (4), pp. 288–289.

Verma, V., **Betti, G.** and **Ghellini, G.** (**2007**): *Cross-sectional and Longitudinal Weighting in a Rotational Houshold Panel: Application to EU-SILC.* Statistics in Transition, 8 (1), pp. 5–50.

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