VARIANCE ESTIMATION OF LONGITUDINAL MEASURES OF POVERTY

Francesca Gagliardi, Tushar Kanti Nandi, Vijay Verma

Working Paper n. 64, December 2006

VARIANCE ESTIMATION OF LONGITUDINAL MEASURES OF POVERTY

Francesca Gagliardi[*], **Tushar Kanti Nandi**[*], **Vijay Verma**[*]

Abstract

Variance of poverty measures has implication for inference and decision making. For the cross-sectional measures of poverty, Jackknife Repeated Replication (JRR) and linearisation methods have often been used for variance estimation. For the cross-sectional measures of poverty, we find that the results from JRR method and linearisation method are comparable. However, the linearisation method is not readily extended to complex longitudinal measures such as persistent poverty rates. In this paper, we describe and apply JRR method for the variance estimation of longitudinal poverty measures. The JRR procedures for cross sectional measures are extended to estimate variance of longitudinal measures of poverty. We argue that this generalisation is straightforward. The key aspect is the definition of a common structure of the sample and of corresponding Jackknife replications covering the longitudinal duration of interest. We also calculate design effects under JRR methodology. The procedure is essentially the same irrespective of the type of measure, whether cross-sectional or longitudinal. It involved decomposition of the design effect into two components – the effect of sample weights, and the effect of clustering, stratification and the aspects of the design – and the estimation of the each component separately.

Keywords: Persistent poverty; longitudinal measures; variance estimation; design effects; Jackkinfe Repeated Replication; Linearisation; ECHP.

[*] Dipartimento di Metodi Quantitativi, Università di Siena. E-mail: francesca.gagliardi@tin.it , nandi@unisi.it and verma@unisi.it.

Authors acknowledge valuable suggestions from Gianni Betti and research assistance from Francesca Pierozzi.

1. Introduction

Measurement of poverty is of paramount importance for policy makers as well as for social scientists for the success of eradication programmes and for the scientific analysis of its cause and extent. Equally important is the variance estimation of poverty measures based on sample surveys for the purpose of inference and decision making. Traditionally poverty measures have been of the cross-sectional form. Over the last decades, it is increasingly recognised that the measurement of poverty has to be extended to the longitudinal dimension in order to capture its persistence and continuity (Bane and Elwood, 1986; Jenkins, 2000; Hills, 1998; Lillard and Wills, 1978; Rodgers and Rodgers, 1993). However, little is known about the sampling errors and design effects of persistent poverty rates and other longitudinal measures of poverty. Often Taylor linearisation is used to estimate the variance of cross-sectional poverty measures. The approximation gives a linear form for the variance formula of a complex statistic. To our knowledge, the linearisation approach has not been developed or applied to longitudinal measures of poverty. The paper explains a methodology – based on Jackknife Repeated Replication (JRR) - for the estimation of variance of longitudinal poverty measures. We also provide illustrative results for the estimation of variance of longitudinal measures of poverty using this methodology. One of the major advantages of JRR methods is that, once the sample structure has been specified and the required measures appropriately computed, procedures for estimating variance and design effects remain essentially unchanged when we move from cross-sectional to longitudinal measures.

The methodological innovation of the paper is the application of JRR method for the variance estimation of longitudinal measures for which linearisation methods are not presently available. The paper also presents, for the first time in the literature, design effects for longitudinal poverty measures estimated using the JRR method.

Longitudinal measures

The measures considered in this paper are the longitudinal poverty rates. The standard 'at-persistent-risk-of-poverty rate' included in the Laeken indicators is defined as follows:

> *Persistent-risk-of-poverty rate:* the share of persons with an equivalised disposable income below the risk-of-poverty threshold in the current year and in at least two of the preceding three years. The threshold is set at 60% of the national median equivalised disposable income.[1]

[1] Eurostat Structural Indicators. Summary Methodology Last update: January 2004

http://europa.eu.int/estatref/info/sdds/en/strind/socohe_di_base.htm

Computational algorithms for the construction of this measure are reproduced in Annex 1, based on a technical document prepared by Eurostat.

In this paper we consider the following set of related measures, constructed using first four waves of Italian component of the European Community Household Panel (ECHP)[2].

Cross-sectional poverty rates for each of the four years (waves 1-4 of Italian ECHP)

1. Average of the cross-sectional rates over the period considered (as a basis for comparison).

Longitudinal poverty rates defined over two consecutive waves (longitudinal sample over waves 1 and 2 of Italian ECHP)

2. Any-time poverty rate – proportion of persons poor during at least one of the two years
3. Continuous poverty rate – proportion of persons poor during both the years.

Longitudinal poverty rates defined over four consecutive waves (longitudinal sample over waves 1-4 of Italian ECHP)

4. Any-time poverty rate – proportion of persons poor during at least one of the four years
5. Persistent poverty rate – proportion of persons poor during at least three of the four years
6. Persistent poverty rate (Eurostat) – 'at-persistent-risk-of-poverty rate' as defined above
7. Continuous poverty rate - proportion of persons poor for all the four years.

The above measures are constructed for balanced panels, that is, the longitudinal samples of individuals present in the survey throughout the duration of interest, weighted so as to represent the corresponding longitudinal population. This refers to the longitudinal population defined over the first two years for measures 2-3, and over the first four years for measures 4-7. Even though measure 1 is a cross-sectional

Laeken indicators refer to the set of common indicators for the assessment and monitoring of poverty and social exclusion in an internationally comparable context.

[2] ECHP involved annual interviewing of nationally representative, panels of households and persons. For a description of methodology see Verma and Clemenceau (1996). The first four ECHP waves correspond to survey years 1994-97, the corresponding income reference period being calendar years 1993-96.

measure, for comparison it also has been computed for the longitudinal populations, in fact, it is computed separately for each of the two balanced panels (1-2) and (1-4).

Variance estimation

The procedure for variance estimation used is Jackknife Repeated Replication (JRR). The extension of the method to longitudinal measures is straightforward. It involves the following steps:

- We construct a balanced panel for the longitudinal duration of interest, and define a *sample structure common to all the cross-sectional samples which are put together to construct the balanced panel.* The sample structure means defining the 'computational' strata and primary selection units (PSU's) to be used in the variance estimation, such that each stratum contains at least two PSU's. The common structure also determines a common set of Jackknife replications to be used for variance estimation.
- The remainder of the procedure is the same as that for any statistic whether cross-sectional or longitudinal. Jackknife replications are formed by deleting one PSU at a time and compensating for it by increasing the weights of the remaining PSU(s) in the stratum of the deleted unit. In this way, each Jackknife replication is also a balanced panel, different from the full sample balanced panel. Just as that for the full-sample balanced panel, the longitudinal measure of interest is constructed for the balanced panel of each replication, and these values are inserted into the standard JRR formula to obtain variance estimation of the measure concerned.

We are not presently aware of alternative procedures based on the linearisation approach for estimation of variances of longitudinal measure. If and when such procedures are available, variance estimates from the two different approaches – replication and linearisation – can be compared and evaluated.

Design effect

This is the ratio of the variance under the given sample design, to the variance under a simple random sample of the same size. (We use deft^2 for the ratio of variances, and deft for the corresponding ratio of standard errors).

With JRR, a slightly indirect procedure is required for estimating the design effect. As will be explained, we can estimate two factors making up the design effect separately: (1) the effect of sample weights on variance, and (2) the effect of clustering, stratification and aspects other than weighting. The product of these two factors gives the required overall design effect for the statistics concerned.

Data set

For the illustrations, we have used first four waves of the Italian ECHP. The data mostly come from the User's Data Base (ECHP-UDB)[3]. Unfortunately, sufficient information is not provided in ECHP-UDB for the identification of the sample structure, in particular, the identification of the stratum and PSU to which a household or person belongs, nor on how the PSU's have been selected. Such information is available in the Production Data Base (ECHP-PDB), D-File, which is not available to the researcher outside the National Statistical Agency or Eurostat. In our case, thanks to the co-operation of Istat we have been able to use the ECHP-PDB that provides information about the structure of the sample.

But even here, the information available on the sample structure in the Italian ECHP-PDB is not complete, and cannot be easily connected to the available documentation on the survey.

In fact the final sample structure as we have used comes not only from the UDB and its PDB version, but also from descriptions of the sample provided in various documents by Istat.

Rest of the paper is organised as follows. Section 2 presents the main results of JRR approach applied to longitudinal poverty measures. Section 3 describes the JRR methodology which applies equally to cross-sectional and longitudinal measures. It also provides illustrations with variance computations for cross-sectional poverty rates, presenting also comparison between the JRR and Linearisation methods. Aspects specific to the longitudinal context such as concerning the construction of balanced panels and definition of various measures of persistent poverty are presented in Section 4. Technical details of all the sections are relegated to the Annexes. Section 5 concludes the paper.

2. Main longitudinal results

Main results of this paper presented in this section concern the analysis of poverty indices from a longitudinal perspective. The measure we consider is the 'at-risk-of poverty rate' or the so-called Head Count Ratio (HCR), and its persistence over-time.[4]

[3] Eurostat (2003a, 2003b, 2003c).

[4] The income of an household/individual in a wave corresponds to the income in the preceding calendar year.

The total disposable income of a household is calculated by adding together the personal income received by all the household members plus any income received at the household level.

Cross-sectional poverty and inequality measures are constructed with reference to the equivalised disposable income distribution of all individuals in the population, considered separately for each cross-section. That is, the poverty (poor/non-poor) status of an individual is purely a cross-sectional measure, defined only by the income distribution at that cross section.

The standard definition of poverty indicator considers persons with equivalised disposable income below poverty lie, i.e. below 60% of the median as poor and the rest as non-poor. HCR is defined as percentage of people with equivalised income below poverty line. Longitudinal indicators of persistence of poverty are defined in terms of patterns of these cross-sectional statuses, i.e. are constructed by appropriately putting together cross-sectional indicators for the period covered at the micro-level (i.e. at the level of individual person). For this latter purpose, the cross-sectional indicators themselves may be defined with references to the population restricted to the longitudinal population common to the period concerned – see Annex I. For the present analysis we have constructed two balanced panels, where the balanced panel consists of all the people enumerated (and with information of equivalised income available) in all the waves of the panel considered. With appropriate sample weights, a balanced panel is meant to represent the corresponding longitudinal population. The first balanced panel in our analysis consists of two waves and it is constructed using the first two waves of the Italian ECHP survey (year 1994 and 1995). The second balanced panel consist of four waves and is constructed using the first four waves of the survey (1994-1997). See Annex 2 on the procedure for constructing the balanced panels and the structure of the sample as defined for the purpose of variance computations.

The equivalised household size is defined according to the modified OECD scale (which gives a weight of 1.0 to the first adult, 0.5 to other household members aged 14 or over and 0.3 to household members aged less than 14).

For each person, the equivalised disposable income is defined as his/her total household disposable income divided by equivalised household size. Each person in the same household receives the same equivalised disposable income.

Table 1

Overall results. Relative standard errors (%se) and design effect (deft) for longitudinal poverty rates.

Measure Head Count Ratio	(1) est	(2) se	**(3)=(2)/(1)** **%se**	Effect of clustering & stratification	Effect of weighting	**Deft**
Panel 2 waves (n=19,984; h=6,656)						
any time poverty	27.8	0.72	**2.59**	0.97	1.21	**1.17**
Mean p1 p2	20.2	0.70	**3.45**	1.16	1.20	**1.39**
continuous poverty	12.5	0.61	**4.89**	1.26	1.18	**1.48**
Panel 4 waves (n=16,960; h=6,022)						
any time poverty	35.8	0.68	**1.90**	0.98	1.15	**1.13**
Mean p1 p2 p3 p4	19.6	0.77	**3.92**	0.96	1.13	**1.10**
persistent poverty	13.8	0.53	**3.87**	1.14	1.12	**1.27**
eurostat persistent poverty	11.5	0.54	**4.72**	1.14	1.11	**1.27**
continuous poverty	7.2	0.58	**8.10**	1.19	1.12	**1.32**

n = no. of sample individual persons in the balanced panel.

h = number of households from which these persons came (at wave 2 in the first panel of the table, and at wave 4 in the case of the second panel)

Deft: it shows the effect on standard errors of clustering within areas, stratification and weighting[5].

Table 1 presents the results obtained using the JRR methodology. (Details on the methodology are given in the subsequent sections).

The results shown in Table 1 are for the longitudinal populations represented by the two balanced panels. We have constructed two sets of poverty rates corresponding to these two longitudinal populations represented by the two balanced panels.

The first set is based on the population covered in the two-wave balanced panel, i.e. on individuals with equivalised household income recorded at both waves. For constructing measures at either wave the data are weighted by the individuals' "base weights" at the most recent wave included in the panel – in this case weights at wave 2. With this common sample (and common weighting), an individual's position in the income

[5] Deft is the ratio of standard error under the actual sample design, compared to that under a simple random sample of households, of the same size (h). Note that the $deft^2$ will be larger by a factor (n/h) if it were defined with reference to equivalent simple random sample of persons, of size (n). With the latter definition, the higher deft values will take into account the fact that equivalised income is measured at the household level and hence is uniform for all persons clustered in the household.

distribution and the status as poor/non-poor are determined at each wave, based on the income at that wave.

The individual's poverty status at each wave, defined as above, can be used to define his/her longitudinal poverty status:

- Whether the person is subject to any-time poverty – poor at either of the two waves
- Whether the person is subject to continuous poverty – poor at both of the two waves.

The (weighted) proportion of "yeses" to these statuses give the corresponding poverty rates.

The first part of the Table 1 shows the results for any-time and continuous poverty rates over two years and, for comparison, also the average of the cross-sectional rates ("mean p1 p2") over the period. All measures are computed using the two-wave balanced panel, weighted with the individuals' base weight at wave 2.

The second part of Table 1 shows results for similar measures constructed with the four-wave balanced panel as the sample base. To represent the longitudinal population over the four waves, the most appropriate sample weights available are the "base weights" of individuals at wave 4.

Measures p1 to p4 are the cross-sectional poverty rates for the individual population over the interval covering the four waves[6]. Their average is shown in Table 1 for comparison with the longitudinal rates. As above, based on the individuals' cross-sectional poverty status at each of the four waves, we can define various measures of his/her longitudinal poverty status. The measures shown in Table 1 include:

- any-time poverty: poor in any wave of the period covered
- continuous poverty: poor in all the waves during the period.
- persistent poverty: poor at three of the four waves.
- Eurostat persistent poverty: poor at 4th wave (the most recent wave in the period covered), and at least at two of the preceding three waves.[7]

First concerning the poverty rate estimates themselves. For the cross-sectional percentage in poverty, Table 1 reports the average over two years and four year periods.

[6] Note that (p1, p2) here are not identical to (p1, p2) for the two-wave balanced panel, because these two sets are defined over somewhat different longitudinal population. The four-wave population is a subset of the two-wave population.

[7] See Annex 1 for complete definition.

For Italy, about one individual out of five is poor at any given time. The slight difference in the averages in the two parts of the table (20.2% vs 19.6%) is, almost certainly, simply the result of selective panel attrition.[8]

One individual out of three is poor in at least one year during the four years, and one individual out of four is poor in at least one of the first two years. At the other end, 12.5% of individuals are always poor (continuous poverty) in the first two waves, and 7.2% are always poor over the four waves.

The results are of course logically consistent in the sense that any-time poverty rates are (much) higher than continuous poverty rates. In fact, as expected, the rates follow the sequence

$$\text{any-time} > \text{cross-sectional} > \text{persistent} > \text{“Eurostat persistent”} > \text{continuous}.$$

Also, the any-time rate goes up and the continuous poverty rate goes down as the period of the observation increases (from 2 to 4 years in the present example).

The main objective of the table is, of course, to report standard errors and design effects for longitudinal poverty measures, and compare those with the corresponding figures for cross-sectional rates. In absolute magnitude of standard errors declines from any-time poverty rate to continuous poverty rate (column 2), but substantially rises in relative terms (column 3, Table 1). This is expected from the fact the estimates itself declines as we move from any-time to the continuous rate. For instance, we know that for a simple proportion (p) in simple random sample, and assuming small p, standard error in absolute terms varies as $\sqrt{p}$, while in relative terms it varies as $1/\sqrt{p}$.

In fact, as we go down the rows of Table 1, the values of the standard error are higher than what would be expected from the above variation with p in a simple random sample.

This is because the design effect is increasing as we move from any-time to continuous poverty measures (last column of Table 1). Design effect (deft) is the rates of actual standard error for a statistics, to what this error would be in a simple random sample of the same size[9].

[8] This is confirmed by the annual figures in Table 4 below.

[9] As noted in Table 1 the denominator in the definition of deft is not the value of standard error corresponding to a completely random sample of individuals (containing the same number of individuals as in the actual sample), but corresponding to a simple random sample of (the same number of) households. Hence, deft as defined in Table 1 does not include the effect of “clustering” of individuals into households, at the level of which income is actually measured.

Design effect can be decomposed into two components as shown in Table 1: the effect due to sample weights; and the effect of clustering, stratification and other aspects of the design.

A major factor determining the effect of weighting is the variability of weights in the sample. For a given sample, this variability tends to be similar for whatever measure we considered – for instance it is very similar for the various statistics with the two parts of the table[10].

Hence, the increase in overall design effect as we move from any-time to continuous poverty rate arises mostly from the increase associated with a higher degree of clustering – i.e. a greater homogeneity on the variable concerned among households within the same sample cluster.

It is not unexpected that there is greater intra-cluster homogeneity when it comes to persistence or continuity of poverty, to the extent that poverty is spatially influenced/determined. The area effects are likely to be less strong when it concerns the more transient movements in and out of poverty which determine, for instance, any-time rates.

3. Methodology and cross-sectional results

This section describes the variance estimation methodology adopted, many aspects of which apply equally to cross-sectional and longitudinal measures. Some cross-sectional results are presented to illustrate the methodology. Some specifically longitudinal aspects will be discussed in Section 4.

Variance estimation for cross-sectional measures has been widely analysed in the literature (Verma and Betti, 2005). In this section we summarise the methodology and provide some illustrative results with the objectives to clarify the methodology used for the estimation of variances and design effects using Jackknife Repeated Replication (JRR).

For completeness, we also compare the results on variances and design effects obtained by using two different methodologies - JRR and Linearisation. Presently, this has been done only for cross-sectional measures of poverty and inequality.

As before, the illustrative results presented in this section pertain to cross-sectional analysis of the four waves (1994-1997) of the ECHP survey for Italy. Each measure

[10] The difference in this effect between the two panels in the table indicates that wave 4 base weights (used in the four-wave panel) are, as provided by Eurostat in ECHP-UDB, less variable than wave 2 base weights (used in the two-wave panel).

(poverty rate) in this section is based on the full cross-sectional sample of the year (ECHP wave) concerned, using the cross-sectional weights for that wave.[11]

3.1 Methodology

Our main methodology, the Jackknife Repeated Replication (JRR), is one of a class of methods for estimating sampling errors from comparisons among sample replications which are generated through repeated resampling of the same parent sample. The JRR method was originally introduced as a technique of bias reduction (Durbin, 1959). However, it has been widely used for variance estimation (Kish and Frankel, 1974). A detailed discussion of Jackknife methodology can be found in Efron and Stein (1981). In the 'standard' version, each JRR replication can be formed by eliminating one PSU from a particular stratum at a time, and increasing the weights of the remaining PSU's in that stratum so that the sum of the weights in that stratum remain the same as that before eliminating the PSU. Each such replication provides an alternative, but an equally valid, estimate of the statistic concerned to that obtained from the full sample.

Let u be the full-sample estimate of a statistic of any complexity (in our example, poverty rate), and $u_{(hi)}$ be the estimate produced using the same procedure, but after eliminating primary unit (PSU) i in stratum h and increasing the weight of the remaining (a_h-1) PSU's in the stratum by an appropriate factor g_h (see below). Let $u_{(h)}$ be the simple average of the $u_{(hi)}$ over the a_h values of i in h. The variance of u is then estimated as:

$$\operatorname{var}(u) = \Sigma_h \left[(1 - f_h) . \frac{a_h - 1}{a_h} . \Sigma_i \left(u_{(hi)} - u_{(h)} \right)^2 \right], \quad [1]$$

where (1-f_h) is the finite population correction, usually ~1.

Concerning the re-weighting of units retained in a stratum after dropping one unit, generally the factor is taken as in [1], $g_h = \frac{a_h}{a_h - 1}$; in this paper we adopt the proposal of Verma and Betti (2005) so that $g_h = \frac{w_n}{w_n - w_{ni}}$, where $w_n = \sum_i w_{ni}$, $w_{ni} = \sum_j w_{nij}$, the sum of sample weights of ultimate units j in PSU i.

[11] In Table 2 below, we have used symbols W1-W4 to denote poverty rates for the full cross-sectional samples. Please note the distinction between these and the cross-sectional rates (p1-p2) and (p1-p4) computed for the balanced panels referred to in Section 2, and described further in Section 4 below.

The second approach, namely Linearisation, is based on the use of Taylor approximation to reduce non-linear statistics to a linear form, justified on the basis of asymptotic properties of large populations and samples (Deming, 1934; Kendall and Stuart, 1958; Keyfitz, 1957). For each ultimate unit (e.g., household or person) in the sample, it seeks a linearised 'indicative' variable, in such a way that the simple expression estimating the variance of the total of this linearised variable, under the given sampling design, approximates the required variance of the original complex statistic.[12] For a cross-sectional poverty rate (p) the linearised variable is

$$u_j = (p_j - p) - f_p \beta \left(\frac{\alpha_i - \alpha}{f_\alpha} \right) \qquad [2]$$

where $\alpha = 0.5, \beta = 0.6, \alpha_i = \delta(y_i \le y_\alpha)$ and f_p, f_α density functions.

The basis of Taylor linearisation is the following simple variance estimation formula for aggregates in multistage stratified samples of large size:

Let $u_{hi} = u(y_{hi}, x_{hi}, \ldots);\ u_h = \sum_i u_{hi}$ be a sample aggregate or a linear function of sample aggregates such as of $y = \Sigma_h y_h;\ y_h = \Sigma_i y_{hi};\ y_{hi} = \Sigma_j (w_{hij} \cdot y_{hij})$. Then its variance is estimated as:

$$\mathrm{var}(u) = \Sigma_h \left[(1 - f_h) . \frac{a_h}{a_h - 1} . \Sigma_i \left(u_{hi} - \frac{u_h}{a_h} \right)^2 \right], \qquad [3]$$

with the quantity u_{hi} defined at the level of primary selection (h,i) as the weighted sum of values of the ultimate units j in the PSU. Here j refers to ultimate sampling unit, i to PSU, and h to stratum; $a_h>1$ is the number of sample PSU's in stratum h; and $(1-f_h)$ is the finite population correction, usually ~1.

[12] Presently the linearised forms for the full range of measures of poverty and inequality are available only for the cross-sectional case.

3.2 Sample structure and computational units

For the analysis of this section we have used the four cross-sectional household data files (H-files) of the first four waves of Italian ECHP. The actual sample structure may be summarised in two parts as follows. Details are provided in Annex 2.

The major part of the sample consists of a two-stage stratified sample. The primary strata are Italian Regions, each of which is then divided into finer strata of approximately uniform size according to the size group of municipalities in it. From each finer stratum, one municipality is selected as the PSU with probability proportional to its population size. The second part of the sample consists of large municipalities all of which are included in the sample automatically. In either case, a random sample of households is selected systematically from each municipality in the sample.

Some aspects of this sample structure have to be redefined to make variance computation possible. The computational structure can differ from the actual sample structure because of various consideration noted in Annex 2. Note that such considerations apply equally to any of the practical methods available for variance estimation – whether JRR or linearisation for instance.

In the main part of the sample, adjacent municipality size strata are paired to provide computational strata, each of which contains two PSU's – that being the minimum number required for variance computation. In the second part, consisting of 'self-representing' municipalities, each municipality in fact forms a stratum, and random groupings of sample households within each such municipality form the computational PSU's. The resulting sample so defined consists of 121 computational strata and 253 computational PSU's.

This defines the sample structure for wave 1 of the panel. In relation to defining the sample for subsequent waves, the basic point is that in a household panel the sample structure remains essentially the same over the survey waves. The structure is defined by where and how the original sample of households and persons was selected at wave 1. Irrespective of any original sample persons moving to new locations over the life of the panel, each person's position in the structure of the sample (the stratum, PSU, SSU, etc., to which the unit belongs) remains unchanged, namely, as it was defined at the time of the original selection into the sample. The location in this structure of a household or a person in subsequent waves is determined by the unit's 'root household id', which identifies the original wave 1 household which the sample person(s) in the current household came from. In this way, a common set of stratum and PSU identifiers can be defined for each household and person in any wave.[13]

[13] The above defines the structure used for computation involving the two-wave balanced panel. A minor exception to the common sample structure across waves arose in wave 4. Because of sample attrition, one of the PSU's in this wave turned out not to contain any

The final ECHP-UDB data have been weighted using a complex procedure taking into account design probabilities, non-response and calibration to external controls. Furthermore, complex procedures have been used by Eurostat or countries to impute missing income components in the presence of item non-response. In our illustrations, the procedure takes the sample weights and imputations as given, and does not include the effect of their sample dependence on the variance estimations.

Sample weights are of course wave-specific: each household and each of its members receive the wave-specific household cross-sectional weight. The unit of analysis for income related statistics is the individual person rather than the household. Nevertheless, since the variable of interest (equivalised household income) is identical for all members of a household, it is possible (and convenient) to use the household level data file for the purpose. For the cross-sectional analysis all that is required is to assign to each household a weight equal to its cross-sectional weight times the household size. For longitudinal measures based on a balanced panel the analysis can also be performed on the household file, using the person's most recent household as the unit of analysis. The appropriate "household weight" to use is the sum of the most recent "base weights" of persons in that household.

3.3 Comparison of variance

Illustrations are provided below only for cross-sectional measures. The main objective is to explain the procedure in detail.

Table 2 presents the estimate of the poverty rates based on the four cross-sectional samples and corresponding standard errors calculated with two methods – JRR and Linearisation. As already noted in Section 2, the panel attrition causes a loss of households and individuals, especially of those at the lowest end of the income distribution.

There is a significant loss in the sample at wave 4 (see Annex 2). The poverty rate estimate for wave 4 is somewhat lower than that for the preceding waves, which may well have been caused by selective loss of the sample due to panel attrition. It also suggests that adjustments to the sample weights have not been able to fully compensate for this loss. See Annex 3 for a brief note on some problems with the data sets used for the present illustrations.

completed interviews. The non-empty PSU of the stratum that contained this empty PSU was moved to the preceding computational stratum, resulting in a total of 120 strata and 252 PSU's in wave 4 sample. This slightly modified sample structure was used for computations involving the four-wave balanced panel.

Concerning standard error estimates, in general the results from the two methodologies are quite similar. The estimated variance is similar in different waves, except for the 4th wave that shows large values of variance, especially with the JRR method.[14]

In any case, an important methodological point should be noted when comparing sampling errors from different estimation methods. For the following reasons we feel that, broadly, the two variance estimation procedures have given comparable results.

Firstly, estimates of sampling error are themselves subject to sampling variability, which can be large depending on factors such as the number of PSU's available in the sample for such estimation.

Secondly, at least with fairly large samples (as in the case for the present illustrations), sampling error of a substantive indicator is a small quantity compared to the indicator itself.

Thirdly, from the substantive point of view, while we need precise estimates of the indicator itself, it is sufficient to have, in relative terms, more approximate values of the magnitude of its sampling error for the purpose of drawing inferences or taking decisions.

For all these reasons, we can accept larger differences – in relative terms – in the estimated sampling error with different methods, than differences in the estimates of the indicator itself. Thus if, for instance, a 2-3% difference in the estimated indicators is considered substantively 'significant', a difference of this order in the estimated sampling errors of the indicator could be justifiably regarded as 'trivial'. Similarly, while a difference exceeding, say, 4-5% in the estimated indicators may be considered 'large' differences, the label 'large' may be reasonably applied to the estimates of the sampling error only if the difference exceeds, say, 10-12%.

[14] Generally, the results involving wave 4 data have been somewhat less satisfactory and stable in our illustrations. This results from some peculiarities (irregularities) in the observed income distribution for that wave. See Annex 3.

Table 2

Estimates and standard errors of cross-sectional poverty rates (HCR) using the JRR and linearisation methods ("Taylor")

Measure			JRR	Taylor	Comparison
HCR based on 60% median	(1) est	(2) se	(3)=(2)/(1) %se	(3)=(2)/(1) %se	%se (Jrr/Taylor)
W1	20.4	0.77	3.77	3.38	1.12
W2	20.5	0.59	2.89	2.97	0.97
W3	20.2	0.63	3.13	2.91	1.08
W4	19.7	0.88	4.45	3.09	1.44
Mean	20.2	0.72	3.55	3.09	1.15

For the calculation of JRR standard errors, extreme values of contribution of individual replications to the total variance estimate have been trimmed so as to ensure that no single replication contributes disproportionately to the total estimate of variance. This precaution is desirable as a protection against the presence of large irregularities or outliers in the empirical data used. See Annex 3 for procedure used.

3.4 Randomisation

We have also randomised our sample in order to estimate particular components of the design effect, as explained in the next subsection. The randomised sample structure consists of only one stratum and 50 PSU's. The randomisation is achieved by generating random numbers from a uniform distribution and ordering the sample households according to these numbers. Then random grouping of households, which serve as computational PSU's, are defined as follows. After randomisation, the sum of all the cross-sectional weights is divided by the number of PSU's required (50 in our case) to obtain the target sum of weights of units to be grouped to construct one PSU. Then we cumulate the household weights in the randomly arranged list of households till we reach the sum required per PSU. The households in the cumulation form one PSU, and cumulation begins again down the list of households for the construction of the next PSU.

3.5 Design effect

As noted in Section 2, design effect (deft) is estimated by the ratio of actual standard error (se) of a statistics under the given sample design, to standard error (se_srs) under a simple random sample of same size. In this section we outline the procedure for estimating design effects, especially under the JRR approach.

Design effect with the linearisation method

This method is based on defining a linearised variable (say u_{hif} for ultimate unit f in PSU i, stratum h) such that the ordinary expression for variance of the total of this linearised statistics gives an estimate of variance of the complex statistic of interest. As noted in equation [3], this is:

$$se^2 = \text{var}(u) = \sum_h \left[\frac{a_h}{a_h - 1} \sum_i \left(u_{hi} - \frac{u_h}{a_h} \right)^2 \right],$$

where $u_{hi} = \sum_j w_{hij} \cdot u_{hij}$, $u_h = \sum_{i-1}^{a_h} u_{hi}$, $u = \sum_h u_h$, and the finite population correction is disregarded.

Once u_{hij} has been defined, it can be used to directly estimate variance under a equivalent simple random sample

$$(se\_srs)^2 \equiv \sum_{j \in S} w_j \cdot u_j^2 / \sum_{j \in S} w_j \qquad [4]$$

where the sum is over all ultimate units j in the sample *S*.

Hence, from the ratio of the two quantities defined above[15],[3]/[4], directly gives the required design effect for the linearised method, at least for cross-sectional measures for which the required linearised variable (u_{hij}) has been developed.

Decomposition of the design effect

For the general application of the JRR approach specially to longitudinal and other more complex situations to which the linearised approach is not easily extended, we have to assume that the required linearised form (u_{hij}) is not available, so that (se_srs), and hence deft cannot be estimates directly. An indirect approach is required. One such approach involves decomposition of the design effect into components, each of which can be separately estimated without requiring the fully linearised variable (u_{hij}). The

[15] There in no need to distinguish different h and i values in this case. Note that [4] defines the denominator of deft ^{2} in terms of a simple random sample of households.

required components are (1) the effect of sample weights on variance, and (2) the effect of clustering, stratification and other aspects of the design.

In fact, the identification of the effect of weighting is in itself of substantive interest apart from its usefulness for the above purpose. A question of great practical interest is the following. How does the weighting affect variances? There are effects in both directions:

(i) Calibration weights and other weighting correlated with the survey variables can reduce, not only bias, but also variances. (Optimal allocation in stratified samples and the corresponding weighting involved is an obvious example.)

(ii) On the other hand, very often weighting is determined on the basis of 'external' factors (e.g., need to over-sample small regions; compensation for high non-response in particular sample areas etc.). Such weighting, essentially uncorrelated with survey variables, results in increased variance.

Generally, the second of the above effects is found to predominate in practice. That is, usually the net effect of weighing is to inflate variances[16].

Again, when the full linearised variable (u_{hij}) is available, variance of an equivalent element sample but retaining the effect of weighting can be directly estimated as

$$(se\_wtd)^2 = n.\sum w_j^2 \cdot u_j^2 / \left(\sum w_j\right)^2 \qquad [5]$$

Hence, deft can be directly decomposed as

$$deft = \left(\frac{se}{se\_wtd}\right) \cdot \left(\frac{se\_wtd}{se\_srs}\right) = D_u \cdot (Kish\_talyor) \qquad [6]$$

The first factor in the above (D_u) is the effect of sample design features other than weighting. The second factor

[16] Proper weighting should of course reduce mean-squared error, by controlling bias even if there is some increase in variance.

$$(Kish\_taylor)^2 = \left(\frac{se\_wtd}{se\_srs}\right)^2_{Taylor} = \left(\frac{n}{\sum w_j}\right) \cdot \frac{\sum w_j^{\,2} u_j^{\,2}}{\sum w_j u_j^{\,2}} \qquad [7]$$

is an estimate of the effect of weighting. We have termed it "Kish_taylor" for the following reason.

This formula can be directly compared with the well-known "Kish Factor", proposed by Kish to estimate the effect of essentially "random" weights on variance:

$$(Kish\_Factor)^2 = D_w^{\,2} = n \cdot \frac{\sum w_j^{\,2}}{\left(\sum w_j\right)^2} = \left(\frac{n}{\sum w_j}\right) \cdot \frac{\sum w_j^{\,2}}{\sum w_j} = 1 + cv^2\left(w_j\right), \qquad [8]$$

where w_j is the sample weight of unit j, and the sum is over n units in the sample; *cv* is the coefficient of variation of unit weights. D_w indicates the design effect (deft) due to weighting: standard errors are inflated by D_w.

The above factor is simple and general as it is determined simply from the variability of weights in the sample, independently of any substantive variable.

The factor (Kish_taylor) is an alternative and in principle a more accurate expression of the effect of weighting than the simple Kish_factor. This applies specifically in situations where the sample weights are not "random" but are systematically correlated with substantive variables of interest – as is the case with calibration and similar types of adjustments applied to weights in ECHP.

Decomposition under the JRR method

Again, an indirect procedure is required if we assume that the full expression for the linearised variable (u_{hij}) cannot be evoked. An alternative estimate of variance of a weighted element sample, (se_wtd), can be obtained by "randomising" the sample (see Section 3.4) and applying the ordinary stratified multistage variance estimation formula to it.

A randomised sample is created from the actual sample by completely randomising the position of individual elements (households, persons) within the sample structure. In principle, this creates a random element sample, which is not subject to clustering or stratification effects, and differs from a true simple random sample simply because of the presence of unequal weights. Random groupings of the elements can be formed to serve as clusters and strata in the variance estimation without affecting the expected varinaces.

We term the standard error estimated from such a randomised sample using JRR as (se_rnd). This, in theory, is the same as (se_wtd) defined earlier, and their empirical closeness has been verified in Verma et al. (2006).

Consequently we can estimate the effect of clustering, stratification and any factors other than weighting as

$$\text{Effect of clustering and stratification} = \left(\frac{se}{se\_rnd}\right). \qquad [9]$$

Now concerning the estimation of the effect of weighting under the JRR approach, again, we have to assume that the full expression for the linearised form (u_j) is not available, as for instance presently in the case of longitudinal measures. In the linearisation approach, the expression for any (u_j) is developed to have two parts. The first part, say (u_{1j}), is simple and corresponds to treating the complex statistics under consideration as a simple ratio. The second part is the effect of complexity of the actual statistics. For instance, for poverty rate (p), the first part is simply

$$u_{1j} = (p_j - p) \qquad [10]$$

where the full expression for u_{j} and the terms involved have been defined in equation [2]. The above is exactly the linearised form for variance estimation of a ordinary ratio. Hence, this first part for the linearised form is always available for any complex statistics without evoking the full linearisation methodology.

In Verma et al. (2006), it has been empirically demonstrated, at least for a wide variety of cross-sectional measures of poverty and inequality, that the expression

$$(Kish\_Jrr)^2 = \left[\frac{n}{\sum w_j}\right] \cdot \frac{\sum w_j^2 \cdot u_{1j}^2}{\sum w_j \cdot u_{1j}^2} \qquad [11]$$

provides a very close approximation to the effect of weighting computed above as

$$(Kish\_taylor)^2 = \frac{n}{\sum w_j} \cdot \frac{\sum w_j^2 \cdot u_j^2}{\sum w_j \cdot u_j^2}. \qquad [12]$$

This can be seen also from our Table 3, where these two quantities are nearly the same.

We hypothesise that this relationship remains valid also in case of more complex longitudinal measures.

It's important to emphasise this point, because when we deal with longitudinal measures for which the linearised form u_j is not available, it is not possible to compute the design effect directly as the ratio of equation [3] and [4].

To summarise, the motivation of introducing (Kish_Jrr) is the following. The full expression u_i is available only in the context of the linearisation method. It may not even be possible (or at least not easy) to develop the required expression, for instance, for certain complex longitudinal measures of poverty. The simplified form u_{1i} is the well-known one for ratios. It is always available for application of the JRR method, even when the full linearised form u_i is not (or cannot be) developed. It is for this reason that we have subscripted the above quantities as "_Jrr".

Finally, the design effect is estimated as the product of the components defined above

$$deft = \left(\frac{se}{se\_rnd} \right) \cdot \left(Kish\_Jrr \right) \qquad [13]$$

Table 3

Standard errors and design effects for cross-sectional poverty rates: comparison of the JRR and Linearisation approaches.

Measure		**JRR**						
HCR based on 60% median	(1) est	(2) se	(3)=(2)/(1) %se	(4) se randomised	%se randomised	(5)=(2)/(4) Effect of clustering & stratification	(6)Kish jrr Effect of weighting	(5)*(6) DEFT
W1	20.4	0.77	3.77	0.64	3.14	1.20	1.31	1.57
W2	20.5	0.59	2.89	0.64	3.13	0.92	1.36	1.26
W3	20.2	0.63	3.13	0.55	2.73	1.15	1.30	1.49
W4	19.7	0.88	4.45	0.68	3.48	1.28	1.27	1.63
Mean	20.2	0.72	3.55	0.63	3.12	1.14	1.31	1.49

Measure		**Linearisation**							
HCR based on 60% median	(1) est	(2) se	(3)=(2)/(1) %se	(4) se_wt	(5) se_srs	(6)=(2)/(4) Effect of clustering & stratification	(4)/(5) kish_taylor Effect of weighting	(2)/(5) DEFT	Kish Factor
W1	20.4	0.69	3.38	0.57	0.44	1.22	1.30	1.59	1.26
W2	20.5	0.61	2.97	0.58	0.43	1.06	1.33	1.40	1.29
W3	20.2	0.59	2.91	0.54	0.43	1.08	1.25	1.35	1.23
W4	19.7	0.61	3.09	0.56	0.45	1.08	1.25	1.35	1.25
Mean	20.2	0.62	3.09	0.56	0.44	1.11	1.28	1.42	1.26

Table 3 demonstrates that, overall, the results from the two methodologies are very close for the estimates of standard errors and design effects of complex statistics such as cross-sectional poverty rates.

This close agreement also applies to the decomposition of design effects despite the different methodologies required for this decomposition in the two approaches[17].

The effect of weights is quite stable across the four waves.

4. Longitudinal aspects

This section considers the extension of the methodology described in the previous section to the longitudinal case.

4.1 Construction of the balanced panels

A balanced panel over an interval is defined to include only those individuals who were enumerated in the survey in all the waves during that interval. We also exclude from the balanced panel individuals for whom data on equivalised income are not available for any of the waves concerned.

As noted in Section 2, we constructed two balanced panels: (1) a two-year balanced panel containing individuals, irrespective of age, present in both waves 1 and 2; and (2) a four-year balanced panel containing individuals present in all the four waves 1 to 4. For all balanced panels the sample structure (PSUs, Strata etc.) is the same, exactly as defined above for the cross-sectional samples.

The balanced panel for the first two waves consists of 19,984 individuals, and contains variables common to the two waves (including sample structure and personal and household identifiers), equivalised income for the first and the second waves, base weights for the second wave, and a set of dichotomous indicators identifying the (yes=1/no=0) poverty status of the person in wave 1 (p1), in wave 2 (p2), in either wave (union of p1 and p2), and in both waves (intersection of p1 and p2).

The balanced panel for the first four waves contains information similar to the above: variables common to the four waves, equivalised income for each of the four waves, base weights for the wave 4, and a set of dichotomous indicators identifying the poverty status at each waves (p1-p4), any-time poverty (union of p1-p4), continuous poverty

[17] Again a clear outlier is the result on deft due to clustering and stratification with JRR in the case of wave 4. This is connected with the data problems referred to earlier (see also Annex 3).

(intersection of p1-p4), as well as indicators of persistent poverty (poverty in at least three out of four waves) and "Eurostat persistent poverty" (poverty in 4th wave, plus poverty in at least 2 of the other waves). This panel consists of 16,960 individuals.

For the longitudinal analysis we have used the base weight of the 2nd wave for the panel with two waves, and the base weight of the 4th wave for the panel with four waves.

Longitudinal measures of poverty

From these indicator variables all the measures of interest (any-time poverty, continuous poverty,…) can be constructed simply as the (weighted) average of the corresponding indicators codes as (poor=1, non-poor=ϕ).

Let $\left(p_1, p_2, ..., p_n\right)$ be the $\{\phi, 1\}$ indicator variables measuring poverty at micro level in n waves of a panel. Then the longitudinal measures of poverty for the individual are defined as:

Any-time poverty = $\max\left[p_1, p_2, ..., p_n\right]$;

Continuous poverty = $\min\left[p_1, p_2, ..., p_n\right]$.

Persistent poverty is defined as poverty during at least a majority of the years in a panel. In general, if there are T periods in a panel then persistent poverty means poverty in at least T' years where:

$$T' = \text{int}(T/2) + 1. \qquad [14]$$

In our case, Persistent poverty $\left[p_1, p_2, ..., p_n\right] = \delta\left((p_1 + p_2 + ... + p_n) \geq (\text{int}(n/2) + 1)\right)$, where $\delta = 1$ if its argument is true and zero otherwise. In case of a four year panel, persistent poverty would mean poor during at least 3 years.

Eurostat persistent poverty is defined as the poverty in the current period and at least two years in poverty during previous three years. It can be expressed by the following formula:

Eurostat persistent poverty=min[persistent poverty[p_1, p_2, p_3], p_4].

The longitudinal poverty measure corresponding to an indicator is, as mentioned above, simply the weighted average of the indicator.

It should be emphasised that once the poverty indicators are obtained at micro level the application of JRR method is straightforward. On the basis of the common structure applicable to all the waves in the balanced panel, Jackknife replications are created in the usual way and the longitudinal measure computed for each replication. These are then simply substituted into the standard Jackknife variance estimation formula.

4.2 Results

The results of Table 4 are obtained using JRR methodology. We have already presented the results for longitudinal measures in Section 2. Here we present, for completeness, the cross-sectional measures of poverty based on the two balanced panels for individual years. And, in the last row, the mean of the estimates for the four waves produced with the pure cross-sectional samples is presented for comparison.[18] We show the estimates of the HCR, standard errors, relative standard errors and the design effect, decomposed as described in the previous section.

Table 4

Measure	(1) est	(2) se	**(3)=(2)/(1) %se**	(4) randomised %se	(5)=(3)/(4) Effect of clustering & stratification	(6) Effect of weighting Kish jrr	**(5)*(6) Deft**
Panel 2 waves							
poverty p1	19.9	0.72	**3.62**	2.95	1.23	1.18	**1.45**
poverty p2	20.4	0.67	**3.28**	2.99	1.10	1.21	**1.33**
Mean	20.2	0.70	**3.45**	2.97	1.16	1.20	**1.39**
Panel 4 waves							
poverty p1	20.2	0.72	**3.55**	4.03	0.88	1.13	**1.00**
poverty p2	19.4	0.68	**3.52**	3.22	1.09	1.13	**1.23**
poverty p3	20.3	0.94	**4.63**	5.40	0.86	1.14	**0.98**
poverty p4	18.7	0.74	**3.97**	3.74	1.06	1.13	**1.20**
Mean	19.6	0.77	**3.92**	4.10	0.96	1.13	**1.10**
Mean of pure cross sectional est. for all the four waves (Table 3)	20.2	0.72	**3.55**	3.12	1.14	1.31	**1.49**

[18] This is the average of cross-sectional poverty rates W1-W4 presented in Table 3.

The table shows that the percentage standard errors are very similar and quite stable for all the measures. (The result for p3 in the four waves balanced panel appears to be an outlier). The design effect, as in Table 3, has been divided into two components: effect of clustering and stratification, and effect of weights. The effect of weights for each of balanced panels is of course stable, because the base weights used for a balanced panel are for any given unit the same for all waves in the panel. Also the effect of clustering and stratification within each balanced panel is quite stable, and of course so is the overall design effect.

Comparing the mean of the measures for any balanced panel with the mean of the pure cross-sectional ones, we find that they are quite close to each other. It is important to note that the cross-sectional poverty lines in the longitudinal analysis are not the same as that in the full cross-sectional samples. This is because for the former are defined for the population represented by the balanced panel, which is a subset of the population of any of the cross-sectional populations.

It is intriguing to note that the effect of weighting is much lower in the balanced panels than in the full cross-sectional samples, and among the former much lower for the 4-wave panel than for the 2-wave panel. Perhaps this results from some peculiarity of the weighting procedure adopted in ECHP, which we need not go into here.

5. Concluding remarks

In this paper, we have applied JRR methodology for the variance estimation of longitudinal poverty measures. The commonly used linearisation approach cannot be readily applied to longitudinal measures. The application of the JRR method to such complex measures is straightforward. The advantage comes from the fact that, once the sample structure (including the creation of Jackknife replications) is defined and measures of interest computed, the procedures for variance estimation of cross-sectional measures can be readily extended to longitudinal measures. Comparison of JRR results with those using linearisation methods showed generally very close agreement in the estimated standard errors and design effects as concerns the cross-sectional measures of poverty. This evidence provides added confidence in the extension of the JRR variance estimation methodology to longitudinal measures of poverty, even though no direct comparison with alternative approaches, as linearisation is presently available for such measures. With the JRR approach, the extension in itself is a straightforward task, once a common longitudinal structure of the sample and the Jackknife replications have been defined.

Annex 1

Computational algorithm for the construction of the Laeken indicator

At-persistent-risk-of-poverty rate (60% median), by gender and selected age groups (extract from a Eurostat document)

Definition

The ‘at-persistent-risk-of-poverty rate’ shows the percentage of the population living in households where the equivalised disposable income was below the 60% threshold for the current year and at least 2 out of the preceding 3 years. The population consists of all the persons that have been living for four years in private households.

Algorithm for the calculation

Linking information for four years

A file should contain for each person his/her equivalised disposable income for the four years.

Only persons that have been in the panel for all four waves should be included in the analysis. Therefore, all persons with missing values for at least one of the four EQ_INC variables are to be excluded[19].

Calculation of ‘at-risk-of-poverty thresholds’ for each year

For each of the four years, the ‘at-risk-of-poverty threshold’ is calculated in the following way:

- Firstly, persons have to be sorted according to their ‘equivalised disposable income’ (sorting order: lowest to highest value).
- Secondly, the median is calculated as explained for the current rate.
- Thirdly, the ‘at-risk-of-poverty threshold’ is calculated as 60% of the national median.

[19] EQ_INC variable defining equivalised disposable income of the household, ascribed to each of its members at the current wave.

Calculation of age-gender breakdowns

Each person is classified in the following categories according to his/her sex and age (in year 'T')

Total	Males	Females
$1 = \text{AGE} \geq 0$	$6 = \text{AGE} \geq 0$	$10 = \text{AGE} \geq 0$
$2 = 0 \leq \text{AGE} \leq 15$		
$3 = \text{AGE} \geq 16$	$7 = \text{AGE} \geq 16$	$11 = \text{AGE} \geq 16$
$4 = 16 \leq \text{AGE} \leq 64$	$8 = 16 \leq \text{AGE} \leq 64$	$12 = 16 \leq \text{AGE} \leq 64$
$5 = \text{AGE} \geq 65$	$9 = \text{AGE} \geq 65$	$13 = \text{AGE} \geq 65$

Calculation of the 'at-persistent-risk-of-poverty rate'

The 'at-persistent-risk-of-poverty rate' is calculated – for each age-gender category – as the percentage of persons with an 'equivalised disposable income' below the respective 'at-risk-of-poverty threshold' for the current year and at least 2 of the preceding 3 years.

The persons who are concerned by one of the following four cases have to be taken into account:

	T	T-1	T-2	T-3
1.	At risk of poverty	At risk of poverty	At risk of poverty	At risk of poverty
2.	At risk of poverty	At risk of poverty	**NOT** at risk of poverty	At risk of poverty
3.	At risk of poverty	At risk of poverty	At risk of poverty	**NOT** at risk of poverty
4.	At risk of poverty	**NOT** at risk of poverty	At risk of poverty	At risk of poverty

Thus:

$$\boxed{\text{At persistent risk of poverty rate} = \frac{\sum_{\text{All persons case1 or case2 or case3 or case4}} weights}{\sum_{\text{All persons } E\text{Q\_INC(T)} \neq . \text{ AND } E\text{Q\_INC(T-1)} \neq . \text{ AND } E\text{Q\_INC(T-2)} \neq . \text{ A}N\!D\ E\text{Q\_INC(T-3)} \neq .} weights}}$$

For this longitudinal at-risk-of-poverty rate, the base weight of the last wave is to be used.

Annex 2

Sample structure, computational structure, and construction of data files

This Annex gives a brief description of the sample structure as it has been implemented by Istat for the purpose of data collection in ECHP Italy. Then it describes how the required sample structure variables have been identified for each unit (household and persons) at each wave in the sample, how the computational units (PSUs and strata) are constructed and finally it describes the construction of the balanced panels.

Sample design

For the selection of the sample Istat has used a stratified two-stage sampling design with municipalities (singly or in groups) and communes serving as primary sampling units (PSU's). Large municipalities (23 in number) with population above a certain threshold were considered self-representing, i.e., taken into the sample automatically. Each of these effectively forms a separate stratum from which samples of addresses were selected directly in a single stage using systematic sampling from the population register.

The remaining PSU's were grouped according to municipality population size within regions. These groups within each region formed the final strata. Within each stratum a single PSU was selected with probability proportional to its population size (PPS sampling). Within each selected PSU, a systematic sample of addresses (households, families) was selected from the population register lists, so as to obtain an approximately self-weighting sample within each region.

Hence, the actual sample structure may be summarised in two parts as follows. The major part of the sample consists of a two-stage stratified sample. The primary strata are Italian Regions, each of which is then divided into finer strata of approximately uniform size according to the size group of municipalities in it. From each finer stratum, one municipality is selected as the PSU with probability proportional to its population size. The second part of the sample consists of large municipalities all of which are included in the sample automatically. In either case, a random sample of households is selected systematically from each municipality in the sample. Generally, this second stage selection probabilities are determined so as to obtain an approximately self-weighting sample of households within each region; however, the final data have been weighted using a complex procedure taking into account design probabilities, non-response and calibration to external controls.

The two parts of the initial sample as selected consisted of a total of 7,989 households, containing according to the registers a total of 24,063 individuals. Table A.1 shows the achieved cross-sectional sample sizes in terms of the number of households interviewed and the number of persons residing in those households. The numbers of households used in the present analysis were a little smaller than these numbers because of missing values on household income.

Table A.1. Cross-sectional sample sizes

ECHP Italy	Number interviewed		Number of households used in analysis
	households	persons	
Wave 1 (1994)	7,115	21,934	6,915
Wave 2 (1995)	7,128	21,757	7,004
Wave 3 (1996)	7,132	21,505	7,026
Wave 4 (1997)	6,713	20,074	6,627

Identification of the sample structure

The ECHP data set consists of four different types of data files. The first type (D-file) provides data on the addresses originally selected, irrespective of whether or not they were successfully enumerated in the survey. These data mostly pertain to design and implementation. In particular, they contain (at least some) information on structure of the sample.

The remaining three types of files provide substantive information at different levels: household (H) file covering all interviewed households; register (R) file with basic information on all members of the interviewed households; and personal (P) file covering all individuals aged 16 or over with completed detailed personal interview.

Sampling error computations require two types of information on the sample: (1) sample weights for the individual units; and (2) information on structure of the sample, specifically the identification of strata, the PSU's, and how the PSU's have been selected.

(1) Sample weights. Information on weights is required for any data analysis and is contained in the substantive files (H, R and P) for each individual record. In ECHP Users' Data Base (UDB), information is provided on the "cross-sectional" weight of each unit in these three files. For any given wave, a household and each of its members all have the same cross-sectional weight. These weights are of course wave-specific, and are used for cross-sectional analysis. For each individual person, ECHP-UDB also provides "base weights" which change for the person concerned from one wave to the next as a result of re-weighting of the sample to keep it representative of the population. There are no "true" longitudinal weights available in ECHP-UDB; however, base weights for the most recent period (the last wave) in the panel provide a close approximation to the weights required for longitudinal analysis.

Hence for the cross-sectional analysis we use the cross-sectional weights of each wave. For the longitudinal analysis we use the base-weights of individuals in the last wave considered in the panel.

(2) Sample structure. Unfortunately, sufficient information is not provided in ECHP-UDB for the identification of the sample structure, in particular, the identification of the stratum and PSU to which a household or person belongs, nor on how the PSU's have been selected.[20] Such information is available in the Production Data Base (ECHP-PDB), D-File, which is not available to the researcher outside the National Statistical Agency or Eurostat. In our case, thanks to the co-operation of Istat we have been able to use the ECHP-PDB that provides information about the structure of the sample.

But even here, the information available on the sample structure in D-file of the Italian ECHP-PDB is not complete, and cannot be easily connected to the available documentation on the survey.

In fact the final sample structure as we have used comes from D-file of wave 1, based not only on the UDB but also from its PDB version, as well as descriptions of the sample provided in various documents by Istat.[21]

After careful examination of the area codes, we were able to identify and separate out most (the largest 21 of the 23) self-representing PSU's. For the other (non-self-representing) PSU's in the sample, the identification codes also helped in ordering the PSU's (hence their strata) according to the municipality size group.[22] Though some elements of uncertainty remain, we believe that we have been able to specify the structure of the sample quite accurately.

Using D-file of Wave 1 of the ECHP-PDB, the above defines the sample structure for wave 1 of the panel. In relation to defining the sample for subsequent waves, the basic point is that in a household panel *the sample structure remains essentially the same over the survey waves*. The structure is defined by where and how the original sample of households and persons was selected at wave 1. Irrespective of any original sample persons moving to new locations over the life of the panel, each person's position in the structure of the sample (the stratum, PSU, SSU, etc., to which the unit belongs) remains unchanged, namely, as it was defined at the time of the original selection into the sample. The location in this structure of a household or a person in subsequent waves is determined

[20] For instance, information is required on whether PSU's have been selected systematically from a list ordered in some way, and if so, what is the order of selection of the PSU's in the sample.

[21] www.istat.it/dati/pubbsci/documenti/Documenti/doc_2004/2004_4.doc.

[22] Incidentally, for the non-self-representing units, we have assigned PSU equal to the value of the variable called "d01smst2" (primary strata) in the D-file.

by the unit's 'root household ID', which identifies the *original wave 1 household which the sample person(s) in the current household came from.* In this way, a common set of stratum and PSU identifiers can be defined for each household and person in any wave.[23]

Specification of computational units

Some aspects of this sample structure have to be redefined to make variance computation possible.

This is because there are certain basic requirements regarding the sample structure for the application of variance estimation procedures in the practical circumstances of large-scale complex surveys. The most basic one is that there must be at least two independent primary selections with replacement from each stratum of the sample. Some aspects of the actual sample structure often have to be redefined to meet these conditions and make variance computation possible. Redefinition of the structure is often also desirable for reasons of convenience and economy. (Of course, any such redefinition is appropriate only if it does not introduce significant bias in the variance estimation.) The computational structure can differ from the actual sample structure because of various considerations such as the following. *Note that such considerations apply equally irrespective of whether the JRR, Linearisation or some other form of variance computation algorithm is used.*

- It may be necessary to regroup ('collapse') strata so as to ensure that each stratum has at least two sample PSU's – the minimum number required for the computation of variance.
- Units which are included into the sample automatically ('self-representing units') are in fact strata rather than PSU's, and computational PSU's have to be defined at a lower stage within each such unit.
- In samples selected systematically, the implied implicit stratification is often used to define explicit strata, from each of which an independent sample is supposed to have been selected. Such strata need to be paired or otherwise grouped to ensure that each resulting computational stratum has at least two sample PSU's.
- In the case of direct (single stage) samples of ultimate units, it is often convenient to form random groupings of such units to construct computational PSU's.

[23] The household identification number of each person selected in the original (wave 1) sample is made up of two parts: the "root", and a wave-specific "split number". The root remains fixed for each such person and is assigned to whatever household the person moves to in any future wave. The sample structure variables, linked to this root number at wave 1, are therefore available for all households – and hence all persons – in all subsequent waves.

- Generally, grouping of small PSU's within and across strata, and grouping of strata to form fewer and larger computational units is common in practice.

In the present case, the computational units have been defined as follows.

Main part of the sample (non-self-representing PSU's)

In the main part of the sample, adjacent municipality size strata are paired to provide computational strata, each of which contains two PSU's – that being the minimum number required for variance computation.[24] Also, it is assumed that the two primary selections within each (computational) stratum are independent and with replacement. The original explicit strata, each containing only one sample PSU, were ordered according to the municipality size groups they represented. It is important to note that the pairing (or other grouping) of sample PSU's for the purpose of constructing computational strata always involved *adjacent units* according to this ordering. In this way the effect of actual stratification was retained to the maximum extent possible in the structure used for variance computations.

With this process we obtained a total of 88 strata and 187 PSUs for this part of the sample.

Self-representing municipalities

In the part consisting of 'self-representing' municipalities, each municipality in fact forms a stratum. The general procedure involved two steps. (1) Within each municipality sample households, in their original order according to Household Identification (HID) number, were grouped into some appropriate number of strata. (The idea was to retain any effect of the implicit stratification provided by systematic sampling of households.) (2) Random groupings of sample households within each such stratum formed the computational PSU's. Note that such grouping is entirely random, and hence does not in principle affect the variance of the sample. It is introduced only for computational convenience.

In most cases, step (1) was not involved: the municipality was treated as a single stratum, and two random groupings of households were created to define computational PSU's. However, more than one strata (and twice as many PSU's) were created in the six largest cities so as to keep the number of sample households per PSU approximately constant: in Genova, Palermo and Torino two strata each; in Napoli and Milano three strata each; and in Roma six strata. Within each stratum we randomised the households and then assigned them to one of the two PSUs.

[24] In the case of a region containing an odd number of PSU's, one of the computational strata was composed of 3 PSU's so as not to cut across regional boundaries.

With this process we obtained a total of 33 strata and 66 PSUs for this part of the sample. Thus the whole resulting sample consists of 121 computational strata and 253 computational PSU's.

This common structure applies to all the survey waves.[25] This structure constructed as defined above is attached to the each H-file of the four waves. The cross-sectional analysis uses data from the household (H) files of four waves (1994-1997).[26]

Construction of the balanced panels

The following describes the procedure of constructing the balanced panel data files.

For each of the four waves we merge the register (R) file containing all individual persons in the enumerated households, and the household (H) file. The objective is to create a new register file (say, R+), with one record per person, containing additional H-file variables such as equivalised income, as well as the variables defining the sample structure. Next, we merge all the four waves. The resulting file contains all individuals ever enumerated in any of the 4 waves. This is the so-called "unbalanced panel" (this file contained 24,033 individuals). From this last data set we can construct balanced panels. A balanced panel over an interval is defined to include only those individuals who were enumerated in the survey in all the waves during that interval. We also exclude from the balanced panel individuals for whom data on equivalised income are not available for all the waves concerned.

In this manner we constructed two balanced panels: (1) a two-year balanced panel containing individuals present in both waves 1 and 2; and (2) a four-year balanced panel containing individuals present in all the four waves 1 to 4. For all balanced panels the

[25] A minor exception to the common sample structure across waves arose in wave 4. Because of sample attrition, one of the PSU's in this wave turns out not to contain any completed interviews. In order to ensure that every computational stratum contains at least two PSU's, the *non-empty* PSU of the stratum that contains this empty PSU is moved to the preceding computational stratum, resulting in a total of 120 strata and 252 PSU's in wave 4 sample.

[26] The unit in income distribution and poverty analysis is the individual person. For convenience, however, the cross-sectional analysis can be conducted using the smaller H-file. This is possible for the following reasons. (1) Equivalised income is identical for a household and each of its members; hence members of a household are classified in the same way, all as poor or all as non-poor. (2) The cross-sectional weight is identical for a household and each of its members; hence with the product of this weight and household size taken as the weight, a household (in the H-file) represents (the weighted number of) all its members (in the R-file) in the cross-sectional sample for the wave concerned.

sample structure (PSUs, Strata etc.) is the same, exactly as defined above for the cross-sectional samples.[27]

The balanced panel for the first two waves consists of 19,984 individuals, and contains variables common to the two waves (including sample structure and personal and household identifiers), equivalised income for the first and the second waves, base weights for the second wave, and a set of dichotomous indicators identifying the (yes=1/no=0) poverty status of the person in wave 1 (p1), in wave 2 (p2), in either wave (union of p1 and p2), and in both waves (intersection of p1 and p2).

The balanced panel for the first four waves contains information similar to the above: variables common to the four waves, equivalised income for each of the four waves, base weights for the Wave 4, and a set of dichotomous indicators identifying the poverty status at each waves (p1-p4), any-time poverty (union of p1-p4), continuous poverty (intersection of p1-p4), as well as indicators of persistent poverty (poverty in at least three out of four waves) and "Eurostat persistent poverty" (poverty in 4th wave, plus poverty in at least 2 of the other waves). This panel consists of 16,960 individuals.

Technical notes.

(1) We emphasise that for the longitudinal analysis we have used the base weight of the 2nd wave for the panel with two waves and the base weight of the 4th wave for the panel with four waves.

(2) The indices p1 and p2 for the two-wave panel are defined with respect to the population included in that two-wave balanced panel. Similarly, the indices p1-p4 for the four-wave panel are defined with respect to the population included in that four-wave balanced panel, which is a subset of the above. This is done so as to be consistent with Eurostat definition of "persistent at-risk-of poverty rates". This means that 'p1' and 'p2' for the two panels are not identical quantities. And p1-p4 differs from W1-W4 used in this Report to indicate poverty rates computed on the full cross-sectional sample for each waves 1 to 4.

[27] For the waves 1-2 balanced panel, the sample structure is the structure of first wave, meaning that the PSUs and strata of this balanced panel derive from the 1st wave. This contains 253 (computational) PSU form 121 (computational) strata. For the four-wave balanced panel, as noted, sample attrition lead to an empty PSU in the 4th wave. We assigned the non-empty PSU of the stratum containing the empty PSU to the preceding stratum. And then this structure of the 4th wave, with one PSU and a stratum less, is assigned to all the other (1st, 2nd and 3rd) waves in the four-waves balanced panel. Therefore, unlike the two wave panel, in the four wave panel we have 252 PSUs and 120 strata.

Annex 3: Data problems

Reported income distribution

Inspecting the household income reveals that there are many imputed values, and more importantly, that there are *blocks of households for which equivalised income is exactly the same.* This appears highly implausible, and must be the result of some undesirable features of the imputation procedures used in the survey. As a consequence, all these identical values of equivalised income generate a very un-smooth income distribution that takes a step-like form. This step-like distribution of equivalised income causes instability in the determination of the median and the poverty line if many households happen to be concentrated at or very near either of those values. The estimated poverty rate is affected. However, the effect tends to be much greater on the estimate of variance, particularly using the JRR method.

This problem appears to be particularly marked in the data for Wave 4 of Italian ECHP, resulting in rather variable and unsatisfactory results when the data from that wave are involved. There is also a particularly large sample attrition between waves 3 and 4.

Extreme values of sample weights

The household cross-sectional weights in the given data set contain some extreme values. For example, in wave 2 the maximum weight coded is 37.16 and the minimum 0.077. This means that the largest weight is more than 450 times bigger than the smallest. As is recommended for practical work, we considered it desirable to trim the extreme values as a precaution against excessive instability in the numerical results. The following simple procedure, adopted after an examination of the distribution of the weights, has been used. Weights above 99% of the weight distribution are replaced by the value of weight at the 99th percentile, and weights at the bottom 1% are replaced by the value of weight at the 1st percentile.

Extreme values of variance of poverty measure

For the calculation of JRR standard errors, extreme values of contributions of individual Jackknife replications to total variance estimate have been trimmed, so as to ensure that no single replication contributes disproportionately to the total estimate of variance. This precaution is desirable as a protection against the presence of large irregularities or outliers in the empirical data used.

Let u be a full-sample estimate of any complexity, and $u_{(hi)}$ be the estimate for replication (hi) produced using the same procedure after eliminating primary unit (PSU) i in stratum h and increasing the weight of the remaining (a_h-1) units in the stratum by an appropriate factor. Let $u_{(h)}$ be the simple average of the $u_{(hi)}$ over the a_h sample units in h. As noted in Section 3, the JRR variance of u is then estimated as:

$$\text{var}\,(u) = \Sigma_h \left[(1 - f_h). \frac{a_h - 1}{a_h} . \Sigma_i \left(u_{(hi)} - u_{(h)} \right)^2 \right].$$

We have trimmed for each replication (hi) the quantity $\left(u_{(hi)} - u_{(h)}\right)^2$, if this was higher than 6 times its mean value over all replications. That is, any value exceeding 6 times the average is revised to equal 6 times the mean of the quantity $\left(u_{(hi)} - u_{(h)}\right)^2$ over all the replications. We decided to trim the replication variance contributions because the results without trimming could be subject to instability, because of irregularities and outliers in the data.

References

Bane M. J., Elwood D. T. (1986), Slipping into and out of poverty: the dynamics of spells, Journal of Human Resources 21.

Deming W. E. (1943), Statistical adjustment of data, New York: Wiley.

Durbin J. (1959), A note on the application of Quenouille's method of bias reduction to the estimation of ratios, Biometrika 46.

Efront B., Stein C. (1981), The Jackknife estimate of variance, Annals of Statistics 9.

Eurostat (2003a), ECHP UDB Description of variables, Doc. Pan 166/2003-12, Luxembourg.

Eurostat (2003b), ECHP UDB Construction of variables, Doc. Pan 167/2003-12, Luxembourg.

Eurostat (2003c), ECHP UDB manual, Doc. Pan 168/2003-12, Luxembourg.

Hills J. (1998), Does income mobility mean that we do not need to worry about poverty?, in A. B. Atkinson and J. Hills (eds.), Exclusion, employment and opportunity, CASE paper 4, London, Centre for Analysis of Social Exclusion, London School of Economics.

Jenkins S. P. (2000), Modelling household income dynamics, Journal of Population Economics 13.

Kendall M. G., Stuart, A. (1958), The advanced theory of statistics, Vol. I, London: Charles Griffin.

Keyfitz N. (1957), Estimation of sampling variance where two units are selected from each stratum, Journal of the American Statistical Association 52.

Kish L., Frankel M. (1974), Inference from complex sample, Journal of Royal Statistical Society B/36.

Lillard L. A., Willis R. J. (1978), Dynamic aspect of earning mobility, Econometrica 46.

Rodgers J. L., Rodgers J. R. (1993), Chronic poverty in the United States, Journal of Human Resources 28.

Verma V., Betti G. (2005), Sampling Errors for Measures of Inequality and Poverty. Invited paper in Classification and Data Analysis 2005 - Book of Short Papers, pp. 175-179, CLADAG, Parma, 6-8 June 2005

Verma V., Betti G., Pierozzi F., Gagliardi F., Ballini F. (2006), Alternative variance estimation procedures (1) Application to cross-sectional measures, Report to Eurostat, Luxembourg.

Verma V., Clemenceau A. (1996), Methodology of the European Community Household Panel, Statistics in Transition 2.