SAMPLE Seminar, Warsaw 24 March 2010

# Cumulation and pooling to improve sampling precision

Vijay Verma, Francesca Gagliardi

**Siena Team for Sample project:**
Achille Lemmi, Vijay Verma, Gianni Betti, Francesca Gagliardi, Caterina Ferretti

**Aspects presented:**

I. **Cumulation**
II. **Aspects of pooling**
III. **JRR for variance estimation**
IV. **Fuzzy measures of poverty and deprivation**

# I. Cumulation

## 1 Modes of using survey data

Policy research requires indicators of poverty and social exclusion *disaggregated to lower levels and smaller subpopulations*. Direct, one-time estimates from surveys are normally insufficiently precise for the purpose, particularly concerning poverty and social exclusion which involve complex distributional statistics based on relatively small-scale surveys.

Survey data can be used in different forms or manners to construct regional indicators.

(1) Direct estimates from survey data – provided regional sample sizes are adequate for the purpose.

(2) Constructing alternative (but similar) indicators utilising available data more intensively.

(3) *Cumulation of data over survey waves to increase precision of the direct estimates.*

(4) Using survey data in conjunction with other sources using small area estimation techniques.

(5) Going altogether beyond the survey by exploiting other sources.

In the context of SAMPLE project, we here address (3) –estimates for subnational regions in EU countries through the cumulation of data over time.

## 2 Gain in precision from cumulation over waves

Consider poverty rates averaged over a number of consecutive waves.

The issue is to quantify the gain in sampling precision from such pooling, given that data from different waves of a rotational panel are highly correlated. For this purpose, standard variance estimation methodology such as JRR can be easily extended, and I will briefly return to that. It is more illuminating to provide here a simplified procedure for quantifying the gain in precision from averaging over two waves. With $p_j$ and $p'_j$ the (1, 0) indicators of poverty of individual j over the two adjacent waves, we have:

$$\mathrm{var}(p_j) = \Sigma(p_j - p)^2 = p.(1-p) = v$$

similarly,

$$\mathrm{var}(p'_j) = p'.(1-p') = v'$$

$$\mathrm{cov}(p_j, p'_j) = \Sigma(p_j - p).(p'_j - p') = a - p.p' = c_1$$

where ‘a’ is the persistent poverty rate over the two years

For the simple case where the two waves completely overlap and $p' = p$

variance $v_A$ for the averaged measure is: $v_A = \frac{v}{2}.(1+b)$

with correlation $b = \left(\frac{c_1}{v}\right) = \left(\frac{a - p^2}{p - p^2}\right)$

The correlation between two periods is expected to decline as the two become more widely separated, and the model can be easily extended to the general case.

For application to pairs of waves in EU-SILC, allowing for variations in cross-sectional sample sizes and partial overlaps, we have:

$$V = \frac{(V_1 + V_2)}{4}.\left(1 + b.\left(\frac{n}{n_H}\right)\right)$$

where $V_1$ and $V_2$ are the sampling variances, n is the overlap between the cross-sectional samples, and $n_H$ is the harmonic mean of their sample sizes $n_1$ and $n_2$.

**Table 1: Gain from cumulation over two waves: cross-sectional and persistent poverty rates. Illustration: Poland EU-SILC 2005-2006**

| Sample base | Poverty rate | Est | n persons | %se* actual | | mean income | HCR: poverty line national | HCR: poverty line regional |
|---|---|---|---|---|---|---|---|---|
| **CS-2006** | HCR 2006 | 19.1 | 45,122 | 0.51 | **(1)** | 0.42 | 0.34 | 0.40 |
| **CS-2005** | HCR 2005 | 20.6 | 49,044 | 0.45 | **(2)** | 1.31 | 1.18 | 1.18 |
| **LG 05-06** | HCR 2006 | 18.5 | 32,820 | | **(3)** | 0.55 | 0.40 | 0.47 |
| **LG 05-06** | HCR 2005 | 20.2 | 32,820 | | **(4)** | 0.60 | 0.48 | 0.56 |
| **LG 05-06** | Persistent '05-06 | 12.5 | 32,820 | | **(5)** | 14% | 30% | 30% |

Rows (1)-(5) of the table are as follows.

| | |
|---|---|
| Standard error of average HCR over two years (assuming independent samples) | $(1) = \frac{1}{2}.(V_1 + V_2)^{1/2}$ |
| Factor by which standard error is increased due to positive correlation between waves | $(2) = .\left(1 + b.\left(\frac{n}{n_H}\right)\right)^{1/2}$ |
| Standard error of average HCR over two years (given correlated samples) | $(3) = (1).(2) = .(V)^{1/2}$ |
| Average standard error over a single year | $(4) = (V_1)^{1/2} + (V_2)^{1/2}$ |
| Average gain in precision (variance reduction, or increase in effective sample size, over a single year sample) | $(5) = 1 - \left(\frac{(3)}{(4)}\right)^2$ |

## 3 Design effect and its components

A most useful concept concerns ‘design effect’ - the ratio of variance (v) under the given sample design, to variance ($v_0$) under a simple random sample of the same size:

$$d^2 = v/v_0\,, \quad d = se/se_0$$

Proceeding from sampling error to design effects is essential for understanding patterns of variation and determinants of sampling error, for smoothing and extrapolating results of computations, and for evaluating performance of the sampling design.

In applications for EU-SILC, there is a special reason for decomposing the design effect. With limited information on sample structure included in the available micro-data, computation of variances cannot be done in many cases.

Decomposition of variances and design effects identifies more ‘portable’ components, which are more easily carried over from a situation where they can be computed, to another situation where such direct computations are not possible. Thus we can at least partly overcome the problem due to lack of information on sample structure.

We may decompose total variance v (for the actual design) into the components as

$$v = v_0 . d^2 = v_0 . \left( d_W . d_H . d_D . d_X \right)^2$$

where

$d_W$ is the effect of sample weights,

$d_H$ of clustering of individual persons into households,

$d_D$ of clustering of households into dwellings, and

$d_X$ that of other complexities of the design, mainly clustering and stratification.

All factors other than $d_x$ do not involve clusters or strata, but depend only on individual elements in the sample. Parameter $d_W$ depends on variability of sample weights, and secondly also on the correlation between the weights and the variable being estimated; $d_H$ is determined by the size or the number of relevant individuals in the household, and similarly $d_D$ by the number of households per dwelling in a sample of the latter.

By contrast, factor $d_X$ represents the effect of various complexities of the design. Hence unlike other components, $d_X$ requires information on the sample structure (clustering and stratification).

**Table 2: Estimation of variance and design effects at the national level. Illustration: Poland EU-SILC 2006 cross-sectional sample.**
(a two stage stratified sample of dwellings containing 45,122 individual persons)

| | Estimate | %se actual | Design effect $d_X$ | $d_W$ | $d_H$ | d | %se SRS |
|---|---|---|---|---|---|---|---|
| (1) | 3,704 | 0.57 | 0.94 | 1.22 | 1.74 | 1.99 | 0.29 |
| (2) | 19.1 | 0.51 | 1.02 | 1.09 | 1.74 | 1.94 | 0.26 |
| (3) | 19.0 | 0.61 | 1.05 | 1.09 | 1.74 | 1.99 | 0.30 |

(1) Mean equivalised disposable income

(2) HCR – 'head count' or poverty rate, using national poverty line

(3) HCR – 'head count' or poverty rate, using regional (NUTS1) poverty line

"%se": for mean statistics e.g. equivalised disposable income – expressed as percentage of the mean value; for proportions and rates (e.g. poverty rates) – given as absolute percent points.

## 4 Extrapolating to regional estimates

As we go from national to regional level, all the values, except "%se SRS" and $d_X$, are computed at regional level in the same manner as the national level. All factors other than $d_X$ do not involve clusters or strata, but essentially depend only on individual elements. Hence normally they are well estimated, even for quite small regions.

The quantity (%se* SRS) can be directly computed at the regional level, just as for the national level. However, very good approximation can be usually obtained very simply without involving JRR computations of variance.

1. For means (such as equivalised income) over very similar populations, assumption of a constant coefficient of variation is reasonable, giving:

$$(\%\mathrm{se} * \mathrm{SRS})^2_{(\mathrm{G})} = (\%\mathrm{se} * \mathrm{SRS})^2_{(\mathrm{C})} . \left(\mathrm{n}_{(\mathrm{C})} / \mathrm{n}_{(\mathrm{G})}\right)$$

2. For proportions (p, with q=100-p), with standard error expressed in absolute percent points, we can take:

$$(\%se * SRS)^2_{(G)} = (\%se * SRS)^2_{(C)} . \left(\frac{p_{(G)} . q_{(G)}}{p_{(C)} . q_{(C)}}\right) . \left(n_{(C)} / n_{(G)}\right)$$

A poverty rate may be treated as proportions for the purpose of applying the above.

Factor $d_{X(G)}$

for a region (G) may be estimated in relation to $d_{X(C)}$

estimated at the country (C) level on the following lines.

1. For large regions, each with a large enough number of PSUs (say over 25 or 30), we may estimate the variance and hence $d_{X(G)}$

directly at the regional level.

2. Sometimes a region involves a SRS of elements, even if the national sample is multi-stage in other parts; here obviously, $d_{X(G)} = 1$

If the sample design in the region is the same or very similar to that for the country as a whole – which is quite often the case – we can take

$$d_{X(G)} = d_{X(C)}$$

It is common that the main difference between the regional and the total samples is the average cluster size (b). In this case we use

$$d^2_{X(G)} = 1 + \left(d^2_{X(C)} - 1\right)\frac{b_{(G)}}{b_{(C)}}$$

If $d_{X(C)} \geq 1$

the above equation is replaced by $d_{X(G)} = d_{X(C)}$

## 5 Gain in precision from averaging over correlated samples for two consecutive waves. Poland NUTS1 regions

**Mean equivalised income**

| | Country | PL1 | PL2 | PL3 | PL4 | PL5 | PL6 |
|---|---|---|---|---|---|---|---|
| (1) | 0.42 | 0.94 | 0.83 | 0.92 | 1.15 | 1.28 | 1.07 |
| (2) | 1.31 | 1.33 | 1.30 | 1.31 | 1.27 | 1.32 | 1.32 |
| (3) | 0.55 | 1.26 | 1.08 | 1.20 | 1.47 | 1.70 | 1.41 |
| (4) | 0.60 | 1.33 | 1.17 | 1.30 | 1.62 | 1.81 | 1.51 |
| (5) | **14%** | 11% | 15% | 14% | 18% | 12% | 12% |

**HCR national poverty line**

| | Country | PL1 | PL2 | PL3 | PL4 | PL5 | PL6 |
|---|---|---|---|---|---|---|---|
| (1) | 0.34 | 0.70 | 0.65 | 0.88 | 0.89 | 1.06 | 0.94 |
| (2) | 1.18 | 1.18 | 1.17 | 1.18 | 1.18 | 1.17 | 1.19 |
| (3) | 0.40 | 0.83 | 0.76 | 1.03 | 1.05 | 1.23 | 1.12 |
| (4) | 0.48 | 0.99 | 0.92 | 1.24 | 1.26 | 1.50 | 1.33 |
| (5) | **30%** | 29% | 31% | 30% | 30% | 32% | 29% |

**HCR regional poverty line**

| | Country | PL1 | PL2 | PL3 | PL4 | PL5 | PL6 |
|---|---|---|---|---|---|---|---|
| (1) | 0.40 | 0.86 | 0.83 | 0.94 | 1.03 | 1.29 | 1.05 |
| (2) | 1.18 | 1.18 | 1.18 | 1.17 | 1.18 | 1.17 | 1.18 |
| (3) | 0.47 | 1.02 | 0.98 | 1.10 | 1.21 | 1.51 | 1.24 |
| (4) | 0.56 | 1.21 | 1.16 | 1.33 | 1.45 | 1.82 | 1.49 |
| (5) | **30%** | 29% | 29% | 31% | 30% | 31% | 31% |

*Rows (1) – (5) have been defined in Table 1.*

For mean equivalised income, generally the coefficient of correlation between consecutive waves exceeds 0.70 – hence much smaller gain from cumulation.

# II. Aspects of pooling

Objectives

(1) Cumulation or aggregation in order to obtain more precise estimates, albeit normally with some loss of detail.

(2) Comparisons of trends and differences across populations and times.

(3) Meeting the more general and broader objective of *common interpretation* of statistical information from different sources and/or for different populations in relation to each other and against common standards.

Prerequisite: comparability

Diverse scenarios

| | *Data source* | |
|---|---|---|
| *Population* | same/similar (s) | different /dissimilar (d) |
| Same/Similar (S) | S.s | S.d |
| Different /Dissimilar (D) | D.s | D.d |

## Pooling of data versus pooling of estimates

We may distinguish between *pooling of data*, i.e. aggregation of micro-level data for the same or different populations, surveys and times, on the one hand, and the *pooling of estimates*, i.e. the production of a common estimate as a function of estimates produced from individual data sources.

Let us consider estimate $\phi_i$

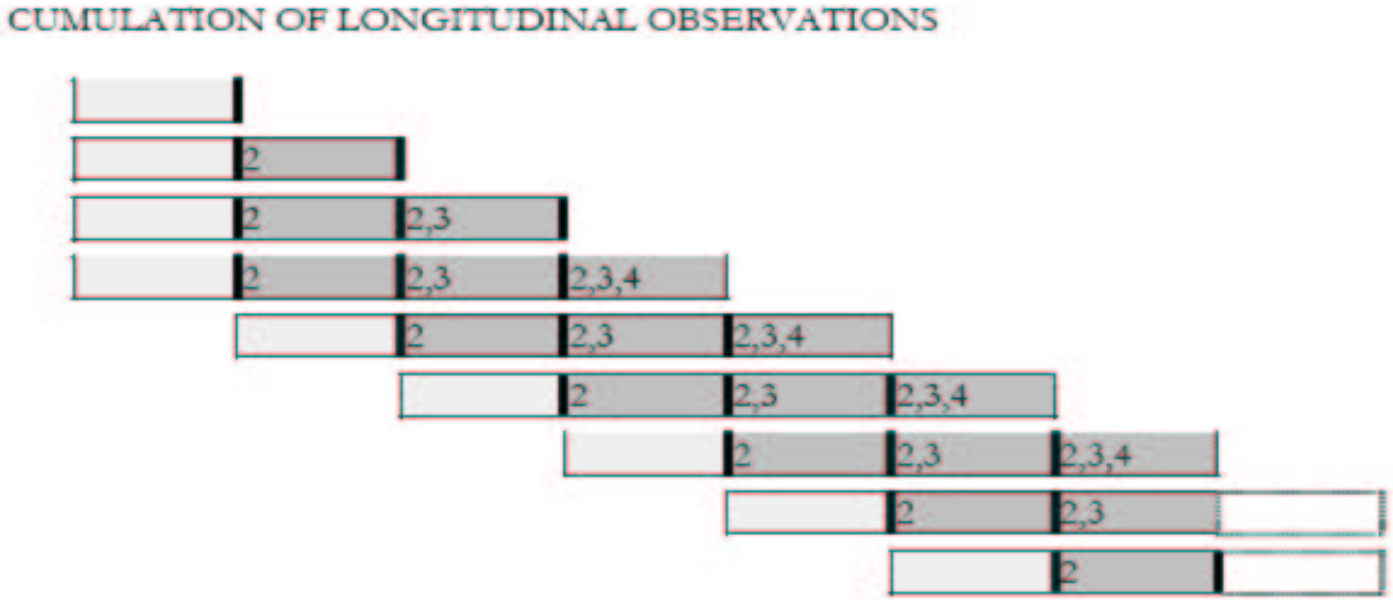
for a certain statistic for country *i*. In comparisons, each $\phi_i$

of course receives the same weight. For estimates aggregated over EU countries, of the form $\phi = \Sigma P_i . \phi_i$

the most common practice by far is to take the weights $P_i$ in proportion to the countries' population size, thus producing statistics for the 'average EU citizen'.
By contrast, in much policy debate, it is the situation in the 'average EU country' that is of interest; this amounts to taking the $P_i$ values as equal.

Whatever the choice of $P_i$, the above formulation involves pooling country-level estimates. Given standardised data sets from all countries, pooling at the micro-level is also possible, with unit weights $w_{ij}$ scaled as $w'_{ij} = w_{ij} . \left( P_i / \Sigma w_{ij} \right)$

## Panels in a rotational design




CUMULATIVE NUMBER OF LONGITUDINAL OBSERVATIONS

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1. SUBSAMPLES PROVIDING CROSS-SECTIONAL DATA | | | | | | | | |
| 4 | 8 | 12 | 16 | 20 | 24 | 28 | ..... | 4*Y |
| 2. SUBSAMPLES PROVIDING YEAR-TO-YEAR TRANSITIONS | | | | | | | | |
| 0 | 3 | 6 | 9 | 12 | 15 | 18 | ..... | 3*(Y-1) |
| 3. SUBSAMPLES PROVIDING 3 YEARS LONGITUDINAL OBSERVATIONS | | | | | | | | |
| 0 | 0 | 2 | 4 | 6 | 8 | 10 | ..... | 2*(Y-2) |
| 4. SUBSAMPLES PROVIDING 4 YEARS LONGITUDINAL OBSERVATIONS | | | | | | | | |
| 0 | 0 | 0 | 1 | 2 | 3 | 4 | ..... | (Y-3) |
| SURVEY YEAR | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | ..... | Y |

*Note:* The numbers in the cells of the diagram indicate the type(s) of observations provided by the subsample. The above numbers may be multiplied by the subsamples size to obtain the cumulated number of observations.

**Reduction in variance by pooling data for subsamples**

Variance decreases in inverse proportion to sample size, provided that the subsamples making up the total sample are *independent* (e.g., if in EU-SILC each subsample is based on a different set of clusters ).

In computing measures for the cross-section, the common practice is simply to pool the cases from the subsamples. This amounts to giving each subsample a 'weight' in proportion to its sample size, i.e. inversely proportion to its expected variance, which is an efficient (optimal) procedure in the absence of bias. Consider two panels with same variance $V^2$, but the second (older) one also subject to bias B due to non-response. Pooling them with weights $W_1, W_2$ respectively ($W_1+W_2=1$) gives MSE composed of

variance $V^2.\left(W_1^2 + W_2^2\right)$ and bias$^2$ $W_2^2.B^2$

Variance is minimised with $W_1=W_2=0.5$, but bias can be reduced by taking $W_2<0.5$, i.e. giving less weight to the older panel.
The optimal choice of the weights depends on the bias ratio $B^2/V^2$

**Reduction of variance from averaging different poverty thresholds**

Consider three poverty line thresholds, with poverty rates

$$p_1 < p_2 < p_3$$

with fixed weights $W_i$, the final rate is computed as

$$p = \Sigma_i W_i . p_i$$

For simplicity, take the sample as SRS and approximate the complex statistic 'poverty rate' as an ordinary proportion. Its variance is given by

$$\mathrm{var}(p) = \Sigma_i W_i^2 . \mathrm{var}(p_i) + 2.\Sigma_{j<i} W_i W_j . \mathrm{cov}(p_i, p_j)$$

**Reduction from averaging over rounds in a rotational design**

Consider a rotational sample in which each unit stays in the sample for n consecutive periods, with the required estimate being the average over Q consecutive periods. The case n=1 corresponds simply to independent samples each quarter. Under the simplifying assumption of uniform variances, variance of the estimate of average over Q period is

$$V_a^2 = V^2 / Q$$

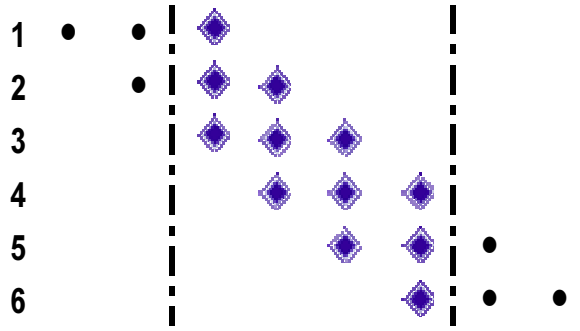
The total sample involved in the estimation consists of (n+Q-1) independent subsamples. Each subsample is 'observed' over a certain number of consecutive periods within the interval (Q) of interest. In principle, for a given subsample the sample cases involved in these 'observations' are fully overlapping. For 'observation' we mean surveying one subsample on one occasion. The distribution of the (n+Q-1) subsamples according to the number of observation (m) provided is:

| *No. of observations (m)* → | *provided by no. (x) of subsamples* | *Total no. of 'observations' provided by all subsamples* |
|---|---|---|
| $m = 1, 2, \ldots, (m_1\text{-}1)$ | $x = 2$ *for each value of m* | $\sum_{i=1}^{(m_1-1)} 2i = (m_1 - 1) \cdot m_1$ |
| $m = m_1$ | $x = m_2\text{-}(m_1\text{-}1)$ | $m_1 \cdot m_2 - (m_1 - 1) \cdot m_1$ |
| *Total* → | *no. of subsamples equal to* $2 \cdot (m_1 - 1) + m_2 - (m_1 - 1) = = m_2 + m_1 - 1 = n + Q - 1$ | *no. of observations equal to* $m_1 \cdot m_2 = n.Q$ |

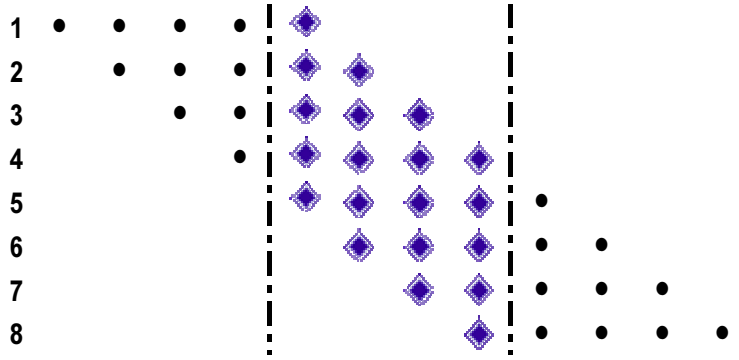
where $m_1$=min(n, Q) and $m_2$=max(n, Q).

***Q=4***

***n=3*** *('observations' provided=3*4=12)*




***n=5*** *('observations' provided=5*4=20)*




*Note: The numbers on the left side of the figures represent the number of subsamples (n+Q-1).*

For illustration, consider $Q=m_1=4$, $n=m_2=5$.
There are 2 contributing subsamples for each number 1, 2 and (m1-1)=3 of observations; and in addition there are $m_2-(m_1-1)=2$ subsamples, each contributing $m_1=4$ observations.
Similarly, for $Q=m_2=4$, $n=m_1=3$, we have 2 contributing subsamples for each number 1 and $(m_1-1)=2$ of observations, and in addition $m_2-(m_1-1)=2$ subsamples each contributing $m_1=3$ observations.

In order to provide a simplified formulation of the effect of correlation arising from sample overlaps, we assume the following model. If R is the average correlation between estimates from overlapping samples in adjacent periods (as defined above), then between points one period apart (e.g. between the 1st and 3rd quarters), the average correlations is reduced to $R^2$, the correlation between points two periods apart (e.g. the 1st and the 4th quarters) is reduced to $R^3$, and so on.

Consider a subsample contributing m observations during the interval (Q) of interest with full sample overlap. Considering all the pairs of observations involved and the correlations between them under the method assumed above, variance of the average over the m observations is given by

$$V_m^2 = \frac{V^2}{m} \cdot \left(1 + f(m)\right)$$

Where

$$f(m) = \frac{2}{m} \cdot \left\{(m-1) \cdot R + (m-2) \cdot R^2 + \ldots + R^{m-1}\right\}$$

The term $V_m^2 \Big/ \left(\frac{V^2}{m}\right) = 1 + f(m)$ reflects the loss in efficiency in cumulation or averaging over overlapping samples, compared to cumulation over entirely independent samples.

For various values of m:

| m | $f(m)$ |
|---|---|
| 2 | R |
| 3 | $\frac{2}{3}(2R + R^2)$ |
| 4 | $\frac{2}{4}(3R + 2R^2 + R^3)$ |
| 5 | $\frac{2}{5}(4R + 3R^2 + 2R^3 + R)$ |

In estimating the average using the whole available sample of $(n \cdot Q)$

subsample observations, we may simply give each observation the same weight.

Taking into account the number of observations and the variances involved, the resulting variance of the average becomes

$$V_a^2 = \left(\frac{V^2}{n \cdot Q}\right) \cdot \left\{ m_1 \cdot [m_2 - (m_1 - 1)] \cdot [1 + f(m_1)] + 2 \sum_{m=1}^{m_1 - 1} m \cdot [1 + f(m)] \right\} \Big/ (n \cdot Q) = \left(\frac{V^2}{n \cdot Q}\right) \cdot F(R)$$

## Concluding remark: objectives of pooling

It may be argued that averaging and similar ‘manipulation’ is not acceptable, or at least that it introduces bias, since it *alters* the measures we obtain. This may be true in a literal sense but this is not a sensible objection in many situations. We need a pragmatic and not an ideological approach to statistics. All statistical measures are constructed for the purpose of conveying some meaning, for providing some interpretation to real and complex situations. The particular forms of measures chosen are always determined by considerations of usefulness and practicality, are always compromises and in themselves not ‘sacred’.

The objectives of pooling include searching for measures which convey essentially the same information but in a *more robust* manner, reducing random variability or noise.

A related objective of pooling is *trading dimensions* – gaining in some more needed directions by losing something less needed for particular purposes – such as permitting more detailed geographical breakdown but with less temporal detail.

A third objective is to *summarise* over different dimensions, providing more consolidated and fewer indicators. Such indicators are of course different from the more numerous ‘raw’ indicators, but are often more, or at least equally, meaningful and useful.

# III. JRR FOR VARIANCE ESTIMATION

## 1 The standard JRR procedure for variance estimation

Jackknife Repeated Replication (JRR) provides a versatile technique for variance estimation. It is one of the classes of variance estimation methods based on comparisons among replications generated through repeated re-sampling of the same parent sample. Each replication needs to be a representative sample in itself and to reflect the full complexity of the parent sample.

The JRR variance estimates take into account the effect on variance of aspects of the estimation process which are allowed to vary from one replication to another. In principle this can include complex effects such as those of imputation and weighting. But often in practice it is not possible to repeat such operations entirely fresh at each replication.

A major advantage of the JRR procedure is that, under quite general conditions, the same and relatively simple variance estimation formula holds for statistics of any complexity.

## The basic model of the JRR

Consider a design in which two or more primary units (PSU) have been selected independently from each stratum. Within each PSU, subsampling of any complexity may be involved. Each JRR replication is formed by eliminating one sample PSU from a particular stratum at a time, and increasing the weight of the remaining sample PSU's in that stratum appropriately so as to obtain an alternative but equally valid estimate to that obtained from the full sample.

Let u be a full-sample estimate of any complexity, and $u_{(hi)}$ be the estimate produced using the same procedure after eliminating primary unit i in stratum h and increasing the weight of the remaining ($a_h$-1) units in the stratum by an appropriate factor $g_h = w_h / (w_h - w_{hi})$

Let $u_{(h)}$ be the simple average of the $u_{(hi)}$ over the $a_h$ values of i in h. Then:

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

## 2 Variance estimation under cumulation using JRR

For the purpose of estimating variance of cumulated measures, the JRR variance estimation methodology is easily extended.

The total sample of interest is formed by the union of all the cross-sectional samples being aggregated. Using as basis the common structure of this total sample, each replication is formed such that when a unit is to be excluded in its construction, it is excluded simultaneously from every wave where the unit appears.

For each replication, the required measure is constructed for each of the cross-sectional samples, and these measures are used to obtain the required averaged measure corresponding to the particular the replication, from which variance is then estimated in the usual way.

Once the set of replications has been appropriately defined, the same variance estimation algorithm can be applied to a statistic of any complexity: for estimating variances for subpopulations (including regions), longitudinal measures such as persistent poverty rates, or measures of net changes and averages over time.

## 3 Estimating design effects under JRR

Analysis of design effects into components is needed for several purposes, but there is another important reason: with JRR design effect can only be estimated by estimating (some of) its components separately.

We decompose total variance v into the components as

$$v = v_0 . d^2 = v_0 . \left( d_W . d_H . d_D . d_X \right)^2$$

All factors other than $d_x$ depend only on individual elements.

By contrast, factor $d_X$ represents the effect on sampling error of complexities of the design (clustering and stratification). To estimate this effect using the JRR procedures, we compute variance under two assumptions about structure of the design:

- variance v under the actual design, and
- variance $v_R$ computed assuming the design to be (weighted) simple random sampling of the ultimate units - estimated from a 'randomised sample' created from the actual sample by completely disregarding its structure other than the weights attached to individual elements.

This gives $\left( d_x \right)^2 = \left( v / v_R \right)$

with $v_R = v_0 . \left( d_W . d_H . d_D \right)^2$

## 4 A major shortcoming: information on sample structure in EU-SILC

Appropriate coding of the sample structure, in the survey micro-data and accompanying documentation, is an essential requirement in order to compute sampling errors taking into account the actual sample design.

Lack of information on the sample structure in survey data files is a long-standing and persistent problem in survey work, and unfortunately affects EU-SILC as well.

Indeed, *the major problem in computing sampling errors for EU-SILC is the lack of sufficient information for this purpose in the micro-data available to researchers.*

We have developed approximate procedures in order to overcome these limitations at least partially, and used them to produce useful estimates of sampling errors. Use has been made of these results in this presentation, but it is not possible here to go into detail concerning them.

# IV. Fuzzy measures of poverty and deprivation: the concept and variance estimation

The introduction of fuzzy measures brings in *additional* considerations and choices such as:

- *Membership functions*: a quantitative specification of the propensity to poverty and deprivation of each person given the level and distribution of income and resource.
- *Rules for manipulation* of the resulting fuzzy sets: defining complements, intersections, union and aggregation of the sets.

To be meaningful both these choices must meet some basic logical and substantive requirements.

**Definition of the membership function based on monetary variables**

(Betti, Cheli Lemmi and Verma (2005, 2006))

$$\mu_i = FM_i = (1-F)^{\alpha-1}.[1-L(F)] = \left(\frac{\sum_{\gamma} w_\gamma \mid y_\gamma > y_i}{\sum_{\gamma} w_\gamma \mid y_\gamma > y_1}\right)^{\alpha-1} . \left(\frac{\sum_{\gamma} w_\gamma y_\gamma \mid y_\gamma > y_i}{\sum_{\gamma} w_\gamma y_\gamma \mid y_\gamma > y_1}\right)$$

Where parameter $\alpha$ is chosen so that the mean of the m.f. is equal to head count ratio H:

$$E(FM) = \frac{\alpha + G_\alpha}{\alpha.(\alpha+1)} = H$$

## Poverty and inequality

Fuzzy Monetary (FM) measure as defined above is expressible in terms of the generalised Gini measures. This family of measures (often referred to as "s-Gini") is a generalisation of the standard Gini coefficient, the latter corresponding to G with $\alpha$ =1.

It is defined (in the continuous case) as:

$$G_{\alpha} = \alpha.(\alpha+1).\int_{0}^{1}\left[(1-F)^{(\alpha-1)}.(F-L(F))\right]dF$$

The authors have defined it as “Integrated Fuzzy and Relative” (IFR)

## Membership function based on supplementary variables (FS)

Quantification and putting together a large set of non-monetary indicators of living conditions involves a number of steps, models and assumptions.

1. selection of indicators which are substantively meaningful and useful: mostly used ‘objective’ indicators
2. identifying underlying dimensions: this is done via factor analysis and sensible considerations; grouping the indicators accordingly
3. assigning numerical values to ordered categories
4. weighting of measures
5. scaling of measures

**Membership function based on supplementary variables (FS)**

Here we have adopted the methodology of the *Second European report on Poverty, Income and Social Exclusion* (Eurostat, 2002)

Elementary indicators are combined to form an index describing an overall degree of deprivation. The individual's score averaged over items (k) is written as the weighted mean:

$$S_j = \Sigma_k \left( w_k . s_{j,k} \right) / \Sigma_k \, w_k$$

where the weights $w_k$ are defined taking into account dispersion and correlation among items.

## Income poverty and non-monetary deprivation in combination

The two measures $FM_j$ propensity to income poverty, and $FS_j$ the overall life–style deprivation propensity, may be combined to construct composite measures which indicate the extent to which the two aspects of income poverty and life-style deprivation overlap for the individual concerned. These measures are as follows.

- $M_j$ manifest deprivation,

representing the propensity to both income poverty and life-style deprivation simultaneously. It represents the individual being subject to both income poverty and life-style deprivation; one may think of this as the '**m**anifest' or the '**m**ore intense' degree of deprivation.

- $L_j$ latent deprivation,

representing the individual being subject to at least one of the two, income poverty and/or life-style deprivation; one may think of this as the '**l**atent' or the '**l**ess intense' degree of deprivation.

## On structure of the JRR variance computation algorithm

**Parameters involved in the definition of a measure**

In the constant parameter version, we write the poverty or inequality measure in the form of an ordinary ratio, treating the parameters involved in the definition of the measures as constants. The micro-level variables ($u_i$), as defined using the parameters L estimated from the full sample, are used unchanged, the only difference being the set of units and their adjusted weights included in the computation of the required statistic (say $U_k$) for each replication k. In other words, these parameters are computed only once based on micro data for the full sample and are used unchanged in each replication.

In the variable parameter (i.e., real) version, the results are produced by treating the parameters involved as variable from one replication to another: the micro-level variables ($u_i$) are redefined in each replication, using the parameters $\Lambda$ estimated for that replication, based on micro data for the sample cases included in that replication.

In general, by repeating the entire estimation procedure independently for each replication the effect of various complexities, such as each step of a complex weighting procedure, can be incorporated into the variance estimates produced.

| | | | |
|---|---|---|---|
| Full sample, S | $\Lambda = \Lambda(s)$ | $u_i = u(s_i, \Lambda)$ | $U = \Sigma_{i \in s} w_i . u_i$ |
| Replication $S_k$ | | | |
| Constant parameter | $\Lambda_k = \Lambda$ | $u_{i,k} = u_i$ | $U_k = \Sigma_{i \in k} (w_{i,k} . u_i)$ |
| Variable parameter | $\Lambda_k = \Lambda(S_k)$ | $u_{i,k} = u(s_i, \Lambda_k)$ | $U_k = \Sigma_{i \in k} (w_{i,k} . u_{i,k})$ |

$S_k$ refers to replication k;
$s_i$ are the values for a variable or set of variables for unit i in the sample;
$u_i$ refers to the variable for unit i, the weighted sum of which gives the statistic of interest U.
$u_{i,k}$ and $U_k$ refers to the corresponding quantities for a particular replication k.
$\Lambda$ is the set of parameters, estimated form the sample, which are involved in the definition of U and $u_i$.
$\Lambda_k$ is the corresponding estimate based on replication k.

**In the case of conventional poverty rate,**

parameter  refers to the poverty line, different by replication k. In the variable version of the method, the individual dichotomous (0,1)  values $u_{ik}$ are affected in the region of changes in the poverty line.

**In corresponding application to fuzzy poverty or deprivation,**

the parameters involved are a and the weights used to put together individual items of deprivation to construct deprivation dimensions. We treat these as ‘external’ parameters, not varied from one replication to another. Nevertheless, the individual $u_{ik}$ values can vary from one replication to another because they depend on the whole distribution (actually on the ranking of individuals in the income distribution).