QUALITY IMPROVEMENTS IN VARIANCE ESTIMATION FOR THE LABOUR FORCE SURVEY

Ceccarelli C., Chianella D., De Filippo F., Graziani C., Guandalini A., Lattanzio M., Loriga S., Martini A., Terribili M.D.

1. Introduction

The analysis of the results of a sample survey should always include an assessment of the accuracy of the estimates, a measure of their dispersion around the true, unobserved value of the target variable in the population. Furthermore it is a good practice for national statistical institutes to estimate sampling errors and then to disseminate them in a transparent and clear way for the users.

Dissemination strategy for Italian Labour Force Survey (LFS) include a quite large set of accuracy indicators for the main results produced. Monthly, quarterly and yearly press releases include evaluation of precision for the main indicators on labour market. These precision indicators are currently transmitted to Eurostat for both the quarterly and the yearly Quality Accuracy Reports. Istat received recently solicitations coming from Eurostat and the European Commission's Directorate-General for Employment, Social Affairs and Equal Opportunities to broaden the accuracy evaluations provided, considering also variance estimations for LFS-based indicators’ annual net changes. The purpose of this document is to present and explain the methodology developed for the calculation of the annual net changes variance’s estimate for Italian LFS and the first results obtained.

2. The Italian LFS sampling design

The Italian LFS (Istat, 2006) sampling design is a two stage sampling design with stratification of the primary units (PSUs). The PSUs are the municipalities, while the second stage units (SSUs) are the households. In each NUTS3 domain, large municipalities with population over a certain threshold, are always included in the sample (self-representative strata: SR-strata). Smaller municipalities (not self-representative: NSR-strata) are grouped into strata, composed by municipalities having similar population size; then one municipality is selected in
each stratum with probability proportional to its population. Households are randomly selected (SRS) from the population frame, in all the municipalities selected at the first stage\(^1\).

The quarterly sample is uniformly spread over the 13 weeks of the quarter. The representativeness of the sample, that is the presence of all the strata, is guaranteed for each month (group of 4 or 5 weeks).

### Table 1 - Main features of the Italian LFS sampling design.
*Q1:2015 and Q1:2016, mean values.*

<table>
<thead>
<tr>
<th></th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall number of municipalities</td>
<td>8.057</td>
<td>8.047</td>
</tr>
<tr>
<td>Number of municipalities in the sample</td>
<td>1.260</td>
<td>1.264</td>
</tr>
<tr>
<td>Theoretical sampling size (households)</td>
<td>71.536</td>
<td>71.536</td>
</tr>
<tr>
<td>Effective sampling size (households)</td>
<td>65.151</td>
<td>65.236</td>
</tr>
<tr>
<td>Population size (households)</td>
<td>25.792.738</td>
<td>25.766.327</td>
</tr>
</tbody>
</table>

*Source: Italian LFS.*

About 10% of the smallest municipalities is rotated every year: a new municipality is randomly selected from each of these strata (with probability proportional to its population). The households are rotated according to a 2-(2)-2 rotation scheme: households are interviewed during two consecutive quarters. Then, after a two-quarters break, they are again interviewed twice in the corresponding two quarters of the following year. As a result, each household is included in four waves of the survey along a 15-months period.

### 3. The estimation procedure

The calibration estimator (Deville & Särndal, 1992 and Särndal, 2007) is used to obtain Italian LFS estimates. The final weights are computed as follows:

1. First, design weights \( \delta_j \) are obtained as the inverse of the inclusion probabilities of any household in the sample.
2. Then, correction factors \( k_j \) are worked out as the reciprocal of the response ratios (computed for several kind of households and territorial domains), in order to take into account households non-response. Intermediate weights,

\(^1\) For methodological details see Cochran (1977) and Conti & Marella (2011).
corrected for non-response, are then computed multiplying design weights by these correction factors \( d_j = \delta_j k_j \).

3. Then, starting from intermediate weights \( d_j \), final weights \( w_j \) are obtained solving a minimization problem under constraints. The function to be minimized is a distance between final and intermediate weights; the constraints regard the estimates of some auxiliary variables that have to be equal to the totals in the reference population derived by external sources. The main constraints are: population by gender and 14 5-years age groups (at NUTS2 level), population by gender and 5 age groups (at NUTS3 level and for the 13 biggest municipalities), monthly population by gender (at NUTS2 level), number of households by wave (at NUTS2 level), foreigner population (Male, Female, EU and Not EU at NUTS2 level). Final weights ensure that all members of a given household have the same weight.

Through the calibration estimator, applying final weights, the sample reproduces the same distribution of the population according to the chosen auxiliary variables.

The estimate of the total of the variable \( Y \) is then computed as follows:

\[
\hat{Y} = \sum_{j \in s} y_j w_j = \sum_{j \in s} y_j d_j y_j = \sum_{j \in s} y_j d_j k_j y_j.
\]

Weights are computed on quarterly basis, whereas annual estimates are computed as averages of quarterly estimates.

4. The variance estimation procedure

With every sampling figure, an estimate of its accuracy must be provided, so a measure of the variability of the estimates around the unknown “true” value of the parameter in the reference population is calculated, depending on both the partial observation of the phenomenon in the sample instead of the whole population and the adopted estimation procedure.

For each quarter, an evaluation of the accuracy of the survey figures is given by the variance of the estimator. Considering that the Italian LFS sampling design deals with SR and NSR strata and that the different strata are independent, the variance estimation is given by:

\[
\hat{\text{Var}}(\hat{Y}) = \hat{\text{Var}}(\hat{Z}) = \text{SR} \hat{\text{Var}}(\hat{Z}) + \text{NSR} \hat{\text{Var}}(\hat{Z})
\]
where \( Z \) is the linearization of \( Y \), \( h \) is the stratum index, \( H_{SR} \) is the set of SR strata, \( H_{NSR} \) for NSR strata taking into account that the calibration estimator is not linear.

In SR strata the estimation of the variance is computed among the observation of the variable \( Y \) in the households, according to this formula:

\[
\text{SR} \tilde{\text{Var}}(\tilde{Z}) = \sum_{h \in H_{SR}} M_h \frac{(M_h - m_h)}{m_h(m_h - 1)} \sum_{j=1}^{m_h} (z_{hj} - \bar{z}_h)^2
\]

where \( \bar{z}_h = \frac{1}{m_h} \sum_{j=1}^{m_h} z_{hj} \). In NSR strata, the corresponding formula for the variance estimation is given by:

\[
\text{NSR} \tilde{\text{Var}}(\tilde{Z}) = \sum_{h \in H_{NSR}} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left( \hat{z}_{hi} - \frac{\hat{z}_h}{n_h} \right)^2
\]

where \( \hat{z}_{hi} = \sum_{j=1}^{m_{hi}} z_{hij} w_{hij} \) and \( \hat{z}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} z_{hij} w_{hij} \). that is in SR strata the variance is computed among the households (SSUs), while in NSR strata it is computed among the municipalities (PSUs) that is considering only the leading contribution.

Considering that the LFS sample is selected on quarterly basis, the variance estimation is conducted for each quarterly estimate. Istat along the years has developed specific IT tools in order to provide accuracy estimation for sampling surveys estimates, that implement the methodology above mentioned. Genesees (Generalized Software for Sampling Estimates and Errors in Surveys) was a SAS-based application to calculate estimates and sampling errors for complex designs, recently ReGenesees (R Evolved Generalized Software for Sampling Estimates and Errors in Surveys) is a full-fledged R software for design-based and model-assisted analysis of complex sample surveys.

In this work we used ReGenesees, as we currently do in the context of the LFS.

5. The net changes variance estimation

The net change \( \Delta \bar{Y} \) is the difference between yearly estimates. The plug-in estimator for net changes can be written as:

\[
\hat{\Delta} \bar{Y} = \tilde{\bar{Y}} - \bar{Y}_{t-1}\]

where \( \hat{Y}_t \) is the yearly total estimator of \( Y \) for the year \( t \) and \( \hat{Y}_{t-1} \) is the same estimate but related to the previous year. In the Italian LFS, the yearly estimates are obtained as average of four quarterly estimates of totals:

\[
\hat{Y}_t = \frac{1}{4} \sum_{q=1}^{4} q \hat{Y}_t
\]

(2)

where \( t \) is the year and \( q (= 1, 2, 3, 4) \) are the quarters. The Italian LFS, as described in paragraph 2, adopts a 2-(2)-2 rotation scheme. It means that a rotation group of households remains in the sample for two consecutive quarters, then it leaves the sample for two quarters, and then it is considered again in the sample for other two consecutive quarters. Then it is dropped out definitely. The theoretical overlapping is equal to 50% for two consecutive quarterly samples and for same quarters in two consecutive years, it’s equal to 25% with a lag of three or five quarters, otherwise there is no overlapping.

**Figure 1** – Rotating design of LFS and overlapping of a quarterly theoretical sample in two years. Overlapping, \( o_q^{(\bullet)} \), has been expressed considering the quarter 1 of year 2 as base.

The variance estimator of the yearly estimate is given by:
\[
\hat{\text{Var}}(\hat{Y}_q) = \hat{\text{Var}}\left(\frac{1}{4} \sum_{q=1}^{4} \hat{Y}_q\right) = \frac{1}{16} \left( \sum_{q=1}^{4} \hat{\text{Var}}(\hat{Y}_q) + 2 \sum_{q \neq q'} \hat{\text{Cov}}(\hat{Y}_q, \hat{Y}_{q'}) \right). \tag{3}
\]

It depends on the \(\hat{\text{Var}}(\hat{Y}_q)\), the variance estimator of quarter \(q\), and \(\hat{\text{Cov}}(\hat{Y}_q, \hat{Y}_{q'})\), that is the covariance estimator between two quarterly estimates. The covariance estimator can be written as the product of the auto-correlation between estimates of two quarters and the square root of the product of the quarterly variance estimates:

\[
\hat{\text{Cov}}(\hat{Y}_q, \hat{Y}_{q'}) = \hat{\delta}^{q(t)}_{q'(t)} \hat{\rho}^{q(t)}_{q'(t)} \sqrt{\text{Var}(\hat{Y}_q)\text{Var}(\hat{Y}_{q'})} \quad q = 1, 2, 3, 4
\]

where \(\hat{\delta}^{q(t)}_{q'(t)}\) is the partial overlapping of quarterly samples of year \(t\) due to the rotation scheme used in LFS and \(\hat{\rho}^{q(t)}_{q'(t)}\) is the auto-correlation between estimates of two quarters. For a more accurate estimate of \(\hat{\text{Cov}}(\hat{Y}_q, \hat{Y}_{q'})\) we compute the effective overlapping between quarters, that is lower than the theoretical one due to the attrition. The auto-correlations, \(\hat{\rho}^{q(t)}_{q'(t)}\), are computed without sampling weights. In the calculation are taken into account three auto-correlations at one quarter and one auto-correlation at three quarters.

Then, the estimator of variance of net change in (1) is:

\[
\hat{\text{Var}}(\hat{\Delta Y}) = \hat{\text{Var}}(\hat{Y}) + \hat{\text{Var}}(\hat{Y}^{t-1}) - 2\hat{\text{Cov}}(\hat{Y}, \hat{Y}^{t-1}) \tag{4}
\]

where \(\hat{\text{Var}}(\hat{Y})\) and \(\hat{\text{Var}}(\hat{Y}^{t-1})\) are the variance estimators in (3), while \(\hat{\text{Cov}}(\hat{Y}, \hat{Y}^{t-1})\) is the estimator of the covariance between yearly estimates. In this case, in the calculation of the covariance term are taken into account seven auto-correlations at one quarter, five auto-correlations at three quarters, four auto-correlations at four quarters and three auto-correlations at five quarters.

All the expression formulas above hold also for the linearized version of \(\hat{Y}\).
Table 2 – 2015-2016 Net changes accuracy evaluation for the list of 23 indicators requested (part 1).

<table>
<thead>
<tr>
<th>INDICATOR DESCRIPTION</th>
<th>V_2015</th>
<th>V_2016</th>
<th>SE</th>
<th>VAR</th>
<th>CHANGE</th>
<th>POINT ESTIMATION</th>
<th>C.I.</th>
<th>SIGNIFICANT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newly employed, age group 20-64, in %</td>
<td>10.64</td>
<td>11.01</td>
<td>0.1233</td>
<td>0.0152</td>
<td>0.4</td>
<td>0.1</td>
<td>0.6</td>
<td>YES</td>
</tr>
<tr>
<td>Employment rate of older workers, age group 55-64, in %</td>
<td>48.24</td>
<td>50.33</td>
<td>0.1553</td>
<td>0.0241</td>
<td>2.1</td>
<td>1.8</td>
<td>2.4</td>
<td>YES</td>
</tr>
<tr>
<td>Temporary employees as percentage of employees, age group 20-64, in %</td>
<td>13.81</td>
<td>13.81</td>
<td>0.1128</td>
<td>0.0127</td>
<td>0.0</td>
<td>-0.2</td>
<td>0.2</td>
<td>NO</td>
</tr>
<tr>
<td>Youth unemployment rate, age group 15-24, in %</td>
<td>40.32</td>
<td>37.77</td>
<td>0.6019</td>
<td>0.3623</td>
<td>2.6</td>
<td>-3.7</td>
<td>-1.4</td>
<td>YES</td>
</tr>
<tr>
<td>Unemployment rate, age group 15-74, in %</td>
<td>11.93</td>
<td>11.72</td>
<td>0.0963</td>
<td>0.0093</td>
<td>0.2</td>
<td>-0.4</td>
<td>0.0</td>
<td>YES</td>
</tr>
<tr>
<td>NEET rate, age group 18-24, in %</td>
<td>27.91</td>
<td>26.04</td>
<td>0.3113</td>
<td>0.0969</td>
<td>-1.9</td>
<td>-2.6</td>
<td>-1.3</td>
<td>YES</td>
</tr>
<tr>
<td>Employment rate 20-44, total, in %</td>
<td>60.33</td>
<td>61.56</td>
<td>0.0863</td>
<td>0.0044</td>
<td>1.0</td>
<td>0.9</td>
<td>1.2</td>
<td>YES</td>
</tr>
<tr>
<td>Employment rate 20-44, men, in %</td>
<td>70.60</td>
<td>71.68</td>
<td>0.0977</td>
<td>0.0095</td>
<td>1.1</td>
<td>0.9</td>
<td>1.3</td>
<td>YES</td>
</tr>
<tr>
<td>Employment rate 20-44, women, in %</td>
<td>50.62</td>
<td>51.59</td>
<td>0.0915</td>
<td>0.0084</td>
<td>1.0</td>
<td>0.8</td>
<td>1.2</td>
<td>YES</td>
</tr>
<tr>
<td>Early leavers from education and training, age group 18-24, in %</td>
<td>14.70</td>
<td>13.78</td>
<td>0.1952</td>
<td>0.0381</td>
<td>0.9</td>
<td>-1.3</td>
<td>-0.5</td>
<td>YES</td>
</tr>
<tr>
<td>Tertiary educational attainment, age group 30-34, in %</td>
<td>25.35</td>
<td>26.17</td>
<td>0.2626</td>
<td>0.0690</td>
<td>0.8</td>
<td>0.3</td>
<td>1.3</td>
<td>YES</td>
</tr>
</tbody>
</table>

V_2015 = estimated value in 2015; V_2016 = estimated value in 2016; SE = relative error; VAR = Sampling Variance; CHANGE (Point Estimation) = V_2016 - V_2015; CHANGE (C.I.) = 95% Confidence Interval of Change, lower bound (LB) and upper bound (UB). SIGNIFICANT = YES if the change is significantly different from 0, NO otherwise.

Source: Italian LFS
### Table 2 – 2015-2016 Net changes accuracy evaluation for the list of 23 indicators requested (part 2).

<table>
<thead>
<tr>
<th>INDICATOR DESCRIPTION</th>
<th>V_2015</th>
<th>V_2016</th>
<th>SE</th>
<th>VAR</th>
<th>CHANGE POINT ESTIMATION</th>
<th>C.I. LB</th>
<th>C.I. UB</th>
<th>SIGNIFICANT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Youth employment rate, age group 20-29, in %</td>
<td>40.29</td>
<td>42.06</td>
<td>0.2453</td>
<td>0.0602</td>
<td>1.8</td>
<td>1.3</td>
<td>2.3</td>
<td>YES</td>
</tr>
<tr>
<td>Employment rate of low skilled persons, age group 20-64, in %</td>
<td>49.39</td>
<td>50.32</td>
<td>0.1178</td>
<td>0.0139</td>
<td>0.9</td>
<td>0.7</td>
<td>1.2</td>
<td>YES</td>
</tr>
<tr>
<td>Employment rate of non-EU nationals, age group 20-64, in %</td>
<td>60.72</td>
<td>61.52</td>
<td>0.2346</td>
<td>0.0810</td>
<td>0.8</td>
<td>0.2</td>
<td>1.4</td>
<td>YES</td>
</tr>
<tr>
<td>Part-time employment rate, age group 20-64, in %</td>
<td>18.24</td>
<td>18.49</td>
<td>0.0832</td>
<td>0.0069</td>
<td>0.3</td>
<td>0.1</td>
<td>0.4</td>
<td>YES</td>
</tr>
<tr>
<td>NEET rate, age group 15-24, in %</td>
<td>21.39</td>
<td>19.87</td>
<td>0.2261</td>
<td>0.0511</td>
<td>-1.5</td>
<td>-2.0</td>
<td>-1.1</td>
<td>YES</td>
</tr>
<tr>
<td>Share of adult population with upper secondary or tertiary education, age group 25-64, in %</td>
<td>59.90</td>
<td>60.14</td>
<td>0.0514</td>
<td>0.0026</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>YES</td>
</tr>
<tr>
<td>Lifelong learning: Percentage of adult population aged 25-64 participating in education and training</td>
<td>7.33</td>
<td>8.26</td>
<td>0.0792</td>
<td>0.0063</td>
<td>0.9</td>
<td>1.1</td>
<td>0.8</td>
<td>YES</td>
</tr>
<tr>
<td>Youth unemployment ratio, age group 15-24, in %</td>
<td>10.56</td>
<td>10.05</td>
<td>0.1943</td>
<td>0.0377</td>
<td>-0.5</td>
<td>-0.9</td>
<td>-0.1</td>
<td>YES</td>
</tr>
<tr>
<td>Long term unemployment rate, age group 15-74, in %</td>
<td>6.93</td>
<td>6.72</td>
<td>0.0782</td>
<td>0.0061</td>
<td>-0.2</td>
<td>-0.4</td>
<td>-0.1</td>
<td>YES</td>
</tr>
<tr>
<td>Involuntary temporary employment, age group 20-64, in %</td>
<td>10.20</td>
<td>10.13</td>
<td>0.1780</td>
<td>0.0817</td>
<td>-0.1</td>
<td>-0.4</td>
<td>0.3</td>
<td>NO</td>
</tr>
<tr>
<td>Involuntary part-time employment as percentage of the total part-time employment, age group 20-64, in %</td>
<td>65.45</td>
<td>64.24</td>
<td>0.3400</td>
<td>0.1156</td>
<td>-1.2</td>
<td>-1.9</td>
<td>-0.3</td>
<td>YES</td>
</tr>
</tbody>
</table>

V_2015 = estimated value in 2015; V_2016 = estimated value in 2016; SE = relative error; VAR = Sampling Variance; CHANGE (Point Estimation) = V_2016 - V_2015; CHANGE (C.I.) = 95% Confidence Interval of Change, lower bound (LB) and upper bound (UB); SIGNIFICANT = YES if the change is significantly different from 0, NO otherwise.

Source: Italian LFS
6. First results

As presented at the December 2014 meeting of the Eurostat Labour Market Statistics Working Group (LAMAS), DG EMPL of the European Commission asked Eurostat to provide from 2017 onwards estimates of the variance of annual net changes for a number of LFS based indicators. The detailed list of indicators for which variance estimates of annual net changes have been requested was defined in the annex 1 of LAMAS document Eurostat/F3/LAMAS/29/15.

According to these definitions and the above described methodology the level estimates for 2015 and 2016 as well as the annual net change 2015-2016 have been calculated.

In particular the same SAS macro statement for indicators calculation and linearization of the target variables have been used. Quarterly standard errors have been calculated applying the usual tool (ReGenesees) and methods in order to guarantee coherence with the accuracy evaluation currently defined and disseminated.

7. Conclusions

With every sampling figure, an estimate of its accuracy must be provided, so a measure of the variability of the estimates around the unknown “true” value of the parameter in the reference population is calculated, depending on both the partial observation of the phenomenon in the sample instead of the whole population and the adopted estimation procedure.

The aim of this work is developing the estimation of the sampling variance of net changes between yearly estimates of the Italian LFS. This information is fundamental in order to analyze the significance of the difference year-to-year of a certain variable, properly taking into account of the sampling error and it has been required by Eurostat and the European Commission for the main policy indicators computed over LFS data.

According to specific National needs, the goal of the work has been furtherly enlarged in order to develop the estimation of the sampling variance of net changes between a quarterly estimate and the estimate in the corresponding quarter of the previous year, in order to allow the evaluation of the significance of annual changes, each quarter. Moreover the territorial domain of the estimates for which these variances have been computed is NUTS2.
Acknowledgements

The present work has been realized within the grant for the project “Quality improvements for the Labour Force survey” supported by Eurostat and the European Commission's Directorate-General for Employment, Social Affairs and Equal Opportunities.

References


SUMMARY

The Italian National Institute of Statistics, along with other European countries, is involved in the improvement efforts for the Labour Force Survey (LFS) in the context of the new IESS (Integrated European Social Statistics) framework regulation.

The IESS framework regulation asks for several actions that have to be undertaken by National Institutes of Statistics in order to achieve improvements in LFS. Among these actions, additional indicators for the annual point and annual net changes are request. Furthermore, the related estimates of variance of the estimates must be produced.

The requested indicators are mainly simple ratios (annual point) or difference of ratios (annual net changes), therefore nonlinear parameter must be linearized to derive the estimates of variance. Moreover, when computing the sampling variance the overlapping among the samples, due to the rotation scheme adopted in the Italian LFS, has to be taken into account. In this work the methodology used to produce annual point, annual net changes and the related variance estimations is illustrated.

Claudio CECCARELLI, Istituto nazionale di statistica, clceccar@istat.it
Diego CHIANELLA, Istituto nazionale di statistica, chianella@istat.it
Filomena DE FILIPPO, Istituto nazionale di statistica, defilipp@istat.it
Cinzia GRAZIANI, Istituto nazionale di statistica, cingraziani@istat.it
Alessio GUANDALINI, Istituto nazionale di statistica, alessio.guandalini@istat.it
Marco LATTANZIO, Istituto nazionale di statistica, lattanzio@istat.it
Alessandro MARTINI, Istituto nazionale di statistica, alemartini@istat.it
Silvia LORIGA, ISTAT, Istituto nazionale di statistica, siloriga@istat.it
Marco D. TERRIBILI, ISTAT, Istituto nazionale di statistica, terribili@istat.it