An IT framework for a quick evaluation of accuracy of Italian LFS.

Cinzia Graziani, Silvia Loriga, Alessandro Martini e Andrea Spizzichino

7th Workshop on Labour Force Survey Methodology

Madrid, May 10-11th 2012

Overview

- **Accuracy analysis in Italian LFS**
- **The prototype for a quick evaluation of sampling error**
- **Prospects for development**

The Issue I

The analysis of the results of a sample survey should always be accompanied by an assessment of the accuracy of the estimates, in terms of MSE, to take into account the estimator’s variability as well its bias.

Calibration estimator is biased but, with increasing size of the sample, the estimator converges asymptotically to the unbiased GREG estimator.

For large samples (such as LFS) we can assume that the calibration estimator has approximately the same properties (accuracy, consistency) as the GREG and the same sample variance.

An exact computation of the estimated variance is easy only for simpler sampling designs.

The Issue II

In all other cases the estimation is quite difficult and requires high-demand procedures in terms of computational complexity:

- Estimator no more linear function of sample data;
- Complex sample designs;
- Questionnaires are very complex.

Publication of estimated variances is very difficult to produce and to interpret for users.

For these reasons regression models may be used to produce synthetic evaluations of sampling errors.

Regression models

The hypothesis is the existence of a relation between relative sampling error $\varepsilon({}_d\hat{Y})$ and the estimation ${}_d\hat{Y}$, in particular for qualitative variables a model specification which shows a good fit is:

$$\boxed{\log \hat{\varepsilon}^2({}_d\hat{Y}) = a + b\log({}_d\hat{Y})}$$

Models are fitted for each domain of interest on a wide set of estimates, taking care to choose heterogeneous levels for them.

$$\boxed{\hat{\varepsilon}({}_d\hat{Y}) = \sqrt{\exp\left(a + b\log({}_d\hat{Y})\right)}}$$

The estimation of relative sampling errors makes it possible to define a confidence interval which, with a given probability α, is likely to include the actual value.

$$\boxed{({}_d\hat{Y} - z_{1-\alpha/2} * {}_d\hat{Y} * \hat{\varepsilon}({}_d\hat{Y}); {}_d\hat{Y} + z_{1-\alpha/2} * {}_d\hat{Y} * \hat{\varepsilon}({}_d\hat{Y}))}$$

An example of calculation for IT-LFS 2010

We can consider the estimation of the total male unemployment in the North, amounting to 196,000 individuals.

We obtain the following values of parameters for the model referred to the North: (a=6,590031 and b=-1,132387), so that:

$$\hat{\varepsilon}(196.000) = \sqrt{\exp(6{,}590031 + 1{,}132387 * \log(196.000))} = 2{,}72\%$$

The corresponding absolute error is:

σ (196.000)=2,72/100 x 196.000 = 5.331

And the bounds of the confidence interval (at 95%) are:

Lower=196.000 – (1,96 x 5.331) = 185.551

Upper= 196.000 + (1,96 x 5.331) = 206.449

If we want to analyze the unemployment rate by region and sex, this should be repeated **84 times**, using an Excel spreadsheet.

IT-LFS regression models methodology

For **relative frequencies** we have to distinguish two cases:

Relative frequency where the **denominator is a calibration constraint:**

$$\boxed{\hat{R}_d = \frac{\hat{Y}_d}{T_d}}$$

Example: Activity rate:

$$\boxed{\hat{A}ctR = \frac{\hat{A}ct}{Pop}}$$

we have to calculate just the sampling error for the numerator (case1)

Ratios where **numerator e denominator are both estimates**:

$$\boxed{\hat{R}_d = \frac{\hat{Y}_d}{\hat{D}_d}}$$

Example: Unemployment Rate

$$\boxed{\hat{U}neR = \frac{\hat{U}ne}{\hat{A}ct}}$$

An approximation is needed (case2):

$$\boxed{\hat{\varepsilon}(\hat{R}_d) = \sqrt{\hat{\varepsilon}^2(\hat{Y}_d) - \hat{\varepsilon}^2(\hat{D}_d)}}$$

Analyzing survey results

Making comparison across time and among different subpopulations is quite common before disseminating data. Analyzing the distribution of unemployment incidence on the female population by macro regions in the 4 th quarter of 2010:

	Stima	Lim.Inf.	Lim.Sup.
Nord Ovest	3.5	3.3	3.8
Nord Est	3.3	3.0	3.6
Centro	3.9	3.6	4.2
Sud	4.3	4.0	4.6
Isole	4.8	4.4	5.2

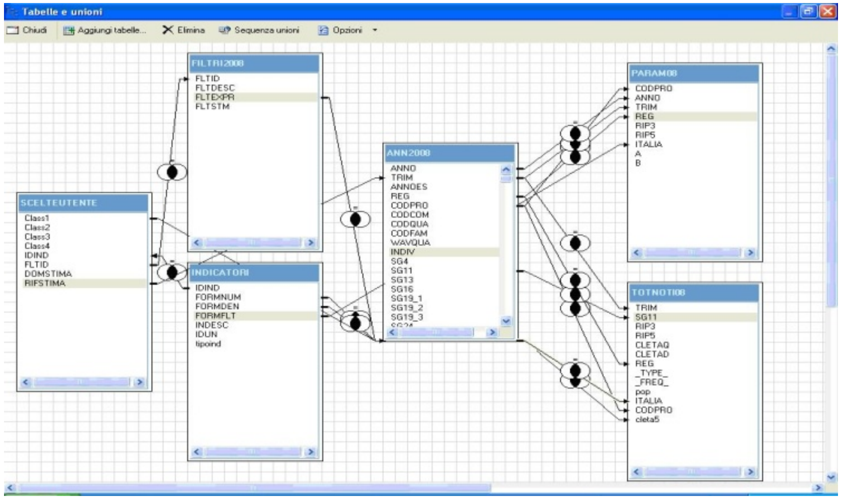
Can we say that?

1. **The percentage of unemployed women in the North-West is lower than that recorded in the South or Islands.** 🟢
2. **The percentage of unemployed women in the center is higher than in the North East** ⚠
3. **The percentage of unemployed women in the South is lower than in the Islands.** 🔴

Elaboration on IT-LFS 2010Q4 data

An IT framework for a quick evaluation of accuracy of Italian LFS

The procedure we developed automates the calculation of the estimates and their sampling errors using regression models, by integrating a set of metadata.



In the "data warehouse" SAS all the information needed to develop this capability for the Labour Force Survey have been stored, since 2006 until 2011:

- Micro data files
- Population totals used as constraints for calibration
- Regression model parameters
- Main indicators definition
- Filters definition for specific subpopulations (Gender, employed, age classes)

An IT framework for a quick evaluation of accuracy of Italian LFS - II

The procedure has been developed in SAS macro language and requires the user to specify some parameters.

For the calculation of the accuracy of LFS estimates, the following parameters have to be specified:

- The indicator of interest (absolute frequencies or rates);
- The classification variables;
- The domain of interest;
- The time reference;
- The filter to apply (including user-defined).

An IT framework for a quick evaluation of accuracy of Italian LFS - III

The flowchart of the algorithm can be summarized in the following steps:

1. Estimates calculation;
2. Extraction of occurrences in the metadata (parameters, domains, totals, filters, indicators);
3. Comparison with population totals;
4. Calculation of the relative sampling error;
5. Definition of confidence interval;
6. Tabulation of results.

An IT framework for a quick evaluation of accuracy of Italian LFS - IV

The choice of the correct method to calculate the sampling error is made during the elaboration, taking into account the results of matching with metadata.

In the metadata we define a classification for ratios in order to distinguish between those having estimates or population total as denominator. This classification allows to apply the correct method for evaluating relative sampling error, using the formula (2) or (1), respectively.

Estimates, once they have been calculated, are compared with known population totals and the correct formula can be applied.

An example:

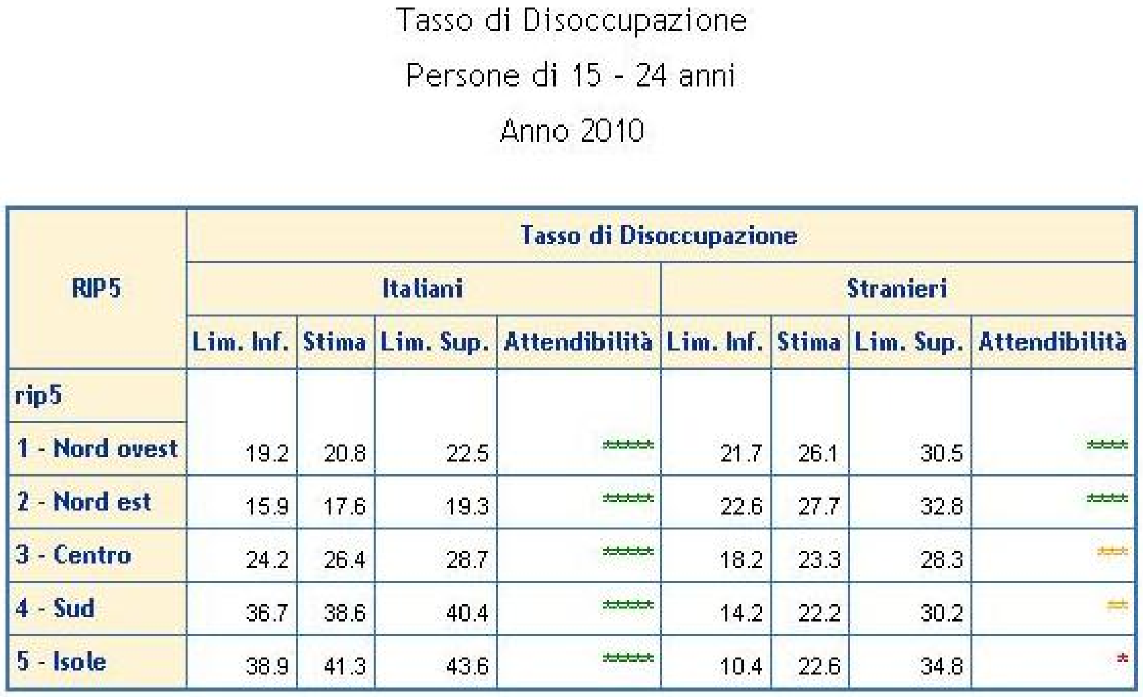
Estimation of the activity rate by region and age classes.
In this case the denominator consists of a population total considered in calibration procedure so sampling error have to be calculated just for the numerator, with formula (1).

However, if we apply a filter, specifying the analysis for married individuals, the denominator becomes an estimate and the formula (2) is required instead of (1).

The output of the procedure

Tables of results report:

- The estimates;
- The bounds of confidence interval (α= 95%)
- An evaluation of estimates accuracy:



Tasso di Disoccupazione

Persone di 15 - 24 anni

Anno 2010

RIP5	Tasso di Disoccupazione							
	Italiani				Stranieri			
	Lim. Inf.	Stima	Lim. Sup.	Attendibilità	Lim. Inf.	Stima	Lim. Sup.	Attendibilità
rip5								
1 - Nord ovest	19.2	20.8	22.5	*****	21.7	26.1	30.5	****
2 - Nord est	15.9	17.6	19.3	*****	22.6	27.7	32.8	****
3 - Centro	24.2	26.4	28.7	*****	18.2	23.3	28.3	***
4 - Sud	36.7	38.6	40.4	*****	14.2	22.2	30.2	**
5 - Isole	38.9	41.3	43.6	*****	10.4	22.6	34.8	*

CV Values	Symbol
CV<5%	*****
5%>=CV<10%	****
10%>=CV<15%	***
15%>=CV<20%	**
CV>=20%	*

Improve interpretability: users can easily get supplementary Information to interpret statistical figures.

Development perspectives

At the moment a first prototype, developed in SAS macro language/ AF forms is shared in a server with researchers of our division who have in charge data dissemination.

Procedure have been developed for other surveys conducted by our division (Adult Education Survey)

We are also studying the feasibility of developing the project within a business intelligence platform. We started a feasibility study to develop those capabilities with an open source tool (Pentaho), which starts to be used in our Institute.

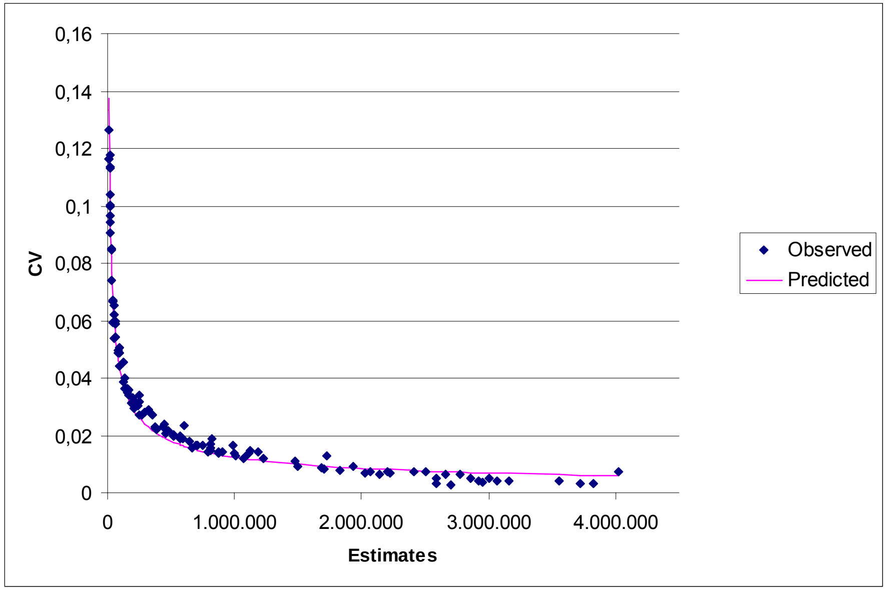
- web-intranet environment, so that access could be granted to researchers that visit Istat to make elaborations on micro data by their own.
- OLAP processing and to enable roll-up and drill-down operations on hypercubes with accuracy evaluation.
- Improve integration with other metadata driven system and dissemination data-warehouse (I.stat)

Istat

Thanks for your attention.

Regression models

			R^2		
Nuts II Area	Q1	Q2	Q3	Q4	2011
Piemonte	95,1	94,4	95,1	94,5	95,5
Val d'Aosta	93,6	93,2	94,0	94,1	94,8
Lombardia	94,4	96,0	94,5	95,5	96,3
Trentino	95,0	94,9	94,8	95,0	95,6
Veneto	93,8	94,2	95,2	94,0	95,6
Friuli Giulia	92,1	93,5	93,1	94,3	95,1
Liguria	92,1	93,2	94,8	94,2	95,7
Emilia R.	92,4	93,8	93,5	92,5	94,7
Toscana	92,5	93,6	93,3	93,4	94,7
Umbria	93,8	93,8	94,4	94,6	95,2
Marche	94,5	92,8	94,9	95,2	96,2
Lazio	94,6	92,2	92,5	95,9	96,2
Abruzzo	92,5	94,3	93,8	92,4	94,9
Molise	94,1	94,2	92,9	94,5	96,0
Campania	95,5	96,1	96,5	95,6	97,2
Puglia	96,2	94,8	94,2	93,9	96,3
Basilicata	95,0	95,8	96,2	95,5	96,7
Calabria	94,2	94,8	95,0	93,7	95,2
Sicilia	94,8	94,7	95,9	95,3	96,5
Sardegna	92,0	93,5	89,7	92,5	95,2
Italy	96,8	96,8	97,0	97,1	97,3


CV
Estimates
Observed
Predicted