EUROPEAN COMMISSION
EUROSTAT

Directorate F: Social statistics
Unit F-3: Labour market and lifelong learning

Subsampling options for structural variables and their representation in the technical variables in the future LFS

Hartmut Schrör, Eurostat

1. INTRODUCTION

Since the introduction of the structural LFS variables in 2006, an increasing number of NSIs have implemented the wave approach in their data collections. Currently, twelve NSIs use subsampling to collect structural variables.

The draft legal act for the new LFS under the IESS (Integrated European Social Statistics) framework regulation specifies several options and combinations when using subsampling for structural variables, however with several constraints. These options and constraints will be relevant to the sampling and data collection methods at national level.

The various subsampling options and constraints will have to be reflected in the technical variables. These have to be defined in a way that makes it possible to identify to which subsamples each observation in a dataset belongs and which validation rules and weights have to be used accordingly. Therefore, the revised INTQUEST variable will be more complete than before to cover all possible subsampling combinations. The set of weight variables will be extended with new ones for the biennial variables and the eight-yearly and ad-hoc subject variables.

Finally, data transmission from NSIs to Eurostat will be streamlined by using quarterly files only for the main variables and one single annual file for all variables. INTQUEST then becomes a key variable for the identification of subsamples and associated variable sets within the annual file.

This paper outlines the proposed rules and constraints for the subsampling and the associated changes to the INTQUEST variable, the weight variables and the data transmission. The focus is on the structural variables while options for household subsampling are not considered here in detail.

2. STRUCTURAL VARIABLES IN THE CURRENT AND NEW LFS

When referring to variable sets and subsampling options, some clarification on terminology is useful.

The legislation on the current LFS distinguishes between quarterly variables, yearly (also called structural) variables and ad-hoc module variables. “Structural” variables are those with a “yearly” periodicity according to Annex III of Commission Regulation (EC) No 377/2008. Ad-hoc module variables have a yearly reference period but are not called structural variables.

The draft legal act for the new LFS under IESS, expected to take effect from 2021, uses the expression “structural variable” in a wider sense. Structural variables comprise annual, biennial, eight-yearly variables and variables collected on an ad-hoc subject. Biennial variables are a new element; eight-yearly variables have also been referred to as “regular eight-yearly modules” replacing former ad-hoc modules, while an ad-hoc subject corresponds to a former ad-hoc module covering a new topic linked to the labour market.

3. SUBSAMPLING OPTIONS

In the current LFS, the subsampling options for structural and ad-hoc module variables are rather straightforward. Structural variables may be collected for a subsample of independent observations covering 52 weeks. When subsampling is used for the structural variables, it must also be used for the ad-hoc module because this sample must provide also the structural variables. Countries that do not use subsampling may collect the ad-hoc module in at least one full quarterly sample.

The provisions for subsampling according to the draft implementing act on the new LFS are more complex. On the one hand, NSIs should have subsampling options to choose from; on the other hand, combinations of the options must be meaningful and manageable (not least for the production of European statistics). Generally, subsampling is considered for all structural variables. The following non-exhaustive list of requirements applies:

- While subsampling is optional for annual and biennial structural variables, it is mandatory for the eight-yearly variables and ad-hoc subject variables. For those, it is no longer an option to use full quarterly samples (usually the second one).

- If the optional subsampling is used for the annual / biennial variables, it must be applied to the complete sets of variables. No selection within them is allowed.

- Subsampling must be applied to complete waves, representing in each quarter at least one eighth of the full quarterly sample.

- Subsamples must consist of independent observations covering all reference weeks of the reference year.

- The sample for biennial structural variables must be part of the sample for annual structural variables. Hereby, it is:
    - possible to use the same (sub)sample both for annual and biennial variables,
    - possible to use a subsample only for biennial while not for annual variables,
    - possible to use a subsample for annual variables which includes another subsample for biennial variables,
    - but *not* possible to use a subsample only for annual while not for biennial variables,
    - and *not* possible to use a subsample for biennial variables which is not included in the (sub)sample for annual variables.
- The subsample for eight-yearly variables and variables on an ad-hoc subject must be included in the (sub)sample for annual and biennial variables.

Further provisions on distributional requirements, age coverage and consistency with quarterly averages are not considered here, as they are not relevant to the possible combinations of subsamples in general. From the list above, it is important to bear in mind that the subsample for eight-yearly variables and variables on an ad-hoc subject must be part of the (sub)sample for biennial variables, which in turn must be included in the (sub)sample for annual variables.

4. WEIGHTS RELATED TO THE SUBSAMPLES

The draft legal act for the future LFS defines the following five weight variables to be reported for the various sets of variables and samples.

COEFFQ	Quarterly weighting factor
COEFFY	Yearly weighting factor
COEFF2Y	Weighting factor for the 2-yearly variables
COEFFMOD	Yearly weighting factor – module (eight-yearly variables, ad-hoc subjects)
COEFFHH	Yearly household weighting factor

There are some changes compared to the current situation. The new weight variable COEFF2Y for the biennial variables is introduced and COEFFMOD becomes a standard variable in the core set of variables, to be used for the respective eight-yearly variables or ad-hoc subject of a reference year. In addition, all weight variables should be reported explicitly for each observation to which they are relevant. For instance, COEFFY should be reported even if it is identical to COEFFQ/4 in countries without subsampling for annual variables; COEFF2Y should be reported even if it is identical to COEFFY because the same (sub)sample is used. The same applies to COEFFMOD, which may be identical to COEFFY and COEFF2Y. COEFFHH should be reported for all observations to be used for household analysis, regardless of whether household subsampling is used or not. Even if not, COEFFHH may be different from, say, COEFFY, because a country may apply a special weighting scheme for purposes of household level analysis. The expected weight variables become clearer when looking at the new INTQUEST variable and its role in data transmission to Eurostat.

5. Data transmission and the new INTQUEST variable

In June 2017, LAMAS agreed with Eurostat's proposal to streamline the way in which NSIs transmit in the future the different sets of variables for a given reference year based on four quarterly and one single annual dataset. The draft legal act on the future LFS reflects this principle.

In the quarterly datasets, only the quarterly variables will be expected. The annual file will contain all structural variables with the replicated quarterly variables for the respective (sub)samples used for structural variables. Where applicable, it will also contain the household subsample.

This setup of the transmission files ensures that all variables collected for an observation are transmitted in a maximum of two records, i.e. one in the quarterly data and one in the yearly data (if structural variables are collected from the respondent). This is important to avoid merging files at Eurostat, which is error-prone, and to simplify validation, while keeping duplication of data to the necessary minimum by duplicating quarterly variables once in the annual file.

The new INTQUEST variable becomes an important technical variable to identify which questionnaire was used for each record in a dataset; in other words, which sets of variables were collected for the observation and for which (sub)samples this observation was selected. This information is important for the correct aggregation of the quarterly and several annual, biennial, eight-yearly and ad-hoc results, for the correct use of the respective weights, and for performing the correct set of validation rules on each observation.

Therefore, the code list for the new INTQUEST variable is more detailed than before:

code	label
01	Quarterly
02	Quarterly and yearly
03	Quarterly, yearly and biennial
04	Quarterly, yearly, biennial and module
05	Quarterly and (originally selected) respondent forms part of household subsample
06	Quarterly, yearly and (originally selected) respondent forms part of household subsample
07	Quarterly, yearly, biennial and (orig. selected) respondent forms part of household subsample
08	Quarterly, yearly, biennial, module and (orig. selected) respondent forms part of household subsample
09	Household - minimum set of variables (for additional household members)
10	Household - restricted set of module background variables (for additional household members)

The full set of codes will be relevant for countries that survey individuals and collect household level information only for a subsample. Countries surveying households will only need codes 01 to 04. Codes 01 and 02 correspond to the current coding, though with currently different labels (“only core questionnaire”, “whole questionnaire”). Code 01 is to be used for all observations in the quarterly files transmitted to Eurostat and only in these files. Codes 02 to 04 apply to the transmission of annual files and reflect the subsampling options for structural variables described in section 3 of this document. A record in the single annual file must always contain the quarterly variables and the set(s) of structural variables according to the subsampling applied in the survey.

The coding reflects the principle that the (sub)sample for the yearly variables must be included in the set of the full four quarterly samples (code 02), the (sub)sample for the biennial variables must be included in those (code 03), and that the compulsory subsample for the ‘module’ (eight-yearly variables or ad-hoc subject) is always included in all the above (code 04). The same logic is repeated in codes 05 to 08 for countries applying a household subsample, followed by two codes identifying the additional household members.

When compiling the annual dataset, it is then important to report the weights expected for each set of variables consistently with the coding of the INTQUEST variable. For instance, a record with INTQUEST = 3 should have weights in COEFFY, COEFF2Y and COEFFHH, while COEFFQ and COEFFMOD are not expected to be filled in. COEFFQ is not relevant for the annual dataset and COEFFMOD is not expected to be filled in as the observation is not in the ‘module’ subsample. COEFFHH however has to be included, even if identical to another weight.

6. CONCLUSION

The proposed legal framework for the LFS under IESS from 2021 provides for various options to apply subsampling to structural variables and makes it even mandatory for eight-yearly variables and ad-hoc subjects. As the data processing and production of European statistics should be as automated as possible, it is necessary to build the information on the subsampling arrangements into the datasets themselves and to report all relevant weights explicitly, even if there are identical ones. At the same time, duplicating data in different datasets or distributing data across them should be kept to the necessary minimum. The “4+1” principle, four quarterly files and one annual file for data transmission, fulfils these requirements best. The new INTQUEST becomes a crucial technical variable in the annual file for the correct processing, validation and aggregation of the survey results. Miscoding of INTQUEST will result in errors assigning observations to the (sub)samples and may trigger false errors in the validation process. With the introduction of the new shared validation system, which is expected in 2020, NSIs will have the opportunity to test transmissions according to the new standard file structure before the first official transmissions are due in 2021.