



7 th WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

DATA PROCESSING AND DATA QUALITY

Madrid, Spain, 10 – 11 May 2012

A. Data processing: Processing of data with panel design and multi-mode data collection. Quality Controls and checks of data

A4–Quality controls and checks of data in Finnish LFS production system

Kalle Sinivuori - Finland

Quality controls and checks of data in the Finnish LFS production system

7th Workshop on LFS methodology, Madrid, 10-11 May 2012

Kalle Sinivuori / Statistics Finland

Background

The production system of Finland's Labour Force Survey was reviewed between 2002 and 2006. The previous production system had been functioning without fault for 20 years, but it was clearly behind time. The old production system operated in the mainframe environment and it was mainly used by IT experts. For LFS experts the system appeared as a "black box", from which the results were run once a month.

The aim for the new production system was that it would be reliable, transparent and managed by LFS experts. It was made as a .NET application in the SQL database environment. The idea was to build as handy as possible application by the 'press play - control report' -principle, whereby all the process steps go through an LFS expert. The project was also successful: In 2008, the Labour Force Survey production system received Statistics Finland's internal quality award, and in two auditing processes conducted in 2011 (auditing of the statistics and risk analysis of the statistics) our production system was evaluated as high-quality and up-to-date.

In this presentation I introduce our production system from the viewpoints of quality controls and accuracy. I focus on the processes after data collection; checks connected to the data collection form and quality controls are excluded from this presentation.

Production system from the user's perspective

The monthly production of the Labour Force Survey is run with an application programmed here at Statistics Finland. It operates by the following principle:

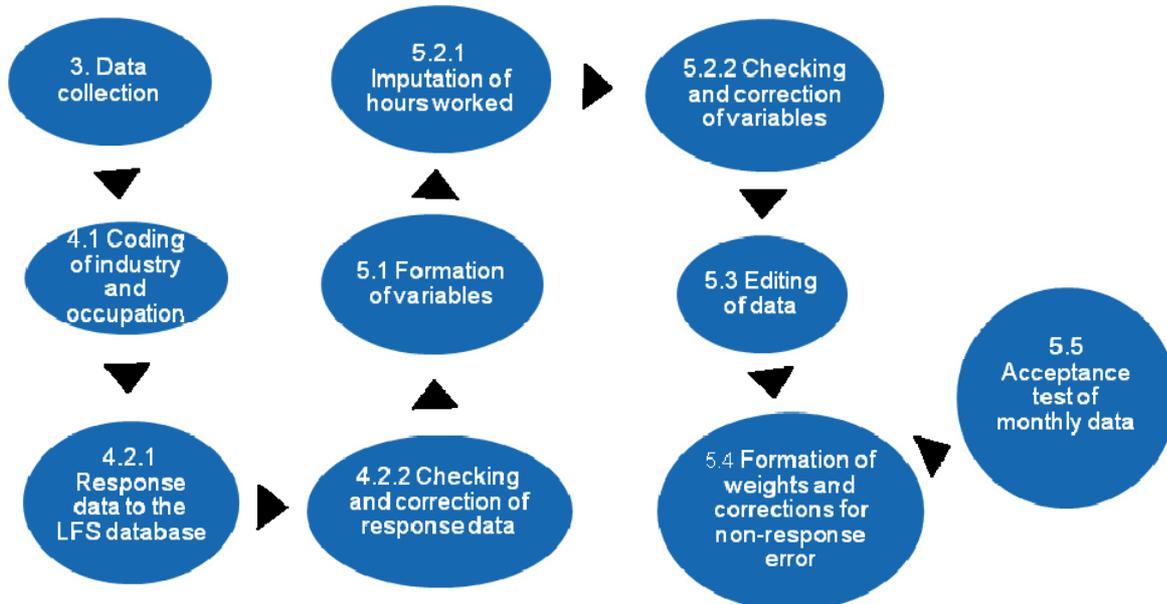
Login => Selection of time period (year/month) => Selection of use case => Run! =>
Report on display => Acceptance/Rejection of report => Storing of data in database
=> Next use case

The expertise of an LFS expert is thus mainly needed in evaluating the results - the technical implementation of runs does not require any special skills. In addition, so called 'task management system' is built to the production system to control that the use cases are made in the correct order (the system will not allow to run a use case too early or too late, only the next one in line). In this sense the system is 'foolproof".

Run reports include distribution data on the use case and the corresponding results of the previous month (as comparison). The reports are saved in the report storage where they can be later examined. Because the whole production system is in the SQL database, during production runs it is easy to examine previous information of the same respondent, or similar variable combinations from the same or previous months.

LFS production process after data collection

Here I describe the monthly production process of the Labour Force Survey from data collection to data acceptance testing. The numbers refer to our process descriptions. After the figure I present briefly process parts, and especially as concern of quality controls and minimization of errors.



4.1 Coding of industry and occupation

Coding of industry and occupation is made with a separate application, that was taken into use a few years before the new LFS production system. The industry and occupation are searched for all those interviewed for the first time and for those whose job has changed between the interview rounds. There are around 2,300 targets per month to be coded.

4.2.1 Response data to the LFS database

The interviews are conducted by Statistics Finland's interviewer organisation and at first the interview data are stored in their database. When data is moved to the LFS database, a set of automatic checks and corrections are made, such as:

- Response data are formed or copied for disabled persons and conscripts;
- Education data are corrected for those aged 15 to 21 (education during the past four weeks);
- If no responses to the first three questions, the respondent is moved to non-response:

4.2.2 Checking and correction of response data

In this use case a checking and correction process is made to the response data, which includes data and value range checks and a few logical checks. The tool is an editor that brings all response data for the target to be checked on display. With the editor the corrections can be made directly to the original data. When needed, new checking rules can be easily programmed. As many data checks as possible should be timed to this use case, so that data checks would be made to the original (response) data.

5.1 Formation of variables

When forming variables, data are combined from three different sources: response data, job coded data (industry and occupation) and register data.

5.2.1 Imputation of hours worked

Averages are needed for hours worked, therefore unknown hours worked need to be imputed. Imputations are made for around 10 to 25 employees per month (less than 0.5% of all employees), so this procedure has a very small effect on the total number of hours worked.

5.2.2 Checking and correction of variables

A similar correction process is made to the variables as to the response data. Logical relations between two variables are checked (for example, employer type with respect to occupational status: a self-employed person cannot have public sector as the employer type). National and EU variables are run in separate checking processes, so there is a risk that there remains a conflict between national and EU variables. For this reason major changes are no longer made at this stage.

5.3 Editing of data

Before calculation of weights, the following three processes must be made:

a) Increasing days worked and hours worked to the monthly level.

The number of days worked and hours worked during the survey week is increased to the monthly level based on various coefficients that depend on how many days the target person has worked during the survey week and whether he/she worked on Saturday or Sunday.

b) Editing of preliminary population figures.

Marginal distributions of the population are needed for calculating weights. They derive from the (preliminary) population data of the last day of the month preceding the survey month. When making weights, we use optimally recent population distribution by region, gender and age group.

c) Handling of the jobseeker register

Each month job seeker register is delivered as a line transfer from the Ministry of Employment and the Economy. Data on the register's job seeker status are linked to respondents, which are used as one marginal distribution in weighting procedure.

5.4 Formation of weighting coefficients and corrections for non-response error

The weights are calculated in two stages:

1) Post-stratification. The sample is weighted by the most recent distribution data: age, gender and area (see previous sub-section 5.3/b)

2) Calibration. Week standardisation and calibration with job seeker register data are made (see 5.3/c). Calibrated weights differ as little as possible from post-stratum weights (correlation > 0.9). Calibration was taken into use in 1997 after careful study and it was thought to have the following benefits:

- * Standard error of unemployed becomes smaller
- * Calibration corrects the bias caused by non-response
- * The LFS estimate for unemployed and the register's unemployment figure get closer.

5.5. Acceptance test of monthly data

The last test before accepting the monthly data is to run and check our publication tables. If the distributions of the publication tables seem reasonable, the user accepts the monthly data, and saves the data in the tabulation database.

Evaluation: The production system from the perspective of quality controls and error sources

For quality controls and minimisation of error sources the new production system - which is not exactly new anymore - involves several improvements from before.

The major improvement is that in the newer system all process stages go through an LFS-expert (not an IT expert). The LFS-expert has a good idea of the whole process and all its steps, which is important if there is something suspicious about the numbers. There have been no errors requiring afterwards correction during the new production system, which can be regarded as excellent performance.

The basic setting of checkings is reasonable. Checks are made in separate processes to both response data and variables, and the system is flexible: it is not difficult to make new checks. Two problems are known, however: 1) Corrected variables are not 'flagged', for which reason we subsequently started to record corrections on an excel spreadsheet. This double-entry bookkeeping is a somewhat unnecessary stage, which should be discontinued. 2) Certain corrections could not be timed to the response data, but only to the variables, which causes a small risk for the compatibility of the data. There are variables with the same content in both national and EU variables, and the system users must take care that there is no contradictions between the two.

In monthly production, we have not been able to get completely rid of our dependence on application engineers (IT), and in 'fatal error' situations we have to call them for help. Annual leaves also need to be planned so that sufficiently skilled support is available during production runs.

In all, the reviewed LFS production system has nevertheless operated well. Our press releases have not once been late, and we have never needed to correct any released data retrospectively. In that sense, checks and quality controls of Finland's Labour Force Survey are satisfactory.