**Quick statistics - how to deal with quality?**

Pertti Taskinen, Statistics Finland, pertti.taskinen@stat.fi

**Abstract**

*Labour Force Survey needs a reservoir of information which is collected directly from the interviewees. Telephone interviews are often quick and even hasty. In this presentation, the object is to tell what kind of quality checks are made to ensure the validity of the data which forms the official monthly unemployment rate - a quick and important statistics. Mostly, checks are mode independent, but the increasing use of the web data collection mode will be taken account in this presentation. During the data collection stage, the best guarantee of the quality is a trained interviewer who uses the data collection software with certain question-specific rules or instructions. The instructions of the interview should be as simple as possible and to be adapted as commonly as possible. For the Labour Force Survey, it is also recommendable to use the data from the previous round to the next one. This makes an interview easier for both sides - an interviewer and interviewee - without to endanger the quality. At the end of each survey month, the data production model uses three separate programmed checks: one for the collected data, one for the domestic variables and one for the EU variables. The software at the data production has been made user-frendly. If there are no unexpected problems e.g. with databases, the software program is possible to execute over a day. Basically, the figures are then ready but during the dissemination step some working days are needed to draft the publication and translations, and to update databases. In this paper, the data production model of the Finnish Labour Force Survey is also examined: could we do something even quicker and possibly in a little less resource demanding way with the existing quality? And eventually, what is the best practice to recommend?*

**Keywords:** Labour Force Survey, data collection, monthly production

**1. Survey Data Will Remain in Labour Force Statistics**

Despite the data available via registers and an extensive use of new methods, like data mining or big data, a part of the statistical data is still needed to collect direct from interviewees. This kind of data collection will not disappear in many cases. Eurostat e.g. demands that survey data is obligatory when collecting information about the labour market status at the reference week, not only under the current regulation but also under the upcoming IESS regulation which includes the labour market domain. Although Statistics Finland is developing the usefulness of the administrative data the importance of the survey data is not going to vanish.

The survey data are often connected to the quick dissemination. Anyway, compromises are not allowed at the costs of quality. In this paper, the quality of the Finnish LFS data is examined by describing the data production process. Table 1 contains some key figures on the Finnish LFS, considering the survey process and its history, from sampling to statistical values of the main estimates.

Table 1. Key numeric information of the Finnish LFS

|  |  |
| --- | --- |
| Yearly individual gross sample size | 150,000 |
| Number of waves | 5 |
| Length of panel | 15 months |
| Size of household data (2017) | 49,000 |
| The number of response variables (EU data) | c. 85 |
| Overall response rate (March 2018) | 67.4 % |
| Relative standard error for employed (March 2018) | 0.7 % |
| Relative standard error for unemployed (March 2018) | 4.0 % |
| Year of beginning | 1959 |
| Comparable time series from | 1989 |
| The number of active interviewers (April 2018) | c. 130 |

**2. Different sources of data**

*2.1. Register variables and**identifiers*

Initially, the source of the sample data for the LFS is the population register, which contains demographic background data on individuals (e.g. age, gender, nationality, region). During the preliminary phase of the survey, many identifiers are to be made for each individual. They are necessary to contact the survey, e.g. information on the reference time and the survey wave, and later on, the weights for the estimation. Also control codes are needed to handle each individual target throughout the survey, and eventually, the technical information on the interviewing situation needs its own variables. In all, this stands over 50 variables.

Administrative data on the highest level of education are received from the Register of Completed Education and Degrees, however, the data has a delay of one to two years. Contrary to that, the data of unemployed jobseekers are driven from the end of the previous month so it can be exploited also for monthly statistics. At the moment, demographic background information and above mentioned two other registers are the main extraneous sources for the Labour Force Survey in Finland.

*2.2. Response variables*

Response variables mean all the information collected in the interviews. The LFS variables are determined for the most part by the EU regulations. On the national level, only a few questions are added, so that the respondent burden would not increase. National variables are mostly in line with the EU variables which helps the survey process.

**3. Timetable for Labour Force Survey – Monthly Data**

*3.1. Before data collection*

The yearly maintenance of the (Finnish) Labour Force Survey, including the declaration of the timetable and other preparations for the next year survey, is to be done every year in good time. The sample is drawn twice a year. In this paper however we concentrate on the items connected directly to the monthly operations.

The overall process is contacted by the GSBPM model (UNECE 2013). Accordingly to the model, the process rules are described in detail in two documents: the use case standards and the processing rules, which are administered by the LFS unit.

Around three weeks before the reference month, the monthly sample is divided to the reference weeks, and the sample is filled with the beforehand information (register data and identifiers, see above 2.1.) and is divided among the interviewers. This stage takes around two days and after it, the search for the telephone numbers is to be made. Questionnaires, including the reference data, are ready around two weeks before the survey month. The sample and the questionnaire are to be delivered to the interviewers in good time so they can make necessary preparations for the data collection.

*3.2. Data collection period*

For every reference week there is a certain time for the fieldwork, i.e. the data collection of the response data. Usually, that is a period of two and a half week. At the end of the month however the data collection period only lasts one and a half week. Immediately after the interview the data is also prepared by coding the branch of industry. The coding of the occupation is usually done by the interviewers during the interview.

After the last interviewed cases in that month, the compilation of the data will be drawn at the last time. Typically, this will be done around the 10th-15th day after the reference month, depending on the calendar.

*3.3. After data collection*

The output run is a crucial phase to turn the raw data into the published ready data. This phase takes one day, unless, IT problems occur which can slow down the process. The output run will be handled in more details later in this paper.

Basically, the results and estimates are ready after the output run. However the preparations for the news release and free of charge data tables take 6–7 working days, including some safety time in case of unexpected problems. Table 2 shows the main details of the phase immediately after the data collection, listing days before the monthly press release.

Table 2. The time table of the monthly data production in the Finnish LFS

|  |  |
| --- | --- |
| **Day no.** | **Task(s)** |
| D-8 | End of data collection (period of 4 or 5 references weeks); compilation of the data |
| D-7 | Coding of the last cases. Controlling and executing of the data (output run), formation of weight etc. > acceptance of the data |
| D-6 | Preparation of the release (draft) |
| D-5 | Updating of database tables |
| D-4 | Seasonal adjustment. Compilation of release > translation |
| D-3 | Analysis of the results |
| D-2 | Translation ready; checking of release texts/figures |
| D-1 | Unit meeting concerning the release |
| D-Day | Press release at 9 am |

**4. Pre-filled Questionnaire**

*4.1. Questionnaire rules*

Let’s go back to the data collection phase and take a look at the main tool of data collection, the questionnaire. At the basis, questionnaire rules are a part of the valid data, including right filters, permissible values, signals and warnings for the unlikely values. The questionnaire must be logical to all labor market status – for employed, unemployed, or inactive. In the target population, each one belongs to one of them.

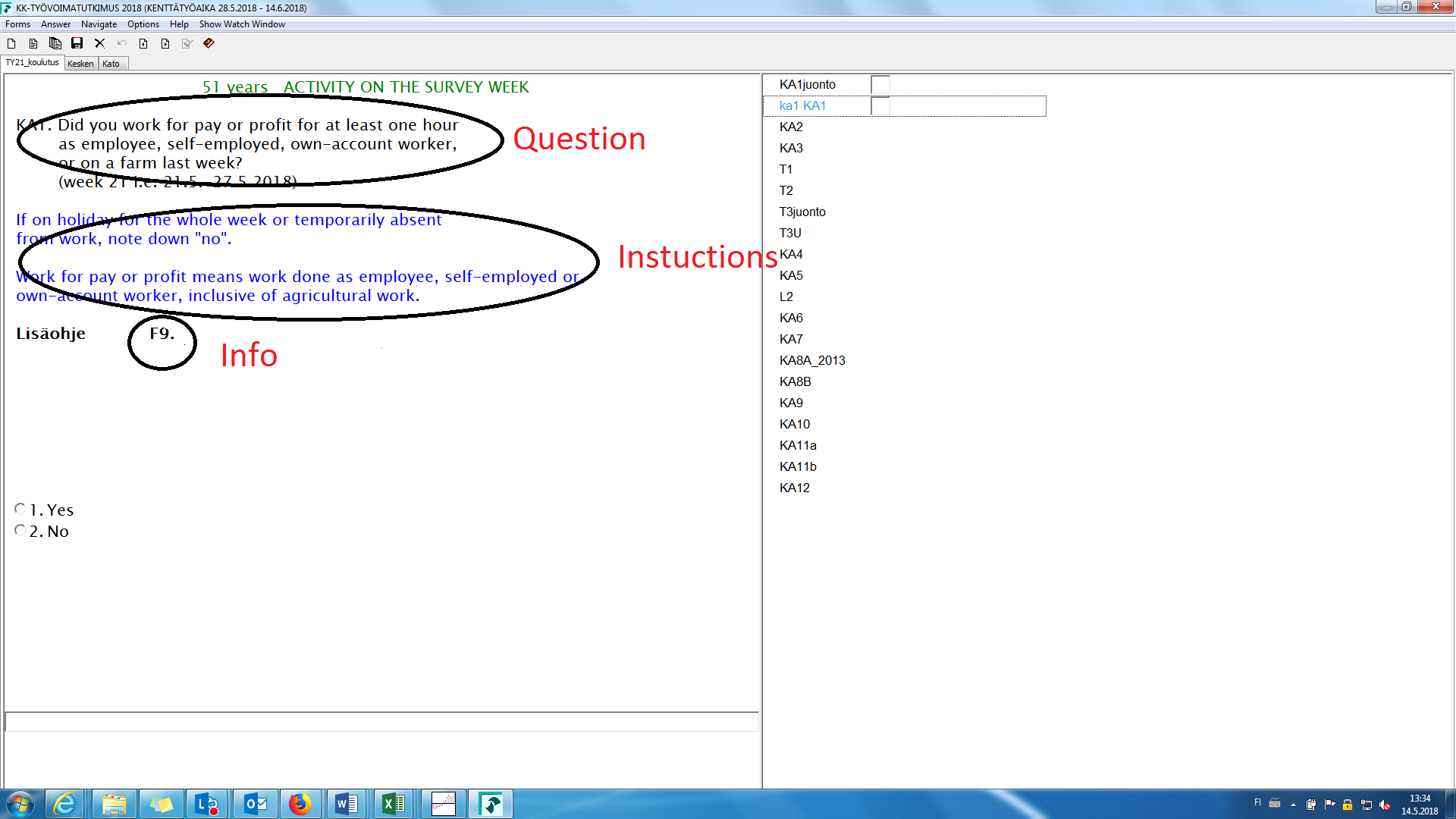
In case of an unclear answer, an interviewer can also make a remark for the further examination, or make a contact with a content expert by phone, instant messenger or email.

Due to a panel character of the LFS (in Finland 5 waves), the data from the previous round(s) are exploited during the next interviews. Information considering the working place of the person and some other details are given at the questionnaire, based on the previous interviews. The beforehand filling of the questionnaire must, however, be checked by the interviewer during the interview, for the insurance of the quality. This kind of pre-filled questionnaire, or dependent interviewing, makes the interview quicker and respondent-friendly.

*4.2. Questionnaire instructions*

For the memory back up, or for the rare case events, interviewers need some instructions at the questionnaire so that the procedure would be uniform among them. At the top edge of the questionnaire page is an actual question to read out. Below that, the essential notes are given. In case of more difficult appreciation of the right answer, additional instructions can be found in help box when pressing F9 (Picture 1).

Picture 1. A page of the LFS questionnaire: the location of the question and the instructions.



**5. Meaning of Interviewer**

*5.1. Telephone interviews*

The interviewer is expected to catch the household or a person who is dedicated to the survey. However, the attaining of the reference person is of course not enough. Interviewers are trained to conduct a good quality interview, using the interviewing methods, rules and instructions. Also the experience of the interviewer is a source of the quality. The interviewer must understand the content of the survey, answer interviewees’ questions and conduct the survey sometimes with a reluctant reference person.

The utmost verifiable quality guarantee in each and every interview is extremely difficult to reach, but the point is that the interviewers have a professional character and ethics.

*5.2. CAWI interviews*

In the near future, the LFS interview will be possible also with the web option. In the case, the telephone mode is planned to apply at the first round (in some cases, the CAPI is also used). That is supposed to help people understand questions better at the later round when they fill the questionnaire themselves in the web.

The content of the CAWI questionnaire comparing with the CATI questionnaire is as similar as possible. However, it is obvious that seeing and reading the questions (CAWI-mode) is different from hearing them (CATI-mode). The mode effect on its harmful meaning is to be minimized by introducing the LFS firstly with a telephone interview and afterwards, during the later rounds, to offer WEB option with dependent interviewing, i.e. the questionnaire containing certain data from the previous rounds.

This principal leads to an idea that interviewers are not the needless part of the data production in the upcoming era of the CAWI mode, or preferably, the mixed mode age.

**6. Output run**

The data production phase called “output run” contains data validation with three step check programs: for the response data, national variables and EU-variables. Eventually, the output run ends with the acceptance of the data.

Most checks should be done for the response data at the first step. However, due to the logical questionnaire structure and permissible values the number of corrections is very limited. During the interview, typing errors with dates at the questionnaire are potential and they will be corrected, if possible, in this step. Also some relations are checked: e.g. if a person is employed and the branch of industry is still missing, it will be checked and corrected.

All corrections at the output run will be executed to the original data. Sometimes it has noticed that a questionnaire contains too much missing information, and then the data for this individual person must be rejected; but, if only single value remains unknown that is not the obstacle to accept the data.

Besides the manual checking and corrections, automatic checking and imputation are made for unknown hours worked, based on the average data according to occupation and industry. Also cross-checking for the EU-variables is automatic, and in this later step, only seldom inconsistences are noticed.

Other checks during the output run are checking the amount of accepted answers and comparing it with the previous month, and furthermore, compare certain figures to the corresponding figures from the previous month. After the checking of publication tables, if the figures seem to be in order, as they usually are, the data will be copied in the tabulation database.

**7. What to do sooner, better or cheaper**

The LFS is already quite a quick statistics at the national level, and at the EU level. It is not realistic to suggest that the dissemination could be fundamentally quicker in its present (and upcoming) form.

The better statistics would mean that the quality would be even better; technically, e.g. the smaller standard deviation for the estimates, better calibration methods fixing the non-response error and so on. This kind of work never ends with a survey. As a better statistics in the sense of content, the LFS has been setting out for the future as the IESS survey.

Automation in some sense is a way to a less cost consuming statistics, using e.g. automatic coding of the branch of industry, or, in some sense WEB option on the interviews.

**8. References**

UNECE 2013. Available at: <https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0?preview=/97356247/97519760/GSBPM%205_0.docx> (Accessed 14.5.2018)