FUNDAMENTAL STAGES IN DESIGNING PROCEDURE OF STATISTICAL SURVEY

PÉTER PUKLI

The mission of National Statistics Institutes (NSIs) is to meet the statistical needs of the different user groups. Consequently, NSIs are required to undertake a complex range of operation such as collection, processing, storage and dissemination of statistical data, in other words they organize and controll the production process of statistics. The precondition for an efficient and satisfactory outcome is to design the whole process in advance.

The standard of the survey design procedure ensures that the setting up of new surveys as well as the redesign of existing ones are based on an appropriate professional preparation and at the same time it is a guarantee for the users that they can get reliable and timely information of their interest.

The first stage is the exploration of user needs and the investigation to what extent they can be satisfied. In general, the following main groups of users can be distinguished:

– governmental agencies,
– businesses and associations, foundations and other non-profit institutions,
– research institutions,
– general public,
– international organizations.

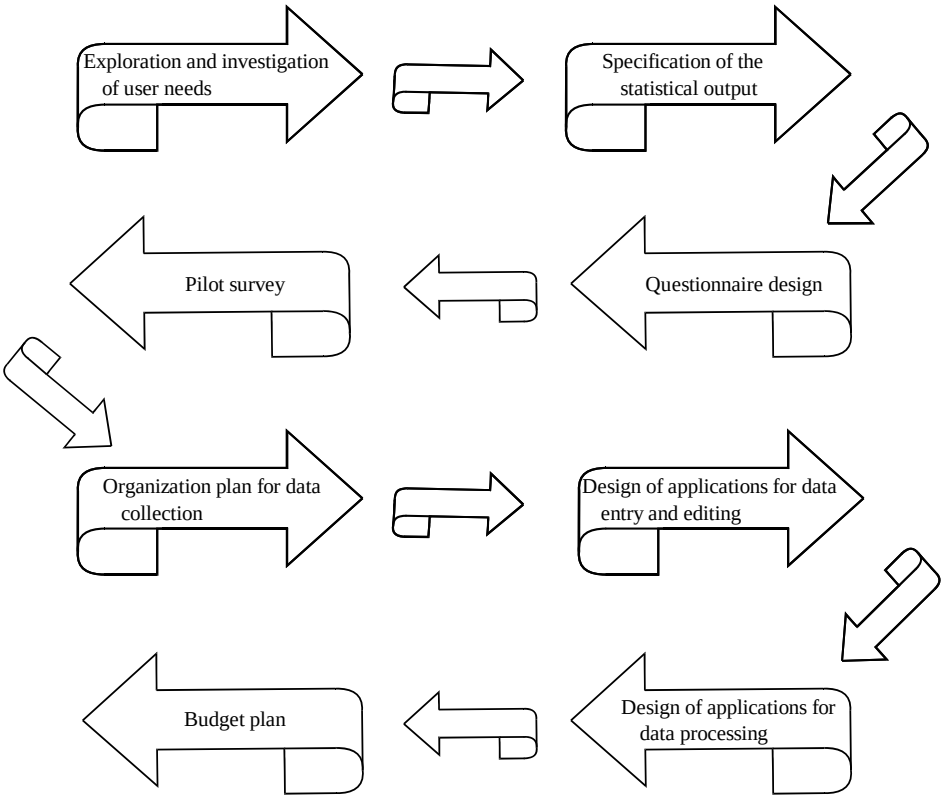
The applied methods for measuring user needs can be diverse relating to the above listed groups, but there are some common characteristics to take into account:

– the statisticians should assist and guide the user to express his needs explicitly and in unambiguous terms;
– to be able to do this the statisticians should be familiar with the intended applications of statistics;
– the users should be introduced to any constraints raising difficulties and additional cost to meet their needs;
– if different user groups indicate deviating needs on a certain issue, the statisticians should try to establish consensus by consultations, during which the users can exchange views;
– the users should be addressed to rank their wishes in order of preference relating to detail, accuracy and quickness.

Besides it is useful to contact the potential data suppliers to examine their full response burden.

Following the measurement and interpretation of user needs, the preparatory work can be started consisting of the following stages presented at the figure.

Stages of preparatory work for implementing statistical surveys



Specification of the statistical output

The final product of the statistical work is realized by releasing the data. The mode and the tool of dissemination has to be selected by taking the potential user group(s) and the costs into account. The specification of the intended statistical output is a key stage of the statistical work and has a determining effect on its whole process.

Before designing the set of tabulations fitting the user needs the target population has to be delineated. This step has a considerable effect on the whole design process, since it determines the survey method including the selection of the statistical register(s) used and the data collection method. It is followed by the choice and definition of the variables which are suitable for measuring the economic and social flows of interest.

This stage of the design process results in the set of tables, helps the analysing work of the statisticians and satisfies the needs of the final users.

Questionnaire design

The well-designed questionnaires – fitting the data collection method and the burden of respondents – can guarantee to meet the survey objectives.

Before starting the questionnaire design, the following things must be determined:

– the scope of the survey,
– the potential data sources of the respondents,
– the method of the data collection (medium and mode),
– the way how to implement the data procession.

For testing and designing the questionnaires it is recommended to establish a working group composed of the questionnaire design experts, the informatics specialists and the representatives of the respondents to attain jointly the optimum version.

The formulation of the questions must be clear and understandable. The unfamiliar phrases, terms, abbreviations must be avoided. In case of questionnaire items the data required are available in the accounting information system of the respondents, the appropriate reference is needed. Also those questions have to be avoided to which various answers can be given simultaneously.

Response burden should be an issue to take into account. In general, the staff for completing the questionnaires is available in the big businesses, while the smaller ones lack of it. The "double-questionnaire-design" (one version for smaller enterprises another for big ones) is the practical way to approach this problem. The questionnaire for smaller ones can be more simple and shorter than the other version. It can improve the response rate.

The sequence of the questions must be developed:

– to guide the respondent from question to question clearly;
– to reveal the respondent the possible information sources;
– to rank the questions in a logical way,
– the questions of similar content to be grouped in a section.

The explanatory notes must be understandable and guide the respondent in filling the form. To enhance the willingness to provide the data, the explanatory material or an introductory letter must inform the respondent about the goals of the survey, the use of the data required and the way how to ensure the confidentiality of data.

Pilot survey

Before introducing a new or a renewed survey, a pilot survey must be carried out in order to test the designed survey. It comprises:

– the quality check of the questionnaire including the explanatory notes;
– defining the size of the field work needed;
– testing the editing and processing applications;
– a clear insight in the cost and efforts needed;
– obtaining sufficient information about the expected response rate.

The size of the pilot survey has to be determined with regard to the characteristics of the target population, of which the most important factor is the degree of its homogeneity. The sample, whose number is usual between 100 and 300, must be selected from the target population by specifying some significant criteria from a statistical point of view:

– e. g. kind of activity, size, legal form, accounting system, regional delineation in domain of economic statistics;
– in case of social statistics, type of institutions and settlements, size of regions, social composition of population etc.

The questionnaire including the explanatory notes and other accompanying materials have to be evaluated in the following context:

– the questions are clear and understandable;
– the sequence of the questions promote the questionnaire to be completed in the desired way at a reasonable time and effort;
– the instructions can be interpreted easily;
– the information about the time is needed to complete the separate sections of the questionnaire.

In case of two or more questionnaire versions to find the optimum, the sample can be divided by these versions during the pilot survey. The most practical way to use this method when two versions are available, consequently the respondents are divided half-and-half.

The pilot survey also results in the information about

– whether the respondents' bookkeeping system can support the statistics needed,
– their willingness to provide the data.

As a result of the pilot survey, the statistical staff involved in the design process analyses the items of the questionnaires completed, evaluates the experiences with the field work and reviews whether the method used for the data collection meets the expectations, whether the training of the interviewers were thorough, which questions were typically misunderstood or incorrectly completed, whether the co-operation and the information exchange was sufficient between the management of the field work (regional offices) and the unit(s) of NSI responsible for collecting the statistical data of interest.

The pilot survey provides the opportunity to test computer programs for data processing. In this phase the following things must also be under consideration:

– whether the questionnaires must be revised and coded manually;
– the application for data capture and editing is user-friendly enough;
– the cost and time of the data capture.

Experiences from the pilot survey leads to the modifications in the design process, but they may also lead to a reduction of the level of output, both in terms of contents and scope, and indicates where new approaches and other methods should be applied to achieve the objectives of the survey.

In each stage the time and the cost of the tasks must be counted separately, which can serve as a basis to extend these calculations to the whole survey. It may occur, that the analysis implies the volume scale of the questionnaire or the number of the sampled units should be moderated to ensure the consistence with the available resources.

Organization plan for data collection

In this stage the steps of the data collection process have to be defined. It comprises:

– defining the survey frame population, in case of sample survey the sample design, as well as selecting the sampled units;
– specifying the data flows and the management tasks relating to the survey and organizing the field work (e. g. determining the number of the interviewers needed, elaborating the training program for them);
– preparations for questionnaires dispatch (including the attached materials),
– receiving the questionnaires completed by respondents, completeness check;
– specifying the method in case of non-response.

Ways to define the frame population or the sample frame are the following.

– By using an algorithm for selecting the units observed. The selection process based on the business register parameters (e. g. sector, size class, legal form, demographical factors etc.) This is the most effective and flexible selection tool, because the changes in units and characteristics over time are also recorded in the frame population (sample frame) automatically;
– On the basis of a former survey, when the criterion of selection can be embodied as an indicator within a maximum and minimum value;
– Individual selection scheme, that means a list of all elements of the frame population (sample frame).

The mode of the data collection (by mail or interview) is already chosen in the earlier stage of the design process, at this point the organizational exercises linked to the dataflows must be worked out to answer

– where the questionnaires are received (regional office, department of NSI);
– where the data entry and editing phase are intended to implement;
– where the data processing are performed.

It is needed to elaborate

– the organization plan for selecting the interviewers, and the mode of their training program;
– the timetable of survey;
– the mailing list (names, addresses and telephone numbers of the reporting units);
– initial contact letter informing respondents in advance about the survey they will be involved in.

At the same time, the statisticians have to specify

– the number of copies of the questionnaire and the other accompanying materials;
– the way how to produce them (printing, copying etc.);
– the schedule of their production and dispatch.

The dispatch can be implemented by

– creation of a dispatch file (from the statistical register) containing the records with each unit surveyed;

– application of a list for mailing purpose which can be compiled by using the outcome of another survey carried out earlier;
– interviewers.

The receiving of the questionnaires completed by the respondents and the completeness check comprise

– collection and registration of the entering questionnaires;
– completeness-check relating to two types of missing information (unit non-response and item non-response);
– determine the reasons of non-response;
– specifying which part of missing information is needed to impute;
– what sort of information are necessary for completeness-check.

Design of applications for data entry and editing

When preparing the data entry and editing process, it must be explored

– which statistical register(s) will be used,
– which nomenclatures, which versions of them are needed for processing the data of the questionnaires,
– whether the metadatabase covers the nomenclatures needed regarding the data collection.

In case if the answer is “no” for the latter question, the metadatabase must be extended to the appropriate version of the nomenclature needed for data collection.

Recommendations for setting up the list of edit rules:

– it is needed to check the correctness and validity of the identification code of the unit observed and to test the relationship between the codes and the respective items,
– it is needed to apply valid values checks and range checks of the data as well as checks taking the form of a ratio between two variables, which should be within specific bounds;
– it is practical to compare the survey data with the same figures at the point of time *t*-1 and with data from other surveys;
– it is needed to apply the arithmetic check based on specifying that the sum of variables should be equal to the total item, if any other relational checks can not be developed;
– each editing rule must be corresponded with an identification code of the error connected with a short message;
– the errors detected must be ranked in the order of significance. For example the following categories can be used: *1*. A warning that the item might be wrong (soft errors). *2*. The erroneous item is revealed and marked, but not corrected. *3*. The error must be corrected.

The following different methods can be applied with respect to the data entering and editing:

– data capture by Optical Character Recognition (OCR) systems, automated editing;
– data capture by OCR systems, correction done by the data typist entering the data form;
– computer assisted data capture and correction by the statisticians, the interviewers or the respondents.

In the specification of editing rules

– an algorithm linked to the different type of errors is needed to provide, when automated correction is used;
– in case of correction by the data typist, the staff has to be instructed the way the correction is implemented;
– a set of tables has to be developed which summarizies the experiences with the data entry and editing.

When setting up the system for the data entry and editing, each application and procedure must be documented, it is the precondition for securing the smooth-running operation.

Design of applications for data processing

The result of the data processing is the intended statistical output fitting the survey objectives. This output can be embodied as

– electronic dissemination (diskettes, on line transmission, etc.),
– printed publication.

The activity of data processing comprises

– creation of the output database,
– imputation relating to unit and item non-response,
– the sample estimation,
– aggregation of micro-data,
– retrospective correction.

When elaborating the application of the data processing, the following questions are to be answered:

– is it necessary for the statisticians to be able to reach the individual data on line;
– whether the missing information to be or not to be imputed and what sort of methods for imputation can be employed in case of unit and item non-response;
– is it necessary for the statisticians to be able to reach the micro data of the sample;
– what estimation method should be developed in case of sample survey;
– if the database has to operate on micro data level, should it cover each item of the questionnaire or only a part of it;
– which aggregation levels are the most commonly used and are the nomenclatures relating to these levels available or not in the metadatabase;
– what indicators should be calculated on micro or aggregated level;
– whether the retrospective correction has to be implemented and what rules are connected with it.

The specification of the data processing must be stored in the computer. Here the main aim is to ensure the easy access. In case the specification is altered, the new version of the documentation including the highlighted changes have to be completed. Titling has to refer to the identification code of the subject and the year of starting point when the given specification was applied at the first time.

Budget plan

Planning the total cost of the surveys helps the decision-making and provides information for the annual budget plan of the statistical agency.

The survey costs have to be calculated in two forms, accordingly a simple and a detailed versions of budgetary plan are compiled. The simple one is derived from the investigation of user needs to what extent they are proven and can be satisfied,

meanwhile the detailed one is incorporated in the design process and based on the information of each stages.

The survey costs have to be divided into groups as direct costs and indirect costs. Direct costs can be measured and listed in the budgetary plan. These items have to be classified by the place where they incurred (the units of NSI), by the time and by their type (wage and salary as well as other expenditures).

*

Regulations on statistical survey design is an important tool for planning and controlling the survey process. The survey design is a chain of actions aiming at satisfying the user needs and minimizing the burden of the respondents. It is the task of the statistician to find the consensus between these two conflicting interests.