This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: The Populations, Frames, and Units of Business Surveys

Contents

General section

1.   Summary

During the design and process of a *survey*, different terms are used for the population of the survey depending on how the user needs can be realised. For this reason *the population of interest, the target, frame and survey populations* can be differentiated. The more complete coverage of the target population we have, the better chances we get to meet user requirements and assure higher quality of survey data.

During the survey process the units of the target population, the observed units and the units being able to provide data can be different. The data collection phase has to manage the *statistical units* and data collection units of the given survey. They have roles in making contact with *data suppliers*, data providers, and controlling, monitoring and validating the expected and provided questionnaires.

The *survey frame* lists and identifies units of the frame population altogether with their contact information, economic and geographic classifications and size categories. For the selection of the sample, the *sampling frame* provides information for stratification. The survey frame is also the basis to get access to the data supplier and to personalise and mail the questionnaires. Furthermore, it helps to control and monitor the data collection phase, to register and validate responses and to urge and evaluate non-responses. The survey frame gives information for the weighting, analysing and micro integration phases of the survey process as well.

The theme summarises the roles and concepts of the above terms.

2.   General description

2.1   *The population of the survey*

2.1.1   *The population and the frame*

The design phase of the statistical processing flow (sub-process 2.4 of GSBPM 4.0) defines the population of the survey, specifies the objects, the units from which information is collected, the set of these units, from which information is needed and the set of units to exclude from the investigation. The population is tied to geographic place and time. The reference time is always part of the definition of the population (in the national statistics the geographic place is usually the country, therefore it is an inherent part of the definition).

Example: A survey wants to observe the enterprises undertaking building construction activities in 2011. In this case, the unit of the population is the enterprise; the extent of the population is defined by the building construction activities of the units and the reference time of the population is 2011. That will give the population of the survey.

The user needs cannot always be met for different reasons (because of cost, time, conflict in standards, conflict in different user needs and other constraints). In that case the *population of interest,* the population in the user needs is differed from the *target population* set in the design phase.

Example: The population of the above example can be considered as the population of interest, then some practical considerations can be added.

1. It is a decision whether the survey observes only the enterprises where building construction is the *principal activity*, and excludes those which do it as secondary activity (because, e.g., there is no information about them); or it observes all construction activities.

2. The way of the observation of the enterprises with very little contribution to the survey can also be decided (because of the relatively high cost). They can be excluded from the data collection.

If both constraints are applied, the target population includes the enterprises where principal activity in 2011 is building construction (enterprises in class 41 according to the NACE Rev.2.) and where the number of employees is more than 5.

Note: in case of cut-off survey (where the small units are not involved into the sample) it is proposed to give a model-based estimation for the excluded part. If this estimation is feasible, the enterprises with less than 6 employees also belong to the target population. Only if it is not possible, then these units may be excluded from the target population, because this solution increases the gap between the population of interest and the target population, and decreases the usefulness of the survey results.

To carry out the survey the list of the population units is needed that fulfils the prescribed properties in the given period. The *survey frame* describes the units with all those attributes that are necessary to their identification and to get into contact with them. One of the most crucial aspects of a survey is to find or compile this list because the completeness of the list greatly determines the success of the survey,

Usually the survey frame has the same unit type as the target population. In that case the set of the survey frame units give the so called *frame population*. But there are cases when they may differ from each other, if there is no possibility to ask directly the unit of the target population or the list of the population is not available. Then the survey frame describes the units through which information about the population units can be accessed.

Example: We need to prepare a survey frame that contains all enterprises in class 41 according to the NACE Rev. 2, with more than 5 persons employed. This set of enterprises gives the frame population. The survey frame for this frame population has to contain the identifier, name and address of the enterprises as a minimum condition for the accessibility of the population units.

Example: We would like to obtain information about the local units of the enterprises with construction activity (target population) but we do not have the list of the local units. In that case there is a possibility to set up a survey frame of the enterprises to be observed, and we can get the frame population by asking them about their local units.

2.1.2   The error of the frame population

For various reasons the frame population usually does not cover the target population and the difference introduces error to the survey (Eurostat, 2009, STATCAN, 2003a, Eurostat, 2008a). In the absence of units, some parts of the population might be out of the observation. In cases when some parts of the target population are missing, we talk about *under coverage*. When the listed units do not suit the properties of the target population, then this part of the population will be observed in vain. Sometimes certain units are duplicated in the list, possibly with different identifier. In cases when some parts of the population are not required for the observation but are still on the list, we talk about *over coverage*. Both under and over coverage distort the results and raise the cost of the survey. So during the preparation of a survey one of the most important tasks is to compile the list of the population units.

The common part of the target and the frame population is called *survey population.* That is the part of the target population for which information from the survey can be obtained (see Figure 1).
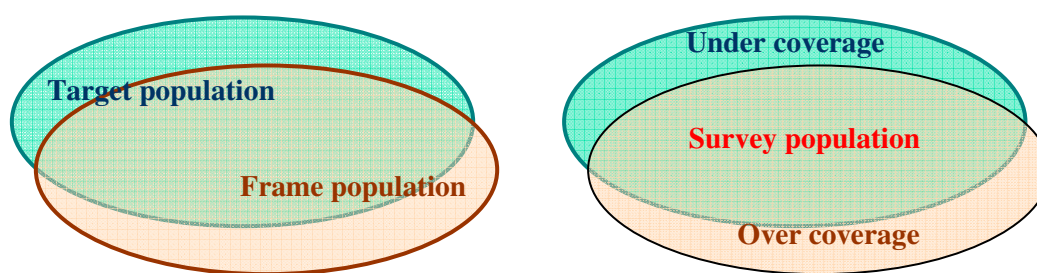

Figure 1. The relationship between terms of population

What are the causes of this coverage error?

–   One of the error sources may lie in the timeliness of the frame when the *reference period* of the frame does not coincide with the reference period of the population. In this case the frame shows an earlier status of the population units than it would be in the reference time of the survey. Therefore it contains dead units, new units are missing, and some attributes are out of date.

–   The selection attributes for the population units may be erroneous; therefore units are missing or are involved in the population needlessly.

A special aspect of the coverage problem is the inconsistency of the target populations within the domains of business statistics, when they have to be compared, processed together. The differences can be caused both by the time and other attributes of the target population. Another case of the inconsistency that causes coverage error is the union of different surveys (e.g., a certain statistical domain together for all European countries), when the target populations do not refer to the same time or they do not cover the whole population, or overlap each other. (See the horizontal and vertical integration in the modules "Statistical Registers and Frames – Survey Frames for Business Surveys" and "Statistical Registers and Frames – The Design of Statistical Registers and Survey Frames").

2.1.3 The source of the frame

An important task of the survey design is to find the suitable source for compiling the survey frame. Usually it is not an easy task, especially if there are no *register*s available for the target population.

- It occurs that the frame has more than one possible source (one can get information from the chamber, union of the given type of businesses or the tax office or other authorities). When selecting the suitable source one has to consider the coverage error, the accuracy of the source, the costs, the availability in time, the repeated usability of the source, etc.

- Sometimes the population cannot be covered by one frame using only one data source. In these cases multiple sources have to be used (for example one for business corporations and one for agricultural farms). These types of frames are named *multiple frames*. In case of using multiple frames, duplications in the population might be the most important error which should be avoided.

- The third case is when there is no source available and it is too costly to create a frame for the units of the target population. To observe the target population one has to find another unit being able to supply data about the unit of the target population. In that case the survey frame will contain these "responding units".

Example: A survey intends to observe the building construction activity in the settlements in 2011. In order to create a survey frame, different sources can be selected.

- One of the possible frames can be the list of the local authorities, as they have information about the construction activity on the base of building permits. Usually scarce and not up-to-date information is available from local authorities, therefore, this source might not be adequate in most cases.

- The best would be to have a list about the local activities of the enterprises. In this case those particular local units which have building construction activities can be asked to give the information needed.

- If there is no such list then the building construction enterprises have to supply information about their activities by settlements.

The observation based on the above frames can give very different survey results; therefore the choice should consider the possible outcome of the survey.

2.1.4 The sampling frame, subpopulation

It has been decided in the design phase whether the survey observes all units of the frame population (census type survey) or the selected units of the population only (sample survey). In latter case the survey frame serves as a base for sampling, therefore the survey frame will be the *sampling frame (also known as sample frame).* The observation will be made only for those units that are selected into the *sample* (see Figure 2).
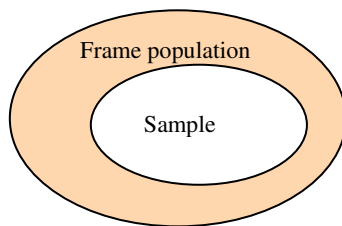
Figure 2.
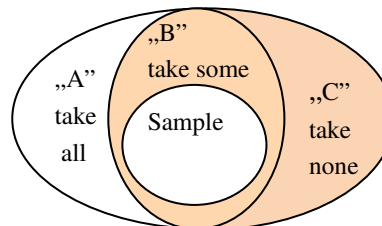*Survey frame as a sampling frame*

Figure 3.
A, B, C subpopulations of the frame population

Different phases of the survey process demand to create subsets of the population, because different processes refer to different parts of the population. These subsets can be defined by the attributes of the population units like *activity*, geographical position, size, etc. These subsets are called *subpopulations*.

Example: Let's assume there is no possibility to make a total enumeration on the enterprises where the number of people employed is less than 20. The observed units of the population will be constructed from three subpopulations. Subpopulation "A" contains the enterprises where the number of employees is at least 20. The full-scope of subpopulation "A" will be observed with total enumeration (take all units). Subpopulation "B" contains the enterprises with 6-19 employees which we want to observe by sampling. Subpopulation "C" contains the enterprises where the number of employees is less than 6 (see Figure 3), they are not observed. If it is possible, their data can be estimated based on, e.g., data collected from administrative source.

2.2   The unit types of the survey

In a survey different entities with different roles have to be managed. The two groups of the entities, units are:

–   the units of the data collection: those units that report, supply data and complete the questionnaire.

–   the statistical units: those units which the previous ones report about.

Usually the two unit types coincide with each other, but sometimes they are different.

2.2.1   The statistical units

The aim of the business surveys is to collect or compile information about the well-defined units of the economy. This collection serves as a base to create aggregates according to the common attribute types of the units. In the ESS, the cornerstones of statistics are the statistical units. The *statistical units* have to be used to collect, transmit, publish and analyse statistical data on the production system.

For internationally comparable statistics, the European Economic Community regulated the concept of the statistical units ((EEC) No 696/93) in 1993. In 2007 the UN edited a reference compendium presenting the different types of statistical terms, their definitions and adaptation in international

7

standards (UN DESA, 2007). Nowadays, in the ESS there is a verification process of the statistical unit types for the sake of international comparability and practical applicability.

The definition of the statistical unit in the above documents is the following: "*a statistical unit is an entity about which information is sought and for which statistics are ultimately compiled. It is the unit at the basis of statistical aggregates and to which tabulated data refer*".

The statistical units are defined on the basis of three criteria:

- the legal, accounting or organisational criteria

- the geographical criteria and

- the production criteria.

According to the above criteria the main types of statistical units are (EUROSTAT, 2010):

- *Enterprise group*

- *Enterprise*

- *Kind of activity unit (KAU)*

- *Local unit*
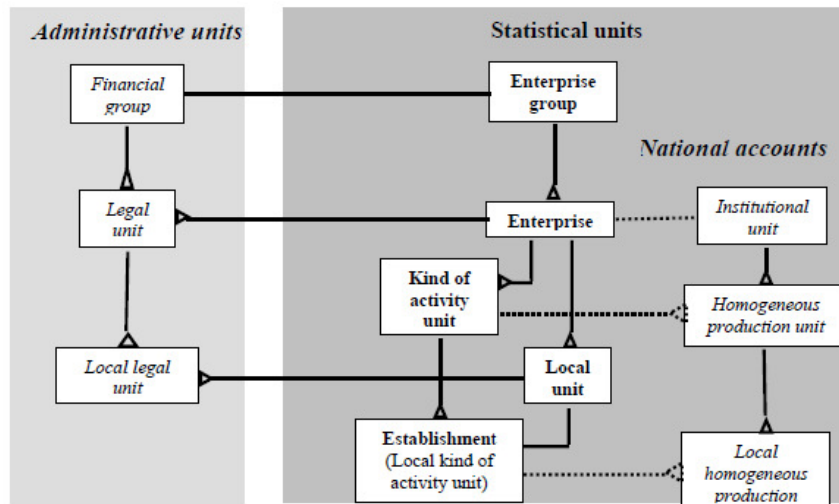
- *Local kind of activity unit(establishment) (LKAU)*

The statistical units can be created from the administrative units built on *administrative source. These unit types are the financial group, the legal unit and the local legal unit.*

To form the statistical units from the administrative units is not a simple process. Because of the huge number of the legal units currently in a lot of countries the legal units correspond directly to the enterprises. Complex correspondence between the legal units and enterprises is applied by some countries only for a small part of the legal units.

The National accounts statistics use other unit types as *institutional unit, homogeneous production unit, local homogeneous production unit*. The homogeneous production units are the so called analytical units that are created by statisticians, often by splitting or combining the observed units in order to compile more detailed and more homogeneous statistics as it is possible by using data on the statistical units only. Analytical units are not able to report data about their activities.

From data collection point of view only the identifiable legal/organisational or physical entities can be taken into account as population units, which are able, actually or potentially, to report data about their activities. The analytical units are the results of the statistical processing.

The connection of the three groups of the units is illustrated in Figure 4. The content and definition of the different types of statistical units are detailed in the theme module "Statistical Registers and Frames – The Statistical Units and the Business Register".

Note: A triangle means one or more entities can be linked to the entity on the other side of the line

Figure 4. Relationship between different types of statistical units (Source: UN DESA, 2007)

Selecting the suitable statistical unit of a given survey is not always unambiguous and simple. There are many factors which have to be considered:

- targeted level of data collection,

- the particular structure of the businesses involved into the survey,

- available business register structure in the country,

- the statistical data collection system in the country,

- regulations for the organisations of the businesses.

Besides the aim of the observation, other practical aspects of the data collection also have effects on the selection of the statistical units of the survey. For international comparison the aim is that the differences in the above aspects should be reduced among the countries so that the applied statistical units for a survey should be comparable as much as possible.

2.2.2 The actors, units of data collection

The units of data collection can be different from the units of the target population. There are several causes of this difference, for instance,

- because of book-keeping practices no data are available in the given unit, only at another level of the organisation

- another organisation is charged with reporting (for example: accountancy firms)

- the partial autonomy of the unit or dependency from another unit (the other unit could be responsible for reporting only)

- the costs of data collection demand to get pre-collected information about a group of the units

9

    −   there is no frame information about the targeted unit, only about the unit it belongs to.

As a consequence, for operational/frame purposes, a distinction is made between *observation unit*, *reporting unit* and *collection units*.

The *observation unit* is an entity about which data can be obtained and for which data are recorded. From different reasons the unit, that the questionnaire describe and identify, can be different from the observation unit. The data collection process has to manage rather the unit of the questionnaire than the observation unit. Therefore a new term is needed, the *reporting unit* to which the questionnaire is tied and for which the questionnaire is filled in. This is the case when, for example, information about a phenomenon is collected by products or by settlements within an enterprise. If the questionnaire refers to the enterprise and the different products, settlements appear as units within the questionnaire then the reporting unit is the enterprise, and the observation unit is the product or the settlement.

The reporting unit acts as the object of the statistical information, e.g., the identifier of an instance of a questionnaire. The reporting unit is the base for controlling the unit level response/non-response.

In other, e.g., OECD glossary of terms, Canadian Survey methods and practices (STATCAN, 2003b) - the *reporting unit* acts only as the supplier of statistical data for a given survey instance and the *unit of reference* is the object of the survey ("the unit about which information is provided"). This meaning of the reporting unit is called data supplier in this handbook.

The collection units are the data suppliers and the data providers. The *data supplier*s have primary role in reporting the surveyed information. They are the organisational entities which are responsible to provide data about the reporting unit. They are the units whom the NSIs are in legal relation with and whom the answer can be required from. The data supplier, in most cases, coincides with the reporting unit, but they can be different.

In the case of census type survey all of the units of the survey frame are assigned to data suppliers, in the case of sampling survey only the units belonging to the sample are assigned to them. The data suppliers involved into the sample give the *scope of data suppliers*. They have special role in the analysing of the response burden.

Sometimes the questionnaire is filled in not by the data supplier but by another firm (e.g., central administrative office or an accountancy firm) who provides this service to its client. Such information providing entities are *data providers*. In this sense data providers are either the data suppliers if they fill in the questionnaires themselves or the other firms who fill in the questionnaires on behalf of the data suppliers.

During the data collection the *surveying departments* need to be in contact with the data provider. Practically it has to know the personal data (name and other contact information) of the persons who are responsible for filling in the questionnaire. These persons are the *respondents* of the questionnaires.

Example: The survey observes the kind of activity units of the enterprises in the industrial sector. The data collection system expects the questionnaires from the enterprises; the enterprises report about the kind of activity units, except the enterprise charges a book-keeping agency with the reporting. In this example the statistical unit and the reporting unit is the kind of activity unit. The data supplier is the enterprise and the data provider is either the enterprise or the book-keeping agency, depending on who completes the questionnaire.

2.3   The role and attributes of the survey frame

The *survey frame* provides information for the identification and access of the units of the population in a given reference period as it was mentioned above. The survey frame is not merely a list of population units, but it also plays different roles in the different phases of the survey process. Usually five particular roles can be distinguished (Györki, 2012):

– The most typical role is related to the assignment of the population to the survey, in other words, the compilation of the survey frame. The given types of units are included in the survey frame for the given reference period with all the necessary attributes for further processing phases.

– If the survey is a sampling survey then the survey frame serves as a sample frame for selecting the reporting units and data suppliers. To create the strata (the subpopulations), mostly the classification and categorical attributes of the survey frame are used. The survey frame has to contain information on whether the unit is part of the sample.

– To access the collection units (data suppliers, data providers) and to mail the questionnaires in the data collection phase, contact information is provided by the survey frame. The personalisation of the questionnaires is built upon the survey frame as well.

– Completeness checking of the responses on unit level is also controlled by the survey frame. The validation of the responses both in the data collection and the data editing phases also uses the survey (or sampling) frame to control whether the reporting unit belongs to the population and to the sample. The validation rules of statistical data differentiated by subpopulations use the attributes of the survey frame.

– The imputation of non-responding units also use the information provided by the survey frame. In the case of sampling surveys, the survey frame provides size data of strata for estimations. The survey frame also contains classification and categorical information for grouping, aggregating the data observed.

In order to carry out the above functions, the survey frame has to characterise the frame units with different attributes. For the accessibility of the units, the minimum requirements for the survey frame content are:

– the identifier of the units, that assures unambiguous, unduplicated identification and

– their contact information (name, address, telephone, e-mail, fax, name of leader, name of contact person, etc.)

Categories and classifications usually support the stratification for sampling, estimation and the creation of other subpopulations, groups and aggregates. In business surveys, standard classifications have to be applied, particularly:

- *NACE classification* for the principal activity of the unit (presently NACE Rev.2)

- *NUTS classification* for the geographical position of the unit.

Beside the standard classifications, national classifications may further typify the units, for example, the *legal form* or the local administrative unit (LAU) of the geographical position.

It is a frequent demand to characterise the units by different size categories, typically:

- by the number of employees or

- by the value of turnover in national currency or

- by the value of turnover in Euro

Standard code lists of EUROSTAT may belong to the categories but they can also be applied or modified according to national demands. Such standard code list is available, e.g., for the number of employees.

The survey frame describes the population units. Besides the identification and classification attributes the survey frame can contain attributes on the role and strength of the unit within the population. For example after sampling, only a part of the units become reporting units and data suppliers. The "flag" of the units that are part of the sample and are expected to respond is also an attribute of the survey frame.

3.   Design issues

4.   Available software tools

5.   Decision tree of methods

6.   Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7.   References

Council of the European Communities (1993), Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community. *Official Journal of the European Union*, L 076, 30/03/1993, 0001–0011. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31993R0696:EN:HTML

Eurostat (2008a), *NACE Rev.2. - Statistical classification of economic activities in the European Community, 2008 edition*. Eurostat Methodologies and Working papers. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-07-015/EN/KS-RA-07-015-EN.PDF

Eurostat (2008b), *Survey sampling reference guidelines, 2008 edition - Introduction to sample design and estimation techniques*. Eurostat Methodologies and Working papers. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-08-003/EN/KS-RA-08-003-EN.PDF

Eurostat (2009), *ESS Handbook for Quality Reports, 2009 edition*. Eurostat Methodologies and Working papers. http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-EHQR_FINAL.pdf

Eurostat (2010), *Business Registers - Recommendations Manual, 2010 edition*. Eurostat Methodologies and Working papers. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-10-216/EN/KS-32-10-216-EN.PDF

Györki, I. (2012): GÉSA: The Tool for Survey Control, Quality Assessment and Data Integration. *Hungarian Statistical Review*, Special number **15**, 48–78. http://www.ksh.hu/statszemle_archive/2012/2012_K15/2012_K15_048.pdf

STATCAN (2003a), *Statistics Canada Quality Guidelines* - Coverage and frames, 17–20. http://www.statcan.gc.ca/pub/12-539-x/index-eng.htm

STATCAN (2003b), *Survey Methods and Practices*. http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.pdf

UN DESA (2007), *Statistical Units*. United Nations, New York. http://unstats.un.org/unsd/isdts/docs/StatisticalUnits.pdf

Interconnections with other modules

8.   Related themes described in other modules

1. User Needs – Specification of User Needs for Business Statistics

2. Overall Design – Overall Design

3. Repeated Surveys – Repeated Surveys

4. Statistical Registers and Frames – Main Module

5. Statistical Registers and Frames – Building and Maintaining Statistical Registers to Support Business Surveys

6. Statistical Registers and Frames – Survey Frames for Business Surveys

7. Statistical Registers and Frames – The Design of Statistical Registers and Survey Frames

8. Statistical Registers and Frames – The Statistical Units and the Business Register

9. Statistical Registers and Frames – Quality of Statistical Registers and Frames

10. Sample Selection – Main Module

11. Data Collection – Main Module

12. Response – Response Process

13. Micro-Fusion – Data Fusion at Micro Level

9.   Methods explicitly referred to in this module

1.

10.   Mathematical techniques explicitly referred to in this module

1.

11.   GSBPM phases explicitly referred to in this module

1. "2.4. Design frame and sample methodology" for frame design

2. "4.1. Select sample" for establishing the frame

3. All phases for register maintenance

12.   Tools explicitly referred to in this module

1.

13.   Process steps explicitly referred to in this module

1. "2.4. Design frame and sample methodology"

2. "4.1. Select sample"

3. All phases for the register maintenance

Administrative section

14. Module code

Statistical Registers and Frames-T-Population, Frames, and Units

15. Version history

Version	Date	Description of changes	Author	Institute
0.2	29-05-2012	first version as a separate module	Ildikó Györki	HCSO
0.3	16-07-2013	revised version after EB review	Ildikó Györki	HCSO
0.4	17-11-2013	revised version after EB second review	Ildikó Györki	HCSO
0.4.1	11-12-2013	preliminary release		
0.5	25-01-2014	revised version after EB third review	Ildikó Györki	HCSO
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:35