مـركـز الإحـصـاء
STATISTICS CENTRE

Process and analyze – from data to statistics according to GSBPM

Methodology and Quality Guidelines – Guide No. (23)

July 2023

Contents

Version 1.0 of this document was written Spring/Summer 2023 by Methodology Section, Statistics Sector, SCAD.

1. Introduction

These guidelines for processing data and analyzing statistics contribute to enhancing the methodological aspects of the production of statistics by following international standards and best practices. For all statistical processes, both for survey based-, and administrative based statistics, the Generic Statistical Business Process Model (GSBPM) can be used to describe the different phases and sub-phases of the statistical work.

This document presents the GSBPM and descriptions with focus on its phases related to the activities of data processing, and analyzing, are presented under section 2. Section 3 focuses more specifically on two types of surveys, for households and for business related statistics. Note that all sub-processes, including 5 and 6, depend on the design document. The design document, in its turn, depends upon the user's needs.

As mentioned, the focus of this manual is to guide the reader through the different phases of processing and analyzing, with examples from survey-based statistics. It is important, though, to remember that both survey-based and administrative based statistics essentially follow the same process according to GSBPM. Some obvious differences are, e.g., that we need no sampling method or questionnaire for administrative data.

One last remark – here we focus solely on two of eight processes, hardly mentioning the initial phase of the statistical production process which always is to make sure to understand the user needs and what the final statistical outcome will be used for. Without that information it is impossible to produce high quality statistical information.

2. The Generic Statistical Business Process Model

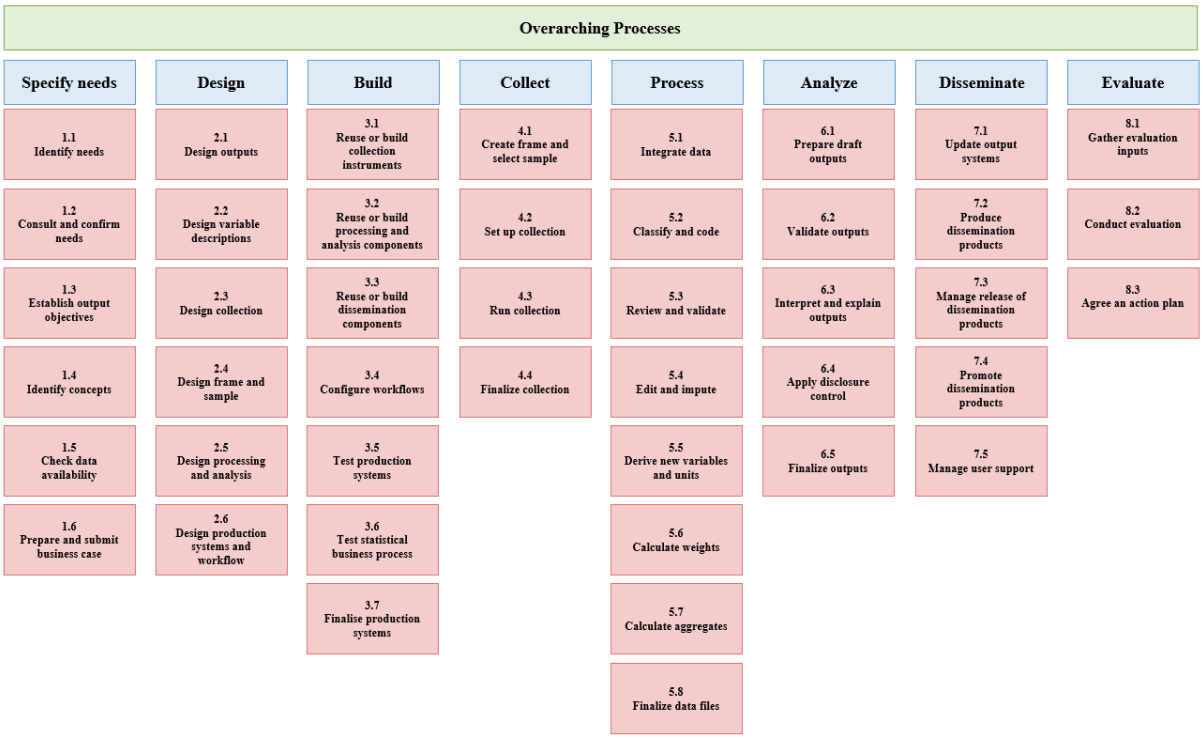
GSBPM[1], adopted by statistical offices of the most advanced countries in this field (Figure 1), is useful for describing all phases of statistical production and dissemination. This model is intended to guide the planning of surveys and other statistical operations through the systematic review of all processes and flow management from the earliest stages of preparation to dissemination, documentation, and archiving.

A statistical business process is a collection of related and structured activities and tasks to convert input data into statistical information. In the context of the GSBPM, SCAD performs statistical business processes to create official statistics to satisfy the needs of the users. The output of the process may be a mixed set of physical or digital products presenting statistics and metadata in different ways, such as statistical products, publications, maps, electronic services, among others.

This document extracts the sections relevant for the statistical process's phases of processing and analyzing.

[1] This section is based on the Statswiki of the UN Economic Commission for Europe (https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1) which is provided for the use of all statistical offices. As the GSBPM is a statistical standard, the definitions of phases and sub-phases are kept as in the original wiki document. The original document on the GSBPM is licensed under the Creative Commons Attribution 4.0 International License. It is attributed to the UN Economic Commission for Europe on behalf of the international statistical community (Downloaded 11th July 2023).

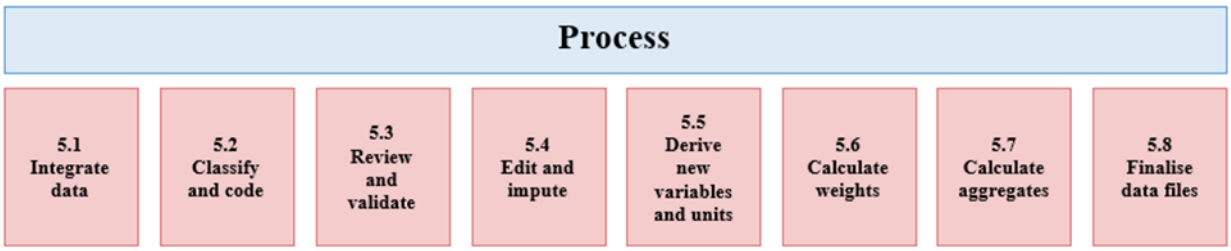
Figure 1: Generic Statistical Business Model



The GSBPM, version 5.1. The phases blue (level 1) and sub-processes pink (level 2). Note that the Process and Analyze phases can be iterative and parallel. Analysis can reveal a broader understanding of the data, which might make it apparent that additional processing is needed.

2.1. Process phase

Figure 2: The Process phase



This phase describes the processing of data and the preparation for analysis. It is made up of sub-processes that integrate, classify, check, clean, and transform input data, so that they can be analysed and disseminated as statistical outputs. For statistical outputs produced regularly, this phase occurs in each iteration.

Activities within the process and analyse phases may also commence before the collect phase is completed. This enables the compilation of provisional results where timeliness is an important concern for users and increases the time available for analysis.

The process phase is broken down into eight sub-processes (Figure 2, page 4), which may be or not, sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes that are relating to the activities of processing, validation and calculating weights of survey data are.

Phase 5.1. Integrate data

This sub-process integrates data from one or more sources. It is where the results of sub-processes in the collect phase are combined. The input data can be from a mixture of external or internal sources, and a variety of the collection instruments, including extracts of administrative and other non-statistical data sources. This sub-process also includes harmonizing or creating new figures that agree between sources of data. The result is a set of linked data. Data integration can include:

- Combining data from multiple sources, as part of the creation of integrated statistics such as national accounts.
- Combining geospatial data and statistical data or other non-statistical data.
- Data pooling, with the aim of increasing the effective number of observations of some phenomena.
- Matching or recording linkage routines, with the aim of linking micro or macro data from different sources.
- Data fusion - integration followed by reduction or replacement.
- Prioritizing, when two or more sources contain data for the same variable, with potentially different values.

SCAD has a data flow model which covers the whole sub-process of integration data, see Guide No. (19) for reference. Note that both survey and administrative data are processed in a similar way.

Phase 5.2. Classify and code

This sub-process classifies and codes the input data. For example, automatic (or clerical) coding routines may assign numeric codes to text responses according to a pre-determined statistical classification to facilitate data capture and processing. Some questions in the survey questionnaires have coded response categories on the questionnaires, others are coded after collection using an automated process (which may apply machine learning techniques e.g., for coding retail sales) or an interactive, manual process.

Coding is the process of assigning a numerical value to responses to facilitate data capture and processing in general.

When determining the coding scheme, the goal should be to classify responses into a meaningful set of exhaustive and mutually exclusive categories that bring out the essential pattern of responses. For some questions, coding may be straightforward (e.g., marital status). In other cases, such as geography, industry

and occupation, a standard coding system may exist. But for many questions no standard coding system exists and determining a good coding scheme is a nontrivial task. The coding scheme should be consistent and logical and consider how detailed the codes should be in light of the purpose of the survey and tabulations or data analysis to be performed. It is best to start with a broad list, since too few categories can be misleading, and a large other category can be uninformative. Categories can always be collapsed, but it might be difficult to split into the original categories after that.

Data capture is the transformation of responses into a machine-readable format. With computer-based collection methods, capture occurs at the time of collection.

Editing is the application of checks to identify missing, invalid, or inconsistent entries that point to data records that are potentially in error. Editing usually identifies non-sampling errors arising from measurement (response) errors, nonresponse, or processing. Editing can occur at several points throughout the process (from collection to dissemination) and ranges from simple preliminary checks performed by interviewers in the field to more complex automated verifications performed by a computer program after the data has been captured.

There are three main categories of editing: validity, consistency, and distribution edits. Validity and consistency edits are applied one questionnaire at a time.

- Validity edits verify the syntax of responses and include such things as checking for non-numeric characters reported in numeric fields and checking for missing values. Validity editing can also check that the coded data lies within an allowed range of values.
- Consistency edits verify that relationships between questions are respected. Consistency edits can be based on logical, legal, accounting, or structural relationships between questions or parts of a question.
- Distribution edits are performed by looking at data across questionnaires. These attempt to identify records that are outliers with respect to the distribution of the data.

Edits during data collection are often referred to as field edits and generally consist of validity edits and some simple consistency edits. The most comprehensive and complicated edits are generally carried out as a separate editing and imputation stage after data collection. During data capture, edits can be carried out by keys or automatically by computer programs, or by the computer application in the case of computer-assisted collection methods. Generally, editing during data capture is minimized since responding to an edit failure slows down data capture. Edits during this stage of processing are mainly validity edits and simple consistency edits. In this manual we will not describe the editing connected to the macrolevel (aggregated data or statistics).

Phase 5.3. Review and validate

Data validation is an activity verifying whether or not a combination of values is a member of a set of acceptable combinations. This sub-process examines data to identify potential problems, errors, and

discrepancies such as outliers, item non-response and miscoding. It can also be referred to as input data validation. It may be run iteratively, validating data against pre-defined editing rules, usually in a set order. It may flag data for automatic or manual inspection or editing. Reviewing and validating can apply to data from any type of source, before and after integration, as well as imputed data from sub-process 5.4 (Edit and impute). Whilst validation is treated as part of the process phase, in practice, some elements of validation may occur alongside collection activities, particularly for modes such as computer assisted collection. Whilst this sub-process is concerned with detection and localization of actual or potential errors, any correction activities that actually change the data is done in sub-process 5.4 (Edit and impute).

Phase 5.4. Edit and impute

Where data are considered incorrect, missing, unreliable or outdated, new values may be inserted, or outdated data may be removed in this sub-process. The terms editing and imputation cover a variety of methods to do this, often using a rule-based approach. Specific steps typically include:

- Determining whether to add or change data.
- Selecting the method to be used.
- Adding/changing data values.
- Writing the new data values back to the data set and flagging them as changed.
- Producing metadata on the editing and imputation process.

The auditing stages can be summarized as follows:

- Field Editing: The first phase of editing, where enumerators and controllers review the completed questionnaires before submission to the office.
- Office Editing: Completed questionnaires are reviewed by the editing team who reports back to field team on errors requiring correction in the field and any persistent errors by enumerators. After that, the questionnaires are coded based on economic activity and loaded in the database through the data entry system.
- Electronic Editing: Once the data is in the system, automated checks are run to report systematic field data errors. The following types of editing rules are applied to check for anomalies: structure edits, and consistency and validity checks. Automated edit checks are used for: total values validations, related variables validations, and logic validations.
- Other checks: Consist basically in the analysis of time series consistency of micro-data and the survey results.

Imputation Methods

Imputation is the process of replacing missing data, invalid values, and extreme values (outliers) with estimated values – using different imputation methods, then analyzing the full dataset considering the imputed value as actual values. Missing data, data errors, and extreme values (outliers), are common

challenges in statistical analysis, which can be found during data validation and data analysis processes. Note that the experienced subject matter expert is most often best suited to assess the outliers. When it comes to deciding on method to use for handling outliers, the subject matter expert and the methodologist need to cooperate.

The missing data is the case when the value of one or more observations (records) is not available in a variable of the dataset. Data errors are the case where the occurrence of invalid value in one or more observations (records) in a variable or group of variables in the dataset. Outliers is the case of extreme values that are significantly distinct from other observations (records) in the dataset.

Deciding on which imputation method to implement depends on various aspects such as the dataset size, variable type, rate of missing data, invalid data, outliers, patterns, variable distribution, time series, consistency, growth rate, the availability of historical data, data noise, and data classification.

In most surveys, two types of missing data are usually distinguished: object (unit) nonresponse and variable (item) nonresponse. Imputation applies for the latter, while object nonresponse is dealt with by reweighting. Object non-response refers to the failure to collect any information from some survey objects (e.g., people or companies). Variable (item) non-response refers to missing data in a returned questionnaire.

- Object (unit) non-response: The usual statistical practice for correcting object non-response is to change weights to compensate for non-responding objects.
- Variable (item) non-response: In the case of more technical or sensitive non-response may be decreased if it is specified that a person with subject-matter knowledge or higher responsibility should respond for the business.

Expert Judgment

In this technique the missing value (and invalid values) is imputed with the projected or estimated value provided by subject matter experts (SME). This method consists of replacing missing values by values that are specified by subject-matter experts, using a manual procedure, or a few rules of thumb. This method is recommended only in the case of very few missing values. Subject matter experts have projections or estimates for each missing value. This estimated/projected missing value will be used for imputing.

For example, the missing values in the labor wages or duties for specific tasks. Subject matter experts can get the information from market experience and suggest the data that must be imputed for missing variables.

Phase 5.5. Derive new variables and objects

This sub-process derives data for variables and objects that are not explicitly provided in the collection but are needed to deliver the required outputs. It derives new variables by applying arithmetic formulae to one or more of the variables that are already present in the dataset or applying different model assumptions.

This activity may need to be iterative, as some derived variables may themselves be based on other derived variables. It is therefore important to ensure that variables are derived in the correct order. New objects may be derived by aggregating or splitting data for collection objects, or by various other estimation methods. Examples include deriving households where the collection objects are persons or enterprises where the collection objects are legal units.

Phase 5.6. Calculate weights

This sub-process creates weights for object data records according to the sampling methodology developed for each survey. For example, weights can be used to "gross-up" data to make them representative of the target population (e.g., for sample surveys or extracts of scanner data), or to adjust for non-response in total enumerations. In other situations, variables may need weighting for normalization purposes. It may also include weight correction for benchmarking indicators (e.g., known population totals).

Phase 5.7. Calculate aggregates

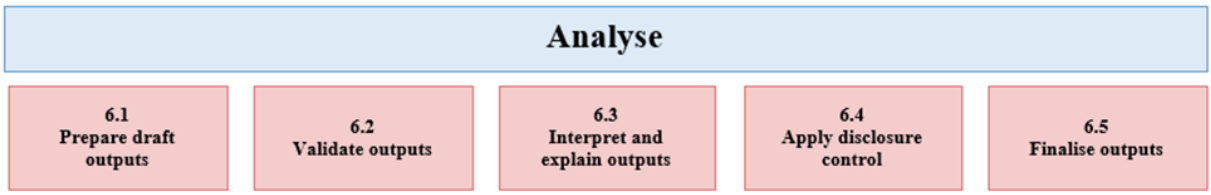
This sub-process creates aggregate data and population totals from microdata or lower-level aggregates. It includes summing data for records sharing certain characteristics (e.g., aggregation of data by demographic or geographic classifications), determining measures of average and dispersion, and applying weights from sub-process 5.6 (Calculate weights) to derive appropriate totals. In the case of statistical outputs which use sample surveys, sampling errors corresponding to relevant aggregates may also be calculated in this sub-process.

Phase 5.8. Finalize data files

This sub-process brings together the results of the other sub-processes in this phase in a data file (usually macro-data), which is used as the input to the analyze phase. Sometimes this may be an intermediate rather than a final file, particularly for business processes where there are strong time pressures, and a requirement to produce both preliminary and final estimates.

2.2. Analyze Phase

Figure 3: The Analyse phase



In this phase, statistical outputs are produced and examined in detail. It includes preparing statistical content (including commentary, technical notes, etc.), and ensuring outputs are “fit for purpose” prior to dissemination to users. This phase also includes the sub-processes and activities that enable statistical analysts to understand the data and the statistics produced. The outputs of this phase could also be used

as an input to other sub-processes (e.g., analysis of new sources as input to the design phase). For statistical outputs produced regularly, this phase occurs in every iteration. The analyze phase and sub-processes are generic for all statistical outputs, regardless of how the data were sourced.

The analysis phase is broken down into five sub-processes (Figure 3, page 9), which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. The only sub-processes that refers to activities of validation of survey data is sub-process 6.2.

Phase 6.1 Prepare draft outputs

This sub-process is where the data from sub-processes 5.7 (Calculate aggregates) and 5.8 (Finalise data files) are transformed into statistical outputs such as statistical values (estimates), indexes, seasonally adjusted statistics, e.g., trend, cycle, seasonal and irregular components, accessibility measures, etc., as well as the recording of quality characteristics (e.g., coefficients of variation). The preparation of maps, GIS outputs and geo-statistical services can be included to maximize the value and capacity to analyze the statistical information. This is also the phase where preparation and planning are checked for the final quality assessments to be done at an aggregated level, e.g., macro-editing and disclosure control. At this stage the draft of the final quality report has to be written (no later).

Phase 6.2. Validate outputs

Sub-process 6.2 consists in validating the quality of the outputs produced, in accordance with a general quality framework and with expectations. Validation activities can include:

- Checking that the population coverage and response rates are as required.
- Comparing the statistics with previous cycles (if applicable).
- Checking that the associated metadata, paradata and quality indicators are present and in line with expectations.
- Checking geospatial consistency of the data.
- Confronting the statistics against other relevant data (both internal and external).
- Investigating inconsistencies in the statistics.
- Performing macro editing.
- Validating the statistics against expectations and domain intelligence.

Phase 6.3. Interpret and explain outputs

This sub-process is where the in-depth understanding of the outputs is gained by statisticians. They use that understanding to interpret and explain the statistics by assessing how well the statistics reflect their initial expectations, viewing the statistics from all perspectives using different tools and media, and carrying out in-depth statistical analyzes such as time-series analysis, consistency and comparability analysis, revision analysis (analysis of the differences between preliminary and revised estimates), analysis of asymmetries (discrepancies in mirror statistics), etc.

Phase 6.4. Apply disclosure control

This sub-process ensures that the data (and metadata) to be disseminated do not breach the appropriate rules on confidentiality according to either organization policies and rules, or to the process-specific methodology created in sub-process 2.5 (Design processing and analysis). This may include checks for primary and secondary disclosure, as well as the application of data suppression or perturbation techniques and output checking. The degree and method of statistical disclosure control may vary for different types of outputs. For example, the approach used for microdata sets for research purposes will be different to that for published tables, finalized outputs of geospatial statistics or visualizations on maps.

Phase 6.5. Finalise outputs

This sub-process ensures the statistics and associated information are fit for purpose and reach the required quality level specified in the design document and are thus ready for use. It includes:

- Completing consistency checks.
- Determining the level of release and applying caveats.
- Collating supporting information, including interpretation, commentary, technical notes, briefings, measures of uncertainty and any other necessary metadata.
- Producing supporting internal documents.
- Conducting pre-release discussion with appropriate internal subject matter experts.
- Translating the statistical outputs in countries with multilingual dissemination.
- Approving the statistical content for release.
- Produce and finalize the quality report. This is done by methodology and subject matter in collaboration (subject matter is responsible).

3. Specificities for each type of survey

In this section, the methodological aspects of data processing, validation and imputation of survey data are presented for the two types of surveys: Household and Business surveys.

3.1. Household surveys

The processing, validation and imputation methods applied for household surveys mentioned above are described below.

For household surveys, it is usually the case that the survey questionnaire contains one section for household-level data, and one section for each individual in the household (replicated as many times as there are individuals in the household). This implies that data processing is carried out using the aggregation of two types of statistical objects: households and individuals.

Data validation methods

One important validation method is by analyzing the coherence of survey results with other data sources, in particular, for household surveys, the following may be used to check the coherence of the results:

- Population and Housing Data
    - Total number of households in Abu Dhabi in a calendar year
- Administrative data sources (e.g., population register)

Imputation methods

Please refer to section 5.4 above for a description of processes to impute missing values.

Calculation of weights

Individual records of the sample have to be weighted to represent the total population. The calculation of sampling weights derives from the sample design (stratification, multi-stage sampling, clustering, etc.) and therefore cannot be decided at this stage. It is however recommended that the sample design follows that of other household surveys (e.g., Labor Force Survey).

Note that if the final sampling object (unit) is the household (i.e., all members of the household are interviewed, which is the international recommendations), sample weights are calculated at the household level. If there is a further sampling (i.e., selecting one informant per household), sample weights have to be calculated at the individual level.

Note that more complex weights have to be calculated if the sampling strategy includes stratification, more than one stage and/or is clustered as recommended in the methodological documents prepared for each survey. Indicators are generally calculated as ratios of grossed-up figures after sampling.

3.2. Business surveys

This section describes specificities of the processing, validation and quality control methods applied for the business surveys mentioned above.

In the case of business surveys, it is often the case that the survey questionnaire includes a section to collect data at the level of enterprise, and a section to collect data at the level of each establishment or Local Unit, even at the level of Local Kind-of-Activity Unit. This translates into a relational database that contains the data at different levels, linked through unique identifiers of businesses and establishments.

Data validation methods

Data editing covers the sub-processes referred to as micro-editing and macro-editing (also sometimes referred to as input and output editing):

- Micro-editing refers to controls, validations and modifications applied to the data of a given business. The process includes the treatment of incomplete or missing data and the detection and treatment of answers that are internally inconsistent with other questions.
- Macro-editing refers to controls, validations, and modifications of whole datasets by means of the analysis of aggregations. The aim of the process is to check whether certain estimates are jointly compatible and are consistent with another knowledge. A sophisticated macro-editing procedure consists of readjusting sample weights according to frame errors detected during the survey. Here the subject matter experts need to use their experience on how to benchmark against other relevant statistics.

For many reasons, statistical information provided by businesses, whatever the instrument of data capture, can contain errors. These include erroneous or missing data, incorrect classifications, and inconsistent or illogical responses. To minimize such errors, it is important to apply techniques which optimize the effectiveness of data capture instruments and collection procedures. In addition, robust data editing techniques should be used to transform raw data provided by respondents into valid and coherent (clean) data that can be used to produce aggregated statistics.

Treatment of internal inconsistencies and errors

Data editing involves checking and often manipulation of the original data. Such processes can introduce errors that affect aggregate data. Thus, although the process of data editing is essential, it is very important that practices be established that decrease the incidence of incomplete or inconsistent data, so that the impact of data editing is minimized. Quality controls already embedded in data collection instruments or at the data entry stage will directly improve the quality of raw data and reduce the task of data processing.

The choice of collection instrument has a direct impact on data quality. Both computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI) can be expected to improve the quality of input data since they provide automatic controls for detecting response errors.

Validity control of an individual data item consists of checking if the answer belongs to a predefined set (or range) of valid responses. In order to check questions for validity, it is necessary to check them against those defined valid responses. To check the internal consistency of a questionnaire, it is necessary to establish and apply rules that define the relationships between questions, so that certain answers restrict the valid values that other questions can accept. Arithmetic checks (for instance, that percentage distributions add to one hundred) may be applied during data entry or later run-in batch mode across a set of records.

Treatment of misclassified units

A frequent problem affecting the quality of business statistics is that some responding businesses may be initially included in the wrong stratum in the population frame from which the sample is drawn. This is more

likely when the frame (and the underlying business register) is of poor quality. Statistical business registers maintained by NSOs usually contain information on size (usually in terms of number of employees and/or turnover), industry and location (based on business address). It is possible that misclassified objects are erroneously included as eligible, and that eligible units are misclassified such that they do not appear on the frame or appear in the wrong stratum. In the first case, if a surveyed business is eliminated from the sample because of non-eligibility, this will reduce the effective sample size unless a reserve list is prepared.

Elimination of misclassified objects should only be considered if the rate of misclassification is small. In the second case, the object is eligible, but was included in the wrong stratum or omitted from the frame altogether. For example, a business selected in the size interval (stratum) of 10 to 20 employees may report that, in fact, it has only eight employees. The technical solution consists of recalculating sample weights. A new estimate of the size of strata must be produced and weights corrected accordingly. Clearly, the establishment and maintenance of an up-to-date business register from which to draw a reliable population frame is of utmost importance.

Where it is impractical to re-contact respondents, missing data could be estimated (imputed). If re-contacting the interviewed business is out of scope for cost or time reasons, item nonresponse can be treated by mathematical techniques for data imputation. Imputation consists of assigning a plausible value to a question for which the selected unit has not provided a response, or to a question whose answer is logically or arithmetically inconsistent with answers in the rest of the questionnaire. When the answer to different questions is inconsistent, the problem of deciding which one is incorrect may be very difficult. Usually, a hierarchy among questions, or blocks of questions, is established, so that answers to some of them are considered ‘dominant’.

One method of imputation used in recurrent business surveys is to assign the answer given by the same object in a previous survey (historical imputation). The same principle can be applied to object non-response. This technique would more frequently be applied to large businesses, because they are more likely to be in successive surveys. Please refer to section 4 below for a description of processes to impute missing values.

Imputation methods

Please refer to section 5.4 above for a description of processes to impute missing values.

Calculation of weights

It is well known that the sampling object (unit) weight is the inverse of the probability of selecting the concerned object from the sample. Weights can be used to make survey sample representative of the target population for normalization purposes. For stratified sampling design, the sampling weights for the sample establishments within each stratum will be calculated independently. The sample size of the economic establishments within a certain stratum is equal to the product of dividing the number of economic establishments in the frame by the number of responded economic establishments to the survey data.

As for the large establishments' stratum, mainly the first partial stratum, the weight of the economic establishment shall be equal to 1, meaning that it represents itself only, given that it has been selected rather than probably chosen. As for the rest of the establishments within the second partial stratum, the weight of each shall be calculated using the same previous technique.

4. References

UNECE Statistics Wikis - Generic Statistical Data Editing Models, https://statswiki.unece.org/display/sde/GSDEMs. (Downloaded 6th July 2023)

مـركـز الإحـصـاء
STATISTICS CENTRE

الرؤية: ببياناتنا نمضي نحو غدٍ أفضل
Vision: Driven by data for a better tomorrow

www.scad.gov.ae
adstatistics