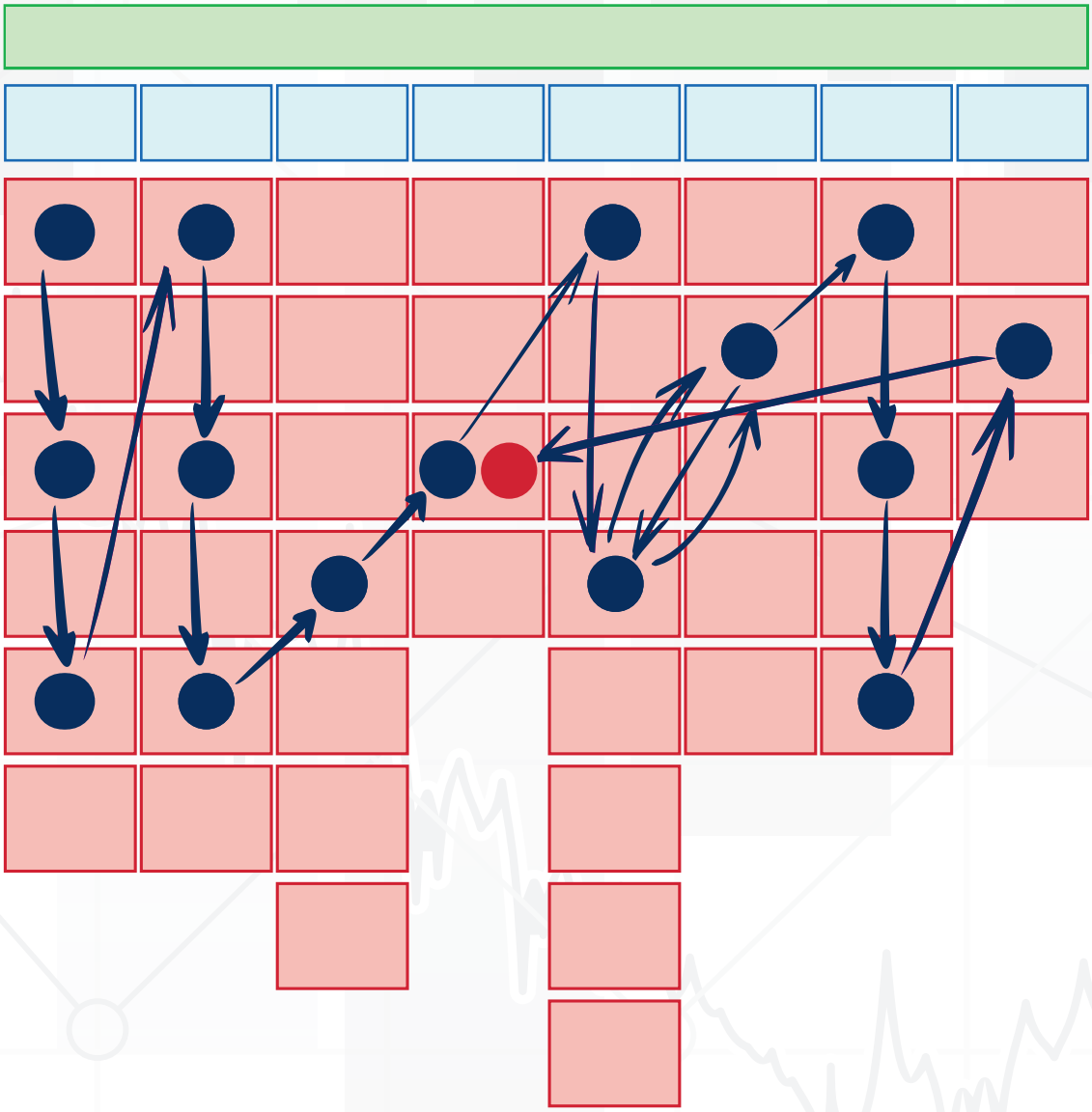
INSTRUCTIONS ON QUALITY ACCORDING TO THE GENERIC STATISTICAL BUSINESS PROCESS MODEL (GSBPM)

June 2022



DRŽAVNI ZAVOD ZA STATISTIKU
CROATIAN BUREAU OF STATISTICS

Printed in electronic format

Contents

FOREWORD

The Croatian Bureau of Statistics is the main producer, disseminator and coordinator of the official statistics system of the Republic of Croatia and the main representative of the national statistical system in front of European and international bodies competent for statistics.

Statistical data collected, processed, published and disseminated by the Croatian Bureau of Statistics are the result of complex procedures and processes of conducting statistical surveys prescribed by the annual implementation plans of statistical activities of the Republic of Croatia.

Effective and timely preparation and dissemination of quality statistical indicators that reflect economic and social phenomena and processes as well as provide users with a reliable source for analysing the current situation and decision-making are among the main tasks of the Croatian statistical system.

Taking into account all the determinants of quality in statistics, the Croatian Bureau of Statistics decided to apply the customised GSBPM model because it comprehensively describes and defines the set of business processes required for the production of official statistics. This model provides a standard framework and unique terminology that allows the modernisation of statistical production processes and the exchange of methods and components. It is also used for data integration and metadata standardisation, as a template for process documentation, harmonisation of statistical computing infrastructures, providing a framework for assessing quality processes, and for further improvements.

On the grounds of the GSBPM model, the Croatian Bureau of Statistics conducted an analysis of the possibility of its application in the Croatian statistical system and it resulted in the creation of an adjusted model as well as in this publication, which comprises, in one place, all business processes and sub-processes, including practical examples and pointing out accomplishment of higher quality of statistical products and services.

This publication outlines the statistical activities of all processes and sub-processes from defining data needs to their evaluation. Readers will be able to gain insight into comprehensive statistical operations and plans for further improvements using modern technologies in one place.

ABBREVIATIONS AND ACRONYMS

APIS-IT	agency providing strategic, professional and implementation services to the public and government sector organisations of the Republic of Croatia
GDP	gross domestic product
CAPI	computer-assisted personal interviewing
CATI	computer-assisted telephone interviewing
CAWI	computer-assisted web interviewing
CD	compact disc
Eurostat	statistical office of the European Union
DVD	digital versatile disc
DV-PO	Annual Report on Kindergartens and Other Legal Entities Implementing Preschool Education Programmes
CBS	Croatian Bureau of Statistics
ESA	European System of Accounts
eSPRi	Statistical Business Register
ESS	European Statistical System
ESSC	European Statistical System Committee
EU	European Union
Fina	Financial Agency
GeoSTAT	portal of the Croatian Bureau of Statistics with cartographic projections
GSBPM	Generic Statistical Business Process Model
PO-31a/Q	Quarterly Report on Sale of Agricultural Products from Own Production – for Legal Entities and Tradesmen
PO-31b/Q	Quarterly Report on Sale of Agricultural Products from Own Production – for Legal Entities and Tradesmen
PO-32	Report on Yields of Early Crops and Fruits
PP/T-11	Quarterly Report of Maritime and Coastal Transport
PR/T-11P	Quarterly Report on the Transhipment of Goods
RAD-1G	Statistical Report on Persons in Employment and Paid-off Earnings
RB-1	Statistical Report on Divorce
SIT	Information Technologies Directorate
SOP	vocational training in enterprises
IACS	Integrated Administration and Control System
IDU-OK	Investments in Environmental Protection and Expenditures on Goods and Services in Environment

IKT-DOM	Annual Survey on Usage of Information and Communication Technologies in Households and by Individuals
IKT-POD	Annual Survey on Usage of Information and Communication Technologies in Enterprises
INOV	Innovation Activities in Enterprises
INSPIRE	Infrastructure for Spatial Information in Europe
INV-P	Annual Report on Gross Investment in Fixed Assets of Legal Entities
KLASUS	a tool designed for use by all users of classifications, which enables browsing and searching of classifications by name and code
IMF	International Monetary Fund
ID	identification number
IDS	subject entry identification number
IMF	International Monetary Fund
NKD	National Classification of Activities
OG	Official Gazette
NSK	National and University Library
OECD	Organisation for Economic Cooperation and Development
OIB	personal identification number
PA/M-11	Quarterly Report on Road Transport
PA/T-11	Statistical Survey on Road Transport of Goods
PC AXIS	free-of-charge softwer developed by the Statistics Sweden
VAT	value added tax
PO-22/STR	Report on Structure of Agricultural Holdings
SGA	State Geodetic Administration
SPR	Spatial Statistical Register
SPC	Statistical Programme Committee
SRPG	Statistical Register of Farms
SBS	short-term business statistics
Š-O/KP	Statistical Sheet on Basic Schools
Š-S/KP	Statistical Report on Upper Secondary Schools
TU-11	Monthly Report on Tourist Arrivals and Nights
TU-14	Raport on Travel Agencies
UNECE	United Nations Economic Commission for Europe

GENERAL PRINCIPLES

STATISTICAL SURVEY

In implementation of statistical business processes, the statistical survey holds the central place. A statistical survey is any comprehensive set of activities designed for data collection purposes. When talking about direct data collection on the selected random sample of observation units, then it is a classical concept, as opposed to a broader concept that includes various additional business processes, in particular during the data collection stage. Statistical surveys can therefore be divided according to:

TYPES OF COVERAGE:

- census – the whole population is observed
- sample survey – the selected population sample is observed
- derived statistics – previously known statistical results are used
- existing statistical aggregates (e.g. statistics on national accounts)

METHODS OF DATA COLLECTION:

- interview (by phone or directly)
- self-completion (analogue or electronic questionnaire)
- other means of collection (observation)

PERIODICALS:

- monthly
- quarterly
- annual
- different periodicals (weekly, perennial)

DATA SOURCES:

- statistical (primary) sources
- administrative and other secondary data sources.

BASIC PROCESSING MODEL

The basic structure of the document was prepared according to the adapted generic model of statistical business processes (GSBPM), which describes in detail and defines the set of business processes required for the production of official statistics. This model provides a standard framework and unique terminology that allows the modernisation of statistical production processes and the exchange of methods and components. The same model is used for data integration and metadata standardisation, harmonisation of statistical computing infrastructures, providing a framework for assessing the quality of business processes, and for further improvements.

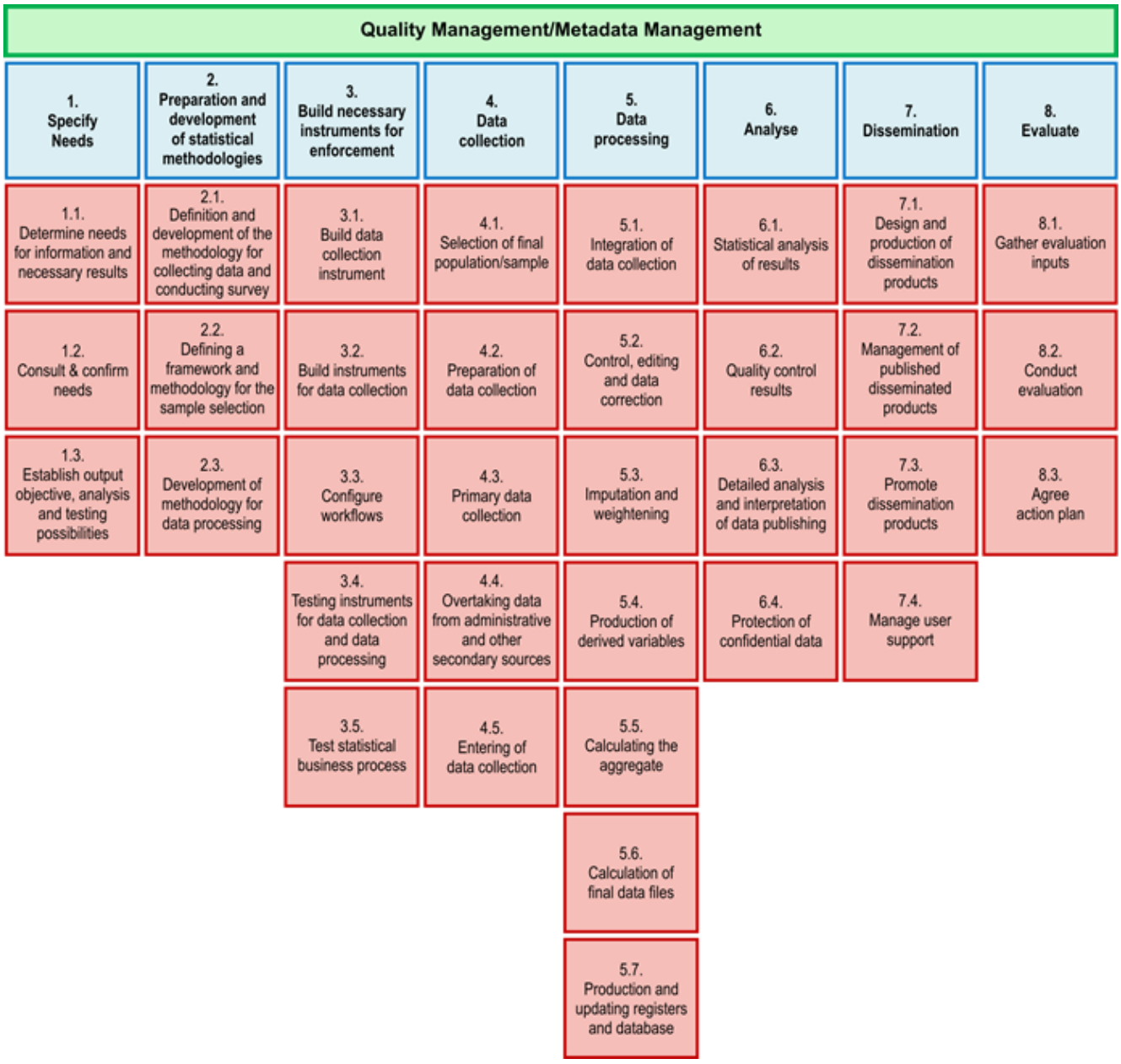
On the grounds of the GSBPM model, the Croatian Bureau of Statistics conducted an analysis of the possibility of its application in the Croatian statistical system and, based on the conducted analyses, developed an appropriate national model that follows the recommendations of the original UNECE/Eurostat/OECD[1] model of the Work Session on Statistical Metadata (METIS) and is based on a business process model used by the New Zealand statistics.

[1] Link: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

As a model, the GSBPM is designed independently of the data source and is adapted to the needs of the Croatian statistical system. It can be used to describe and assess the quality of the process based on surveys, censuses, administrative records and other non-statistical or combined data sources.

This model includes eight processes that are divided into different sub-processes. The titles and subheadings in the document are taken from the general model. Each chapter contains a brief description of the whole process, all subsections consist of two parts: the first part contains a general description, while the second part contains quality assurance guidelines that should serve as a kind of list of quality elements for statisticians. Although the model in actual application follows the sequence of processes and sub-processes, the phases are often interconnected and intertwined. Therefore, they could not be represented by a simple linear model. In general, national statistics should have an established model that should be followed as much as possible in practice.

Image 1: Generic Statistical Business Process Model – GSBPM



ASSESSMENT OF STATISTICAL PROCESSES AND PRODUCTS

In the last few years, models and structures for monitoring and measuring the quality of statistical processes and products have been continuously improved in the production of official statistical data.

The importance of quality is shown by the joint initiative of the national statistical institutes to adopt and harmonise different practices in publishing data on the estimated level of quality. In previous years, statistical quality was focused on the accuracy of statistical results, while today the quality is multidimensional and refers to the adequacy of statistical results in terms of applicability for different purposes and for different user groups. Thus, precision has become just one of several quality components that are measurable and exactly defined.

International organisations such as Eurostat, the IMF and the OECD have played a key role in development and a clear vision in methodological and practical terms in the introduction of modern quality measurement concepts. In this sense, the ESS has in recent years developed a standard and unique model for assessing quality in statistics, based on six dimensions of quality:

Relevance is the dimension of quality within which the following questions are sought to be answered:

- Will the results of the set statistical survey meet the expectations of users and provide complete answers to user requirements?
- Can the scope of the research meet user needs?
- Within this quality dimension, it is important to regularly monitor customer needs using tools such as customer satisfaction.

Accuracy is a dimension that we have estimated to (not) match the recorded condition and does not correspond to the actual condition because the population values are unknown. Inconsistency is also called statistical error. The different types of errors that occur when conducting a survey need to be assessed and these should be clearly explained and made available to users.

Timeliness and punctuality − timeliness refers to the interval between the reference period to which the data relate and the date of publication, while punctuality is the interval of measures between the actual and announced date of publication determined in the release calendar of statistical data according to the legal basis of the statistical survey.
Coherence in statistics refers to the ability to reliably combine data in different ways and for different users. This quality dimension has the following sub-dimensions:

- coherence between provisional and final results
- coherence of annual and structural statistics
- coherence within the same statistical domain
- coherence with national accounts results.

Comparability means the extent to which the data presented are comparable in time and space. The following are particularly important sub-dimensions:

- geographical comparability (possibilities of comparison between countries, regions, etc.)
- temporal comparability (possibility of comparing data in different periods)
- comparability between domains (e.g. comparison of data from two industry branches, two different curricula, different age groups, etc.).

Availability and clarity of data − availability refers to the physical conditions in which data are available to the user (possibility of ordering, schedule of job positions, pricing policy, etc.). Clarity describes the data environment in which the user can access information. It should be described in particular whether appropriate metadata are available and whether the user has been granted access to the quality of the published results.

Based on the described dimensions of quality, the national statistical institutes, organised in the ESS, developed standardised documents for the needs of users to monitor the quality of published statistical data. These documents are called standard quality reports (link Državni zavod za statistiku - Quality reports by statistical domains (gov.hr).

The European Statistics Code of Practice is an important step in ensuring the quality of the statistical results of the European Statistical System (hereinafter: the Code). The Code was first adopted by the SPC in 2005 and revised by the ESSC in 2011 and 2017. The last revision was confirmed by the ESSC on 16 November 2017. The revised Code has 16 principles concerning the institutional environment, statistical processes and statistical results. The aim of the Code is to ensure that statistics produced within the ESS are relevant, timely and accurate and in line with the principles of professional independence, impartiality and objectivity. A set of best practice indicators and standards for each principle provides guidance and references for evaluating the application of the Code.

GENERAL INSTRUCTIONS FOR QUALITY ASSURANCE

The chapters present the Quality Assurance Guidelines in each of the sub-processes so that they are adapted to the overall business process model. General instructions and procedures for the development of a new survey as well as procedures for refining existing surveys are presented.

For the effective establishment of an individual business process, it is necessary to plan in a timely manner on an annual and multi-annual basis. First of all, it is necessary to assess the actual possibilities of implementing business processes with regard to the availability of human, financial and information potentials and to prepare a realistic schedule of activities in accordance with the preconditions.

Procedures within business processes should allow for repeatability and consistency as much as possible. The repeatability of statistical business processes means that final results in each version of business processes where the input data are unchanged should always be the same. Consistency means that any change in the data in this business process is clearly and unambiguously recorded. This allows recording of all changes to the data in each execution process and monitoring of the impact of these changes on the final results.

It is definitely recommended to use good examples from practice. Employees should be regularly informed about the formation and implementation of statistical business processes and the existence of standards. Taking over good practical examples from other domains can be encouraged by improving the flow of information between different statistical domains. This can be most easily achieved by organising workshops, seminars and other forms of internal training.

1. Defining needs for data and the required results

When analysing requirements and needs for data, it is necessary to focus on external users of statistical data, i.e. on data collected according to statistical principles, that is, on data that are necessary for decision-making and implementation of public policies.

Needs and requirements analysis provide the basis on which future activities will be built and decisions on guidelines for survey design will be made. Users' expectations and needs for statistical results are growing and narrowly specialised due to changes in society, global processes and new initiatives at the international and national levels.

Timely identification of needs for data and preparation of appropriate data are key to systematic and strategic decision-making. Needs for data are defined directly with users through various communication channels, taking into account the capabilities and frameworks of the national statistical system.

When studying data, already existing statistical as well as administrative data sources should be checked, as this can significantly affect the rational use of data. It is important for everyone to be able to choose the right source of information that will contain the definition of the observation unit, possibility to determine identifiers, relevant content, quality resources and other characteristics.

When deciding between direct data collection in the field and the possibility to use administrative data sources, it is necessary to take into account that using administrative data sources is cheaper. The decision to use an administrative data source, the quality of resources and their relevance as well as methodological applicability and sustainability play an important role.

Before planning the implementation of a statistical survey, it is necessary to check the methodology, i.e. all phases of the survey. The verification of the methodology may refer to the existing survey or to the methodology for a new statistical survey.

1.1. Defining needs for information and the required results

Description

Identifying data needs should be initiated when data do not exist or when existing data do not meet all the needs of data users.

The need for data is determined by various users: ministries, the CNB, national and international institutions, researchers, scientists and the general public. In the process of data needs, it is possible to determine exactly what a user can expect from the national statistical system. To this end, interested users in various forms of cooperation, from statistical advisory committees, statistical councils to requests to national and international institutions, should be involved in this consultation process. The Croatian Bureau of Statistics is familiar with user expectations, especially with the types of information the user needs, when he needs them and for what purpose. Therefore, it is necessary to establish continuous cooperation with users, because if the previously identified needs of users change, it is necessary to change the procedures as well.

After that, the business processes necessary for the delivery of data sets required to meet the needs of users are determined in accordance with the options provided by the Croatian Bureau of Statistics. At this stage, it is necessary to reach a mutual agreement regarding the information on the quality of statistical data based on its six dimensions in order to avoid misunderstandings in the delivery of the final data set.

Instructions for quality insurance

Determining data needs requires a comprehensive and systematic approach that includes all interested users. In order for the Croatian Bureau of Statistics to learn the user needs, it is necessary to regularly exchange information with users and get to know the needs of users, especially to find out what data the user needs, when, for what purpose and in what way. In the analytical approach, one should also take into account all related user needs that can be easily met with a little extra effort and specific data needs that require the analysis of costs and benefits that data collection can bring.

1.2. Checking available data sources

Description

After the analysis of new user requirements, it is necessary to check all statistical surveys that use administrative data sources, and then check the availability of data in some other sources. If other sources are available, it is necessary to check the conditions, then methodological differences need to be determined in relation to statistical methodology as well as periodicals related to data collection, and other information. Checking the available administrative data sources determines the extent to which administrative data can be used as a direct data source. If the data from the user request are not available at all, it is necessary to organise the collection of these data sets and include them either in one of the well-established statistical surveys or prepare a new statistical survey.

Instructions for quality insurance

The quality of administrative data sources should always be checked, and in the case they are of appropriate quality, the use of administrative data sources in any case has an advantage over the direct collection of data from respondents.

Regarding the qualitative aspect, it is necessary to take into account the quality of administrative data sources, frequency of updates, relevance of data sources, methodological comparability of data source content and time availability of data. The level of data stability in administrative data sources should be regularly monitored and assessed. The quality of administrative data should be checked regularly by comparing individual consecutive data sets or by a sampling method on a survey specifically designed for this purpose.

Special attention should be paid to data users, especially in cases where data are completely taken from administrative sources for statistical purposes.

1.3. Establishing a concept for production of results and analysis and testing of possibilities

Consideration should be given to whether the existing survey can be supplemented to obtain the desired data, whether an administrative data source can be used or whether a new survey should be planned. Furthermore, an analysis needs to be prepared in order to determine the hierarchy of data sources.

In addition, it is necessary to prepare a comparison of administrative and statistical data for individual available variables in terms of content. Compliance with the methodology as well as the management of administrative data sources for the purpose of presenting statistical data should be checked.

In accordance with legal provisions, it is necessary to ensure compliance with statistical confidentiality as well as physical and information protection of data taken from administrative sources by limiting access to data and organising data access monitoring.

Instructions for quality insurance

It is necessary to conclude an appropriate agreement on the exchange of data with the owner of administrative data sources and, depending on the type of data, it is necessary to define a catalogue of data. A good practical example is the Catalogue of Mutual Demands for Statistical Data from the records of the Tax Administration and the Croatian Bureau of Statistics. The Catalogue, for example, contains the following information: name of tax/form, data set, purpose, periodicity of delivery, deadlines for submission of first data results, deadlines for submission of final data results, form and method of data delivery, admission department at the Croatian Bureau of Statistics, user department in the Croatian Bureau of Statistics, person responsible for using the file in the Croatian Bureau of Statistics, responsible person in the Tax Administration, person in charge of admission in the Tax Administration, responsible person in APIS-IT and person responsible for data formation – APIS-IT. In this case, the Croatian Bureau of Statistics is obliged to regularly inform the holders of administrative data sources about the level of data quality. The interest of statistics must be taken into account when identifying sources and proposing changes.

Data download and quality should be monitored throughout the year. If there are illogicalities, they need to be reported to the data owner.

2. Preparation and development of statistical methodologies

The analysis of user needs and other observations identified the needs for statistical surveys that are periodically planned. When preparing statistical surveys within the ESS, care is taken to ensure that they are carried out on the basis of European regulations or agreements and that they comply with prescribed standards. The purpose, content of the survey, target population and other elements must be in accordance with the needs and stated in the annual implementation plan. A detailed plan for each phase of the implementation of the statistical survey must be set out in the work plan. Concerning interviews based on sample that require additional funding (e.g., for interviewing people, households and farms), additional equipment and additional staff, all necessary items (human resources, finance, IT equipment) need to be planned several years in advance. For interviews, in addition to the work plan, it is necessary to develop a detailed financial plan.

The initial stages of planning statistical surveys and interviews are usually as follows: defining the content of variables and survey results; defining the target population and the observation population; determining data sources and data collection methods.

Further steps are determined in the text below, depending on the above definitions. If information is taken directly from reporting units, it is necessary to determine the method of selecting units (probabilistic sampling method), prepare a draft questionnaire and a list of supporting documentation. When part of the data or information for a particular statistical survey is taken from existing sources, it should be checked whether the data are already available in the Croatian Bureau of Statistics. If data on the required variables are not available in the Croatian Bureau of Statistics, it is necessary to initiate a procedure for concluding an agreement and, depending on the situation, to prepare a catalogue for data exchange and other necessary documents in agreement with the institution from which we want to obtain data.

In the early phase of planning and preparation of statistical survey, it is necessary to define in detail the methodology, all the way from the selection of observation units to statistical data processing. For the quality implementation of the survey, it is important that implementation of an experimental statistical survey is already plannded in the design of the survey in order to test all phases, or at least those phases in the implementation of statistical survey processing that have not been previously tested.

When planning and preparing the statistical survey, the Croatian Bureau of Statistics ensures that the burden on respondents is as low as possible as well as that the retrieval of data from reporting units is agreed to mutual satisfaction and with the lowest possible costs of conducting the survey, and with appropriate quality that will provide statistical indicators in the domain observed by the survey.

2.1. Defining and developing a methodology for data collection and conducting the survey

Description

Defining and developing a methodology for data collection and conducting a survey includes developing of all necessary methodologies (methods, collection instruments, variables, definitions, descriptions, instructions, agreements and contracts with data providers, content of the questionnaire, dissemination plan, etc.).

Preparing metadata descriptions of collected and derived variables and classifications is a key prerequisite for the following phases. The most appropriate methods and instruments for data collection are identified. Activities depend on data collection methods (CAPI, PAPI, CATI and CAWI), including testing of instruments. All formal data provision agreements are being drafted, such as memoranda of understanding and confirmation of the legal basis for data collection.

Verification of statistical survey methodology includes verification of theoretical part of methodology (e.g., target population, definitions, sampling methodology, use of existing data source and data processing) and verification of practical part of a statistical survey methodology (e.g., questionnaire format, data collection method, data processing procedure, types of publications and the like).

The methodology verification procedure refers to the adequacy of the methodology in relation to the identified needs. The course of this procedure also depends on whether the methodology is prepared on the basis of EU regulations or not. If it is prepared on the basis of EU regulations, it must be taken into account whether this regulation has already been adopted or it is in the process of being adopted or amended. In cases where it is necessary to apply the provisions of regulations, then the possibility of adopting the requirements of the regulations is examined, while, in the other case, it is determined whether the survey can be conducted as proposed in the regulation or in the amendment to the regulation that is in the process of being created and adopted. In the case when a statement is being prepared for the Republic of Croatia, the positions must be adopted at different levels, beginning with working, directorial, expert working groups in the ESSC and in the Council of the EU.

The implementation and application of the methodology is checked by the survey holder, i.e. the person responsible for the survey and statisticians who are in charge of a particular domain and participate in the process. Regarding the feasibility and suitability of the methodology, the survey holder may propose the convening of an expert working group to discuss the proposed solution and give his opinion on the methodological proposals. Before making a final decision on the methodology, the proposal should be considered by an expert panel.

Instructions for quality insurance

When testing the methodology, preference is given to existing methods and good practices, while in the preparation phase care must be taken to ensure that definitions are written in an understandable way, to clearly define what is covered and what is not.

When collecting data using a questionnaire, it is important that the questions are understandable and the instructions for filling them in are precise. However, when retrieving data from existing administrative sources, it is important to conclude an appropriate agreement on the manner of retrieving data with each designated body that holds administrative data sources required for conducting statistical surveys that are determined by the annual implementation plan of statistical activities of the Republic of Croatia and, at the same time, to define the catalogue of data in terms of types of demanded data, manner and deadlines for data submission and persons responsible for the preparation and retrieval of data.

When checking the implementation of individual phases of research, it is always necessary to take into account the set possibilities for implementation or look for other solutions.

In cases where the statistical survey is conducted at the EU level, the Croatian Bureau of Statistics should be actively involved from the very beginning, i.e. from the moment the topic is discussed in a working group operating at the EU level for each individual statistical survey. Therefore, the Croatian Bureau of Statistics can influence the methodology (target population, definitions, classifications, periodicity of publication, etc.). In addition, the part of the methodology that is not prescribed is checked (e.g. method of data collection, statistical processing methodology).

For successful implementation of a statistical survey, a plan of necessary financial, information and human resources has to be prepared. A detailed schedule of activities is the most important for the effective implementation of a statistical survey.

Each statistical survey must be listed in the annual implementation plan of statistical activities of the Republic of Croatia. Depending on the manner of implementation in the annual implementation plan of statistical activities of the Republic of Croatia, statistical surveys are structured as follows:

I statistical survey based on direct data collection (quarterly, semi-annually, annually, multiannually)

II statistical survey based on administrative sources or the observation and monitoring method

III developmental activities, censuses and other more extensive statistical surveys.

In cooperation with organisational units, the holder of the suvey defines in the work plan the list of activities, the statisticians in charge of their implementation, the sequence of activities and deadlines.

Before starting the statistical survey, it is necessary to prepare a plan of data sources and a financial plan. Activities need to be included in the budget in due time and needs for additional resources planned (e.g. laptops). This is a common procedure in the implementation of individual statistical surveys, then in larger revisions of data, multiannual statistical surveys, in statistical surveys that are continually required through the so-called ad hoc modules and in surveys that require the participation of a survey laboratory.

When planning data sources and activities, it is important to organise cooperation of statistical survey holders with the organisational unit for statistical survey support and to achieve an optimal ratio and distribution of tasks with regards to the data source and available working hours.

The implementation of activities should be closely monitored throughout the year. Identified illogicalities should be reported to the responsible persons in accordance with the current organisation chart and responsibilities for individual sub-processes of the GSBPM model. All identified irregularities and unfavourable circumstances in the conduct of the statistical survey must be taken into consideration when preparing the work plan for the next budget periods. When planning financial resources, it is necessary to determine the amounts of individual budget items for external interviewers who are included in the statistical survey on the basis of concluded employment contracts. The amount of the budget item for the calculation of financial items is obtained on the basis of the sample size and the estimated response and non-response rates for each individual statistical survey. In the structure of interviewers, according to the method of payment for this purpose, it is necessary to use the overview of cost simulations for each individual segment of the survey. In certain surveys, it is possible to plan certain additional financial items (e.g. compensation to households for a properly completed consumption diary, etc.).

2.2. Defining the framework and methodology for sample selection

Description

Defining the sampling selection frame and methodology identifies the target population, determines the sample frame (and, where necessary, the register from which the sample is taken) and specifies the most appropriate sampling criteria and methodology (which may include the entire target population). The most common sources are geospatial, administrative and statistical registers, and censuses. These sources can also be combined. An analysis should be performed to verify that the sample frame covers the target population. A sampling plan should also be developed.

After identifying the target population, it is necessary to obtain a list of population units with a larger number of their approximations, since the data sources used to create the sampling frame are not of appropriate quality. The sample selection frame variables that describe the properties of the units are called auxiliary variables[2].

[2] In the literature, it is possible to find more flexible definitions of the auxiliary variable. For example, a variable whose population value has been known before the survey is conducted, and data from selected sample units were obtained from the survey, is also called an auxiliary variable.

The method of data collection is taken into consideration when preparing the unit selection methodology. The requested data can be obtained from various administrative data sources, derived or modelled from existing statistical data sources, by using innovative approaches, including scientifically based and well-documented methods such as imputation, estimation and modelling or direct/indirect contact with units of the observed population.

Regardless of the way the data are collected, the statistical survey can be conducted either as a pilot survey or as a census. If sample-based survey is organised, it can be done in two ways: by using non-probabilistic or probabilistic sampling procedure. Selection of observation units for a probability sample is usually very simple, inexpensive, and it is not time-consuming, but, at the same time, it is subjective. The methodologist should determine the representativeness of the sample for the subject-matter survey, depending on the characteristics of the whole population.

Unlike non-probabilistic sampling, the use of probabilistic sampling[3)] is more demanding because sample units are selected with known, positive probabilities that are determined before selection, so selection also takes more time. However, since the probabilistic sampling plan is based on the whole sampling theory, much more reliable conclusions can be made about the characteristics of the whole population and sampling errors can be calculated. The selection of units for the probability sample can be carried out in several ways. The method of selecting a probability sample is called a sample plan. The choice of method depends on:

- auxiliary variables from the sample frame
- variability of population units (here we usually mean the variability of the value of the surveyed variable)
- statistics and areas where statistics are published
- research costs.

Part of the sampling plan is to determine the sample size that directly affects the price of the survey. The aim is to obtain the best possible estimates of population parameters with the smallest possible sample size. If there are enough data within the sampling frame, the sample size can be calculated (usually according to the required accuracy of statistical data). Otherwise, it is determined on the basis of experience (e.g. other statistical surveys).

The most common forms of sampling used in the Croatian Bureau of Statistics are:

- a simple random sample
- systematic sample
- stratified sample
- sample groups
- multi-stage sampling.

Instructions for quality insurance

When determining the model framework, it is necessary to assess what resources (registers, databases) are available and how they can be interconnected.

If possible, the quality of the set sample frame should be assessed as soon as it is prepared. If the content of the framework is judged to be of unsatisfactory quality, additional questions should be included in the design of the questionnaire to allow a reliable assessment of the quality of the sample for the next reference period.

If the sampling frame is based on a interviewing method (e.g. persons from the telephone directory instead of all persons from the database of the census of population, households and dwellings), it is necessary to make an assessment of the impact on the quality of the results.

[3] The term "random sampling" is used colloquially instead of "probabilistic sampling", although "random sampling" is mentioned in the literature as a special case of probabilistic sampling in which all samples of a certain size have equal probability of selection by sampling.

The most important sources for the design of the sample frame are the following registers and databases: for the sampling of persons and households it is the database of the census of population, households and dwellings; for company sampling it is the SPR; for sampling of family farms it is SRPG, a database based on Article 117 of the Agriculture Act (Official Gazette, No. 118/18) and IACS.

When designing the sampling frame, attention should be paid to the auxiliary variables because access to the auxiliary variables allows the calculation of the sample size, the implementation of a more complex sampling procedure and the determination of more accurate estimates of statistical data.

The database of the census of population, households and dwellings is a suitable source for selecting natural persons in the sample, with households being used indirectly for sample selection because the selected person, through the residential address, leads to the corresponding household. The probability of selecting households in the sample is estimated using field data on the number of household members.

Preference should be given to the simplest possible sample design. The decision to plan more complex sampling procedures should be well considered. For example, the use of a multi-stage sampling method can reduce survey costs, but also the accuracy of estimates.

When determining the sample size, the expected response rate and results of the non-response analysis from the previous reference period should be taken into account. In other words, it is necessary to assess which level of response provides a satisfactory quality of statistical results.

2.3. Development of methodology for data processing

Description

Statistical data processing refers to all procedures used after the completion of data collection. The development of a methodology for data processing includes determining the procedures for coding, editing, imputation, evaluation, integration, linking geospatial and statistical information, verification and finalisation of data sets with deadlines and holders. The main goal is that the final statistical results accurately reflect the characteristics of the observed population in order to obtain information, i.e. indicators that allow monitoring and comparability for national and international needs. These indicators also serve in policy making, in various scientific analyses, and, in general, for informing the general public.

In the survey planning phase, it is necessary to anticipate which procedures make sense for the implementation of the planned statistical survey, which methodological approaches should be applied in these procedures and which software tools would be the most appropriate for implementation.

Different statistical surveys contain in their implementation the following different statistical processing procedures.

Resolving non-responses. During the survey, it often happens that the response rate is low for a certain set of statistical survey data or that there are no data at all. Procedures to reduce non-response rates should be included in data collection activities. However, in the statistical processing planning phase, it should be anticipated which procedures will be used after the completion of data collection in order to minimise the impact of non-response on statistical results and to reduce the level of bias. Usually, survey results are weighted to accommodate sample design, and, for non-response units, valid results are produced for the target population. The unit of non-response is calculated by re-weighting. The problem of non-response can also occur with administrative data sources.

Data editing procedures. The data editing process applies to all procedures of idetifying and eliminating errors in the data. When planning data editing procedures, it is necessary to choose the procedures that will reduce time and cost structure as much as possible. Therefore, great efforts should be made and a solution found by which the so-called "manual“ data editing procedure will be replaced by introducing more modern approaches, such as selective and automatic data editing.

Data editing can also be applied in the data collection phase by applying data verification checks within the questionnaire and thus reduce the number of processing errors.

Aggregation and tabulation procedures. This part of the process is the „constant point" of statistical surveys, which help in the calculation of statistical aggregates (the so-called final product of the statistical process) from the processed data for statistical purposes at the micro level. In the planning phase, it is necessary to anticipate what types of statistics will be calculated, describe the rules for their calculation clearly and unambiguously and support the accompanying text with mathematical formulae. It is also necessary to ensure a satisfactory level of detail in the scoreboard from the perspective of users and data producers.

Moreover, it is necessary make the following plans in the process of statistical processing, depending on the type of statistical survey or the needs of users.

In the case of sample surveys, procedures must be provided to estimate the sampling error. It is necessary to decide on the use of three basic approaches (analytical approach, model approach and re-sampling approach) and software tools.

Particular care should be taken with financial data, when deflated statistical results are provided to users. Deflation is the process of removing the impact of price changes on valued phenomena. It is carried out by dividing nominal values by a corresponding price index. In the planning phase, it is necessary to anticipate which approach will be applied, for example, whether deflation will be applied at the micro or macro level, then what methodology will be used to calculate deflators, what basic resources will be needed and how this process will fit into the overall statistical process.

The most important economic indicators are often influenced by seasonal and calendar variations that can mask relevant short-term and long-term trends and make it difficult to understand economic phenomena (GDP, industrial turnover, volume of construction works, retail trade turnover and industrial production volume). If it is an occasional statistical survey and if the results of the survey are presented in time series, it is necessary to predict whether the data will have to be seasonally adjusted. The application of these procedures depends on whether there is a seasonal effect and the effect of the number of working days on statistical results. What this is actually about can be determined analytically after a certain number of repetitions of activities. Likewise, the mentioned impact can be predicted when the content of the statistical survey is known in detail.

Instructions for quality insurance

When planning statistical activities, it is necessary to follow modern statistical methods and methodological solutions as well as good practices of other national statistical institutes. When designing statistical processing tools, it is necessary to determine whether there are software tools that could be used or it is necessary to develop tailor-made solutions. If it is necessary to apply new software solutions, it is necessary to choose the most appropriate software tool. The functional capabilities of individual software tools and the good practice of national statistical institutes should be taken into account.

In cases of non-response of the unit in the sample, it is necessary to decide whether to use data weighting or insertion for all variables. The following parameters should be taken into account to make a final decision:

- number of variables to be inserted (insertion methods are usually used to solve the unit non-response problem in the sample only when there are few variables in the survey)
- is it a periodic survey in which historical data can be used to insert data (data of individual examined units in the sample)
- whether there are external (e.g. registers) administrative data sources at the micro level that could be used to insert the data.

Selective and automatic data editing methods should be used as much as possible in editing procedures.

When determining the details of the data publishing level, it is better that this level is not too detailed because, otherwise, it could result in a large number of empty, dimmed or protected cells. Statistical confidentiality should be taken into account, which implies the protection of confidential data that can be determined by direct identification (name and surname, address, publicly available identification number) or indirect identification (any method other than direct identification). This recommendation does not apply in cases where it is prescribed otherwise.

3. Development of necessary instruments for implementation

When determining the target population for a particular statistical survey, it is necessary to decide whether to conduct it on the whole population or only on a part of the population. Since conducting a population-wide survey is expensive, time-consuming and demanding for both those who conduct the survey and those who respond to the questionnaires, most surveys are conducted based on sample selection. This means that, on the basis of the obtained data, only parts of the population with specific characteristics important for the implementation of statistical survey are examined.

Even before making a decision regarding the planning of the sampling procedure, which will determine the method of sampling, it is necessary to theoretically define the population and prepare as many basic characteristics about it as possible. The actual record of population units, together with some characteristics that are important both for sample selection and for calculating estimates of population parameters, is called the sample selection framework. Before preparing the sampling frame, it is necessary to prepare all data sources (registers, data from other surveys) with which the sample selection framework will be compiled. In practice, the sample selection framework is a more or less accurate approximation of the target population (and some of its characteristics). The quality of the sample is reflected in the quality of the final survey results.

Every effort should be made to prepare all data sources needed to compile the sample selection framework in order to achieve the highest possible quality of the sample selection framework. Once the sample selection framework has been determined, a sampling procedure plan is established to select the units in the sample in order to conduct the statistical survey. The pattern of units can be probabilistic, improbabilistic, or a combination of both. Once the sample of observation units is determined, an address book of the units selected in the sample is compiled. The address book contains a list of addresses (and other contact details) of reporting units.

3.1. Development of project requirements

Description

The preparation of project requirements includes requests for the collection and processing of data with deadlines and holders in accordance with the instructions for the preparation of the SIT project requirements, regardless of whether it is a sample survey or a full survey (census).

It is important to understand the difference between the target population and the sample selection framework. The target population of a statistical survey is a set of units whose specific characteristics are observed and to which all the results of a particular survey relate. The sample selection framework is the actual list of target population units available at the time of survey preparation and is used to select the target population units in the sample.

The target population (persons, households, dwellings, family farms, enterprises, etc.) must be precisely defined before setting the sample selection framework. Depending on the characteristics, the target population and the conditions that the unit must meet in order to be part of the population are determined. It is also necessary to determine the geographical position of the units and the reference period (time) with the characteristics of the population important for the implementation of the survey.

Example: Our target population are persons living in private households who were residents of the Republic of Croatia on a certain date and were at least 18 years old on that date. In this matter, it is necessary to take into account the definition of "resident of the Republic of Croatia“ as well as to define the conditions when a person can be considered a member of a private household (what is a private household?; What to do if a person lives in one household at one time and in another household at another time; Which household does such a person belong to?). These are conditions that define the target population and, therefore, the conditions which the unit must meet in order to be or not to be suitable for the survey.

In the statistical survey, one must distinguish:

- sampling units (units from the sample selection framework selected by sampling)
- observation unit (the unit on which data are collected and which is therefore part of the target population)
- reporting unit (the unit that submits the data to be collected).

In some studies, all three traits are combined into one whole, but this is often not the case.

Example: The sampling unit is a selected person from the database of the census of population, households and dwellings who lives in a dwelling at a specific address and has provided the information prescribed for the survey. That person is the reporting unit.

All the features need to be taken into account when compiling the sample selection framework. This means that the sample selection framework often does not consist of units that are subject to observation, but of units that lead statisticians to observation units.

When the target population is known and if units of that population can be accessed, the procedure for defining the sample selection framework can be carried out. Depending on the target population, different data sources are used, either in time intervals or in periods closer to the reference period of the survey or to the time of data collection.

If the data source used covers the reference period of the survey, the population will be recorded more accurately in the reference period. These may be units that were not relevant at the time of data collection. However, they were still present in the reference period of the survey or some characteristics of the units changed (e.g. the company changed business). If the data source is more recent than the time of data collection, the number of units that do not meet the set criteria has to be reduced to a minimum. It can also happen that units have some characteristics that are valid at present time, but not for the reference period, and can be mistakenly put together in groups or clusters.

Tipically, one basic data source is chosen as the sample selection framework. After that, the quality of the sample selection framework is improved by additional sources. The source that best enables the collection of data on the target population is chosen as the basic data source. Additional data sources complement the basic data source.

As a rule, the basic source of data is business statistical surveys, i.e. the SPR of the Croatian Bureau of Statistics, which contains a list of all companies. Additional sources of data may be Fina's annual financial report and VAT data. These data sources allow for determination of the size of the enterprise rangea.

In statistical surveys in which data on persons and households are collected, the main source of data for survey questionnaires used for field survey is the database of the census of population, households and dwellings. A census of the population, households and dwellings is conducted every ten years. The database of the census of population, households and dwellings is updated annually with data from birth and death registers and migration balance; the main source of data for survey questionnaires conducted by telephone may be the telephone directory.

Sometimes one data source is not enough, so it is necessary to create a sample selection framework from several different data sources (administrative data sources, database of the census of population, households and dwellings and other statistical surveys). This is the case when different data sources cover different periods and different sections of the population. In this case, it is important to use a unique identification for all population units and for all types of data sources. Therefore, only data sources in which common identification can be established are relevant. If this is the case, one data source should be defined as the primary data source.

The target population and the sample selection framework rarely match in all units. There are units that are part of the target population that are not listed in the sample selection framework. On the other hand, there are units listed in the sample selection framework that are not actually part of the target population. Both cases can cause errors in coverage that largely determine the quality of the final results of the statistical survey.

Instructions for quality insurance

The reference periods to which the data refer in the data sources that will be used in the development of the sample selection framework should be as close as possible to the reference period of the statistical survey. If the reference periods are different in different data sources, these differences should be recorded and documented.

It is necessary to check whether the variables that appear in several different data sources comply with the methodological definitions. If it is determined that they differ from these definitions or do not match with them, these differences must be recorded and documented.

The field must contain the following information: unique identifier of the observation unit, contact information and classification code of the variable. It would be useful if the sample selection framework also contained the status date of the sample selection framework and the linkage variable.

It is necessary to achieve the consolidation of all data sources from the point of view of formal checks before aggregation, such as the validity of the code list, the appropriate range of variable values and the problem of duplicate units.

The data sources used to create the sample selection framework should be properly documented with a brief description of those sources, which also applies to records that cannot be used to create the framework.

All available data sources should be used in the preparation of the sample selection framework, as each contributes to the maximum harmonisation of the sample selection framework with the target population. Any differences between the sample selection framework and the target population should be documented.

3.2. Development of data collection instruments

Description

It involves the development of instruments to be used in the data collection phase. The data collection instrument is developed on the basis of design specifications made in the preparation phase.

Before conducting a statistical survey with the questionnaire, it is necessary to prepare supporting documentation that is used in communication with reporting units. In the announcement letter, the reporting unit learn about an individual statistical survey, main objectives of the survey, methodological basis of the survey, to whom the survey is focused, obligation of statistical reporting, method of collection and deadlines, confidentiality of data and contact persons if there are any questions or ambiguities. The intention is to encourage reporting units to participate. Therefore, the most interesting survey results are presented to the general public. The results can emphasise the meaning of conducting the survey and stimulate reporting units to participate. Usually, two reminders are sent, and only in exceptional cases three of them. If there is only a small number of reporting units, they can be "urged" by phone and it is not necessary to send another reminder. In cases when the cooperation was terminated due to cessation of operations, or the Croatian Bureau of Statistics redirected its activities to other data sources, the participation of each reporting unit is no longer required. This is also true for surveys with occasional rotations of observed units when a particular reporting unit is no longer included in data collection.

All documentation is prepared for the current year based on standardised templates. Documentation must be proofread before its distribution to the field.

The preparation of the questionnaire is an important part of the statistical process because, if it is sound and well prepared, the collection of reliable data from reporting units will be at a satisfactory level. Thanks to long-term cooperation with reporting units, the Croatian Bureau of Statistics is aware of the form of necessary statistical data held by reporting units.

Before creating a working version of the questionnaire, it is advisable to consult with users and data holders, collect and review questionnaires of all previously conducted statistical surveys and prepare a list of all variables for the survey. For each variable within the statistical activity conducted according to the annual implementation plan of statistical activities of the Republic of Croatia, it is necessary to state the source of data, i.e. whether the source of statistical survey data is a questionnaire, administrative source or another secondary data source. In the case when data are taken over from the register, the reference period and information on the availability of data sources must be indicated. The decision on the method of data collection must be made before the preparation of the working version of the questionnaire and the latest standards applicable to the statistical system of the Republic of Croatia must be taken into account.

The wording of the question depends on various factors, such as: method of data collection (online questionnaires, Excel questionnaires, telephone survey, field data collection using laptops, printed self-completion questionnaires), characteristics of the data providers (respondents), avoiding excessive burden on units and the complexity of the data that are collected with a high level of data confidentiality and information sensitivity. Also, it is necessary to consider comparability with results of other surveys and take account of the relevance. The questions asked should have the same meaning for all data providers (respondents) involved in the survey, as well as other criteria that should be met, as follows:

- availability of necessary data
- willingness of respondents to submit quality data
- probability of non-response
- wording, in the semantic sense, of each question
- the layout of the questionnaire
- source of errors in expressing values in units of measurement and errors in answers
- data entry (survey entry, entry of the speed of filling in the quesionnaire).

Preparation of the working version of the questionnaire

After designing the questions and variables, a version of the questionnaire containing logical checks and skips is prepared. When preparing the electronic questionnaire, developers are consulted about technical possibilities and standards. After that, a working version of the questionnaire is created. The questions in the questionnaire are divided by topics. If it is a questionnaire that needs to be downloaded in accordance with the harmonised questionnaire (for comparability purposes between EU countries), attention must be paid to its translation from English into Croatian, since the terms used in the English version are often unfamiliar or uncommon in the Croatian language.

Verification and revision of the questionnaire

Before testing the questionnaire on the target population, it is important to check all items of the questionnaire. The questionnaire is sent for proofreading and technical editing. It is important to use terms that are professionally well translated. Internal check of the content of the questionnaire should check the grammatical correctness of the questions asked and improve awkwardly asked and incomprehensible questions. Internal check means that the text of the questionnaire is tested by persons who are not directly involved in a particular statistical survey. The persons testing the questionnaire can be experts from other statistical surveys, interviewers or representatives of the target population. All these elements can contribute to improving the final version of the questionnaire.

Technical testing of the questionnaire

In the case that data are collected on an electronic questionnaire (by hiring interviewers or using an online self-completion questionnaire), it is the statistician(s) in charge who technically checks the content of the questionnaire. Such verification includes testing of skips testing and the operation of logical controls. After making sure that everything in the questionnaire works technically well, the content is checked.

Testing and revision of the questionnaire

The following activities are included in this process: informal testing of questionnaires, cognitive methods, focus groups, interviewers' reports, coding of interviewers' or respondents' behaviour, various experiments (e.g. split -sample testing) and pilot testing of questionnaires. It is advisable to perform a cognitive questionnaire test before conducting field pilot surveys.

Completing the questionnaire

Since the design of the questionnaire includes an iterative procedure, it would be advisable to occasionally test the questionnaire.

Instructions for quality insurance

When preparing questions, it is important that the reporting unit understands the questions asked, that it has information to prepare answers to the questions asked, and that it is willing to answer the questions. The manner in which questions are asked needs to be adapted to the method of data collection.

It is always necessary to first check which data are available in administrative data sources, for which reference period and when they are available. The questionnaire, which should be as short as possible, is prepared depending on the availability of data from administrative sources. Furthermore, it is necessary to check whether all basic variables prescribed by regulation or methodology are included in the survey.

Before conducting either pilot surveys or regular surveys, the questionnaire is checked at a detailed level. If it is an electronic questionnaire, first the whole questionnaire is checked internally, followed by technical testing (functioning of logical controls and skips) and finally the content of the questionnaire, i.e. understanding of the questions asked, is checked on the target population.

When preparing the questions, the characteristics of the respondents, i.e. the data providers, should always be taken into account and accordingly adapt the complexity level of the terms and sentence structure used in the questions. The questions intended for the general public or the target population need be clear and understandable in all groups.

The questions asked in the self-completing the questionnaire should be short and clear. The questionnaire must not be extensive. Clear instructions and examples should be provided. It is useful that questions asked are accompanied with practical examples.

There is no mediator for these types of questionnaires, such as an interviewer who could explain the content of the question to the respondent. The decision on how to collect data and how to enter data depends on the financial resources available, the value of the non-response rate, the time frame in which the data were to be collected, the target population and the information available in the sample selection framework. Exceptionally, printed questionnaires can also be used for field data collection if it is estimated that data collection using laptops will take longer than filling in on paper.

Methodologists should make sure that the number of questions in the questionnaire is kept to a minimum. For each question asked, there should be a reason why the question is included in the questionnaire. Therefore, verification in administrative data sources and records is required.

In order to make it easier for the respondent to fill in the questionnaire, it is possible to fill in the questionnaire automatically from another data source and, in that case, the respondent can correct the answer to the question if it is not correct.

For practical reasons, it is sometimes useful to add to the questionnaire a question that is needed in another statistical domain. This way of working is more demanding for the Croatian Bureau of Statistics, while it is less burdensome for the reporting unit.

Particular attention should be paid to the questions raised in order to obtain more complex data. In such cases, questions covering complex topics should be added to the instructions. The instructions help interviewers and respondents to more easily prepare answers to the questions asked.

Confidentiality and sensitivity of the data affect the design of the questionnaire. The questionnaire and the notice in the announcement letter should state that collected statistical data on natural persons or legal entities are statistically confidential and represent an official secret if they can be directly or indirectly linked to those natural person or legal entity. Confidential statistical data are presented in a form in which the statistical unit cannot be identified either directly or indirectly and they can neither be the basis for determining any rights and obligations of reporting units for administrative, legal or tax purposes, nor can they be used for any kind of verification of reporting units.

In the case when sensitive topics or topics that are uncomfortable to discuss are the subject of the survey, more accessible ways of asking questions can be used.

It is also important that questions for foreign respondents are asked in the spoken language of the target population. If the questionnaire is translated from a foreign language, cultural differences and habits must be taken into account when translating. For example, in censuses of population, households and dwellings, questionnaires must be translated into the languages of minorities.

In order to achieve comparability of survey results, it is important that the questions are designed in the same way.

3.3. Software development

Description

It describes the actions to be taken to build new and upgrade existing software components required for the business process designed in the preparation phase. Components may include control tables and reports, databases, result tables, data transformation tools, and tools for managing data, geospatial data and metadata.

All phases of business processes should be clearly and thoroughly documented. For example, in order to process statistical survey with the help of the Survey Processor, it is necessary to prepare a "Project Request", which is also a work order for all tasks related to survey processing submitted by statistical organisational units to the Information Technology Sector (SIT).

The following types of tasks are distinguished:

1. request for new processing
2. supplement to the application already submitted to SIT
3. change in the request already submitted to SIT.

The application contains a number of forms of standard design and content, which serve to document the processing of the survey. It is submitted to SIT for all new processings, as well as for additions and changes in existing processings. The set of attachments must be arranged in a way which enables identification and implementation of all necessary interventions. It is necessary to indicate the type of the work needed on the request.

The documentation should also contain material and technical task descriptions. It should, as far as possible, contain assessments by various stages of the business processes.

Instructions for quality insurance

The CROMETA central metadata database is the core of the system for processing statistical surveys, the so-called Integrated Statistical Information System. It was named CROMETA according to a model developed in collaboration of the Croatian Bureau of Statistics and the Swedish Statistics (SCB), and sponsored by the Swedish International Development Agency (SIDA).

The first version of the model was made in 2005, and it has continuously been evolving ever since according to needs. The CROMETA metadata database model is designed according to the so-called reference model developed in Eurostat 's METANET project (2000 – 2003). The reference model combines different types of metadata that describe statistical data and processes.

Within the CROMETA model, there are different types of metadata according to purpose. Some metadata serve as a declarative description of the content of statistical surveys and data, while others are process-oriented and describe ways of collecting and processing data. In addition, there are global concepts and methodologies used independently of statistical survey, as well as general metadata describing organisational structure, authorisation, and access rights.

The Metadata Manager application is used to manage metadata in the CROMETA central metadata database. The application is available only to authorised users who have permissions to work with metadata, depending on the group to which they belong.

The Warehouse Browser is an application designed to view data stored in the statistical data warehouse and described in the central metadata database. It can be used to generate output tables and save results in the form of aggregated data (cubes) or prepared tables.

3.4. Testing of data collection and processing tools

Description

It includes technical testing and approval of new programs and procedures. It also includes testing the interaction between components and ensures that the production system functions as a harmonised set of components. Furthermore, it includes data collection for experimental surveys to test data collection instruments.

Before launching a survey, especially if it is a new survey or in a situation where neither the content nor the target population has been checked, a pilot survey should be conducted, which helps in determining and preparing the survey methodology. The objectives of the pilot survey must be clearly defined as the plan and purpose of the sample survey need to be adapted to it. In the pilot survey, it is possible to test either the whole procedure or only individual parts of the business process. The following activities are most often carried out by conducting a pilot survey: testing the questions, the required sample size, the sampling procedure plan and the method of data collection. The sample size is usually much smaller than when organising and conducting a regular statistical survey. After the implementation of the pilot survey, the results are analysed in detail and, based on the obtained results, further procedures in the implementation of the survey are determined.

Pilot survey can also be conducted in cases in which data are completely downloaded from administrative sources. Goals of the pilot survey have to be defined regardless of the fact that data are taken over from an administrative source.

Instructions for quality insurance

In accordance with the Official Statistics Act (Official Gazette No. 25/20) and the programming documents, the Croatian Bureau of Statistics collects data using statistical forms, which are the basic tool for conducting statistical field survey.

The statistical forms used in this procedure are of different information/methodological forms and contents. The Croatian Bureau of Statistics is gradually developing a new electronic data collection system in accordance with the available resources. Electronic forms provide reporting units with an alternative way of submitting the data required of them, as opposed to a system based on filling in classic forms in printed form that are usually delivered by mail. The following statistical surveys are included in the electronic data collection system:

- Statistical reports on industry
- Annual Survey on Usage of Information and Communication Technologies (IKT-POD, IKT-DOM forms)
- Statistical reports on energy
- Statistical reports on construction
- Monthly Report on Retail Trade (TRG-1 form)
- Statistical reports on service prices (SPPI forms)
- Monthly Report on Distributive Trade and Other Services (USL-M form)
- Statistical Survey on Road Transport of Goods (PA/T-11 form)
- Statistical reports on transport (PA/M-11, PR/T-11P, PP/T-11 forms)
- Statistical reports on tourism (TU-14, TU-11i forms)
- National Travel Survey
- Reports in the domains of agriculture, fishery and environment (PO-22/STR, PO-32 forms)
- Quarterly Report on Sale of Agricultural Products (PO-31a/Q, PO-31b/Q forms)
- Statistical reports on criminal justice
- Investments in Environmental Protection and Expenditures on Goods and Services in Environment (IDU-OK form)
- Annual Report on Gross Investment in Fixed Assets of Legal Entities (INV-P form)
- Annual Statistical Report on Persons in Employment and Paid-off Earnings (RAD-1G form)
- Vocational training in enterprises in 2020 (SOP)
- Innovation Activities in Enterprises, 2018 − 2020 (INOV form)
- Annual Statistical Report on Basic Schools (Š-O/KP form)
- Annual Statistical Report on Upper Secondary Schools (Š-S/KP form)
- Annual Report on Kindergartens and Other Legal Entities Implementing Preschool Education Programmes (DV-PO form)
- Statistical Report on Divorce (RB-1 form).

In the case of traditional data collection, non-responding reporting units should be provided with a reminder or requested to re-participate in the statistical data collection process. It is common practice to deliver two reminders.

In statistical surveys that includes persons and households, with the participation of an interviewer, the selected reporting unit is informed about the arrival of the interviewer and about the survey in an announcement letter. The announcement letter is usually accompanied by a leaflet with answers to the most frequently asked questions posed to respondents or observed units.

If the target population includes minors, the announcement letter also serves to ask parents or guardians to allow their children to participate in the survey.

In case a combined survey method is used in a statistical survey, e.g. a combination of telephone surveys and field interviews, the normal practice tries to include reporting units that were not covered by telephone surveys and it aims to reduce bias due to non-response rates. In this case, an additional letter must be sent to these units. The content of the letter is adjusted depending on the final reason for non-response in the first survey.

In postal surveys, a reminder or re-application for participation are provided. Care must be taken to submit an envelope for which postage has already been paid in advance and on which the address of the Croatian Bureau of Statistics is written, as well as an additional questionnaire, together with a repeated request to participate (reminder).

Announcement letters should contain basic information about the survey and the way the interviewers work.

The objectives of the pilot survey should be clearly defined. It should examine different methodological aspects: question formulation, individual non-response rates, notification effect, etc. The results of the pilot survey should be used in the preparation of the main survey.

In the pilot survey, the sample design should be adjusted so that differences in the responses of each subgroup can be tested; the effect of notification letters, various forms of questions, etc. can be tested.

The pilot survey is also intended to verify the questionnaire. However, technical tests of questionnaires and cognitive characteristics must be carried out before conducting a pilot survey. When testing the questionnaire, it is necessary to make personal contact with the reporting unit. Sometimes it is necessary to test statistical processing and tabular calculation procedures in a pilot survey.

3.5. Configuring the flow of production processes

Description

Configuring the flow of production processes refers to data collection all the way to archiving the final statistical results. It includes activities that put the process into production, ready to be used by business areas.

All statistical activities carried out in the Republic of Croatia, with the Croatian Bureau of Statistics and other administrative organisations and institutions as their main producers, are described in the GPP. Based on GPP, metadata on statistical surveys were entered into the CROMETA database and supplemented with contact information along with some other metadata. The GPP provides an overview of statistical surveys and activities grouped into five main chapters: Demographic and Social Statistics; Economic statistics; Sector statistics; Environmental and multi-domain statistics and Methodology of data collection, processing, dissemination and analysis.

Quality assurance instructions

The legal framework is the starting point for the organisation and implementation of statistical activities in the Republic of Croatia, i.e., it is the Official Statistics Act (Official Gazette No. 25/20).

The structure of each statistical survey or activity is prescribed by Art. 36 of the Official Statistics Act. The Annual Implementation Plan of Statistical Activities of the Republic of Croatia regulates the implementation of the following official statistics activities:

- statistical surveys based on direct data collection
- activities of official statistics whose data are obtained from administrative sources or by using the observation and monitoring method
- development activities, censuses and other more extensive statistical surveys.

The Annual Implementation Plan of Statistical Activities of the Republic of Croatia determines the titles of official statistics activities, responsible producers of official statistics, reporting units, periodicity of surveys, deadlines for data transmission or collection, indicative deadlines for result publishing, level of result publishing, relevant standards and, if needed, other characteristics of official statistics activities.

The control mechanism used for monitoring statistical activities is the Eurostat's statistical manual called Statistical Requirements Compendium, which specifies current user manuals, standards, technical documentation and professional training courses, e.g. in ESS metadata standards.

4. Data collection

Data collection is a procedure that is carried out at the beginning of a statistical survey and significantly affects the final result. Several ways of collecting data for statistical purposes are known. The most well-known way of collecting data is statistical surveys, which are used to collect data exclusively for statistical purposes. Recently, administrative sources have been used as a second method of collection in combination with field data collection for statistical purposes.

In fieldwork, the first phase involves contacting reporting units, while the second phase involves collecting data using questionnaires in the form of closed or open-ended questions. The questions in the questionnaires for statistical surveys conducted by the Croatian Bureau of Statistics are mostly quantitative, and less often qualitative. Questionnaires and data collection methods must enable the highest possible data quality, control of measurement error, the lowest possible burden on reporting units and the lowest possible costs.

The decision on how to collect data is influenced by several factors: characteristics of the target population, objectives of data collection, available resources and time constraints. In addition, it is necessary to know the characteristics of statistical surveys in order to be able to choose the appropriate method and appropriate data collection technique.

There are two types of statistical surveys: longitudinal and cross-sectional. Longitudinal survey is used to monitor the dynamics of the observed phenomenon. In such statistical surveys, data for the same units are collected more than once at regular intervals, while in cross-sectional statistical surveys they are collected only once or occasionally. Panel survey is a survey that involves the same units in several consecutive phases of the survey.

The most common ways of collecting data are self-completion (postal method and electronic reporting) and interviewing (telephone and personal surveys). The first approach is commonly used in business surveys, while the second one is used in surveys in which individuals and households are observed. There are two types of data collection techniques: recording responses on paper and recording responses on a computer. When printed questionnaires are used, the data must be digitised; digitisation of data can be done by manual input or scanning. In both ways of digitisation, input errors are a problem because both the human and the scanner can make mistakes. Therefore, after digitisation, additional check of the entered data is required.

Combined data collection methods can also be used, such as a telephone survey and then field interviews. Each method of data collection has its advantages and disadvantages, and the following ones are the most typical.

Self-completion is primarily a cost-effective way of completing the questionnaire, i.e. entering the necessary data into the questionnaire. In addition, it is best suited for collecting sensitive data. The reporting unit may complete the questionnaire whenever it wishes. Several questionnaires can be sent to the same recipient at the same time. The set of reporting units can be very large. Because the interviewer has no influence on the answers, the reliability of the answers may be worse. It is often necessary to urge reporting units that do not submit data on time. It is not suitable for open-ended questions or for long, complex questionnaires because they burden the reporting unit. By applying this method, control over the interaction between the reporting unit and other persons cannot be achieved, and this method usually increases the work related to data editing.

Electronic questionnaires can be online questionnaires. Therefore, a message/notification can be sent to the respondent with a link or information on which internet address he can fill in the questionnaire. If the electronic questionnaire is created locally on a computer, it is used as a CATI or CAPI questionnaire. The main disadvantages of this method of completing the questionnaire are non-response and insufficient coverage because the availability of technology depends on various factors (for people, for example, on income or age, while for companies on the type of activity or size of the company). Electronic questionnaires are more appropriate for surveys in which observed units are business entities. Electronic questionnaires allow the use of various more complex forms, skipping questions, access to registers, code lists, etc. In addition, coding of answers can also be included, e.g. occupations. This can reduce the amount of input errors and enable faster compilation of collected data, since logical controls are already part of the questionnaire.

Field interviews (printed questionnaire for interviewer and computer-assisted interviewer) are the most expensive way to collect data due to the use of human data sources, time and financial restraints. The presence of the interviewer helps in directing the situation and explaining the questions asked. However, the interviewer can influence the answers. Due to his presence, sensitive questions can be left without answers, so sometimes it is better for the respondent to answer some questions himself, if there is a possibility to predict such an impact. This method of work reduces the burden on the reporting unit, but some respondents do not want the interviewer to enter their home. By using this method, data are collected at the time the person discovers the event. They are suitable for downloading data in places where there is no telephone, and are also used when the reporting unit is not known. For very short questionnaires, the time required to collect data should also be taken into account. Sometimes it is possible to complete a printed questionnaire faster than turning on a computer, running a programme and completing an electronic questionnaire. Electronic questionnaires are not suitable for outdoor surveys (for example, at border crossings).

Telephone interviewing method helps in achieving a very high response rate if the questionnaire is short and clear. This is supported by the accuracy of telephone numbers and the ability of interviewers to persuade respondents to answer questions from the survey. It differs from field interviews in that the spatial distribution of units does not affect costs, so telephone interviewing is much cheaper than field interviews. On the other hand, there is a high risk of insufficient coverage (households without a telephone or with an undisclosed number). Collecting data through telephone interviewing can be very fast. Due to the fact that it is computer-assisted, the course of the interview can be complex and may contain open-ended questions as well as many skips, but not in the same way as in field interviews. The influence of interviewers is less due to the existence of a certain protocol. It also allows considerable control over the interaction between interviewers and between interviewers and controllers. A telephone interview is a good compromise between postal surveys and field interviews, but choosing between several possible answers (more than six) requires visualisation, which is only possible with other data collection methods.

The main problems in using this method of collecting data and which can greatly affect the result of the survey are non-response and inaccurate data delivery. Non-response reduces the effective sample size, which in turn results in a larger sampling error. However, in most cases the groups that did not respond to the questionnaire may be different, and this can then lead to bias in the results. The response rate must therefore be as high as possible because the higher the response rate, the lower the risk of bias and large errors in the sampling method. These goals can be achieved by careful preparation of the data collection process, testing of measuring instruments, quality execution and control of field work and data processing. A number of constraints must be taken into account, from financial and human resources, deadlines to data sources. Special attention should be paid to the organisation of field work and communication with respondents or reporting units. Errors in the data received can only be detected by editing the procedures carried out after the collection phase.

In his records, the database administrator keeps data that have not been collected for statistical purposes. The use of databases achieves a more efficient use of this type of data source. On the other hand, such an application means dependence on the activities of the database administrator, both in methodological and technical sense. The location and form of data storage depends on the database administrator. When downloading databases, the Croatian Bureau of Statistics applies internal instructions for downloading databases.

Therefore, data collection needs to be well prepared, the work plan needs to be harmonised and, finally, the data need to be converted into an electronic format suitable for further processing. What this process will look like depends on whether the data were collected through self-completion or an interview. Ongoing quality control is paramount when collecting. Questionnaire checks are often performed before entering data into the input database. The download of administrative data sources is characterised by the application of the same phases of work, some of which are much shorter or automated, so this part is covered by a separate point.

4.1. Selection of target population/sample

Description

If a quality sample selection framework is to be prepared, multiple data sources need to be used. Once these data sources have been identified and collected in an appropriate computer format, they should be combined using the appropriate methodology to produce (design) variables that determine the key characteristics of the units included in the sample selection framework. The ultimate goal is to prepare a unique data table, which will contain a list of units corresponding to the theoretically defined target population, in which each unit determines the values of variables that will be needed in the later process of selecting observation units.

A key step in implementing the sample selection framework preparation process is to determine the process for selecting the units to be included in the sampling frame. The starting point of the sample selection frame is the database of observed units at the selected time point. A quality list of units, determined by a selected time point from the register, is sought to be improved by the use of additional administrative and statistical data sources. The primary goal is to detect and then exclude from the list units that are present in the register but are not actually part of the target population. Sometimes a non-registered unit can be included in the sampling frame, especially if it is a business survey. The reason for the existence of such inappropriate units in the register is the insufficient number of procedures for updating the register or it is caused by the administrative nature of the register. As already mentioned, in some cases attempts are being made to identify missing items, i.e. parts of the population not covered by this register, by using additional data sources. The reason for such unsatisfactory coverage usually stems from the definition of the population to which the register relates because it does not meet the completely specific needs of statistical survey.

If the sample selection framework serves its purpose, it must be ensured that, for each of the units, it contains the values of the variables that will be needed in later procedures. In the sampling process, these are mainly stratification variables, variables that determine the group of the first phase (in the two-stage sampling plan) or other variables that will be needed in the sample design or that define the areas where we want to publish results. It is also necessary to ensure that each unit has a unique identifier that will be used later in the survey process.

If the probability sample is selected by the method of persistent Bernoulli numbers[4][5], it is a procedure that is repeated from time to time and which reduces the burden on the reporting units. It is then necessary to transfer all random numbers from the previous sample selection framework to the current sample frame and the new units within the frame. A random number (between 0 and 1) is added by an appropriate computer procedure. The application of the method of persistent Bernoulli numbers is very useful in coordinated sampling, in which samples are selected simultaneously for several surveys. Coordinated sampling helps in monitoring the burden on reporting units in several surveys at the same time.

The sample selection framework is a list of basic set units whose characteristics are described by statistical results. This means that in the next step it is necessary to select among the units in the sample selection framework those units that will be included in the survey and about which the desired data will be tried to be obtained at later stages. For practical and financial reasons, it is seldom affordable to include all units in the sample selection framework in the survey. Usually, the list of sampling units needs to be adjusted to a size that can be financially supported and that, at the same time, allows a sufficiently representative calculation of results for the whole population.

---

[4] Bernoulli's numbers $B_k$represent a series of rational numbers discovered by Jakob Bernoulli, and they are related to the sum.

[5] Jakob Bernoulli (Basel, 6 January 1655 – Basel, 16 August 1705) was a Swiss mathematician. From 1687, he worked as a professor of mathematics in Basel. He made important contributions to the theory of infinite series, solved some of the basic problems of calculus of variations and significantly improved the theory of probability in the posthumously published work The Art of Guessing (lat. Ars conjectandi, 1713), in which the law of large numbers appears as the main lesson.

If it is a probabilistic sampling method, and if the sample size is not determined when planning the survey, and if within the sample there are data for auxiliary variables that are positively correlated with the investigated variables or if data for the surveyed variable (for sample units) from the previous reference period are known, the appropriate sample size is calculated. The criteria for calculating the appropriate sample size (excluding costs) are determined by the required accuracy of the estimated parameters to be published. These requirements are usually given at the level of the whole population, and sometimes at the level of regions.

If the auxiliary variable in the sample selection frame is positively correlated with the surveyed variable, then the corresponding sample size is obtained directly from the formula for the sample variance of the (main) parameter estimator. Of course, population data of the auxiliary variable are used in these formulas because the values of the surveyed variables are not yet available at this stage of the survey. If the estimators are in a complex form, then a Monte Carlo simulation is performed[6]. It is a process in which a large number of probabilistic samples are selected for certain sample sizes. Based on the data from these samples, the sample variance is calculated and the smallest quantity at which the desired accuracy is achieved is selected. If there are several key parameters or key variables, then the mentioned procedures are performed for all parameters or for all variables, and the final sample size is determined as the maximum of the calculated values.

If the data for the subject variable from the survey for the previous reference period are used, then the so-called sample effect (Deff)[7] of the appropriate sample size is calculated. It is the ratio between the sample variance of the applied sample design and the sample variance of simple random sampling without repetition. In practice, this procedure is used very rarely. Based on the calculated estimate of sample variance and the estimated parameter from the data for the previous reference period, and assuming a simple random sample is used, the effect of the Deff sample can be calculated. By further assuming that the effect of the sample remains the same when changing the sample size (usually it is desirable to increase the sample size) and the desired sample variance in the actual sample design, the required sample size is calculated. The calculated sample size is divided by the expected non-response rate to obtain the final sample size.

Once the sample selection framework is ready and the sample size has been determined, the selection of sample units that has already been determined in the survey planning phase begins. Appropriate software and certain algorithms are required for this phase of survey implementation. Special mention should be made of the stratified sampling plan, in which the stratum allocation of the sample must first be calculated. This means that for each stratum the number of units to be selected on the sample must be determined. The most commonly used are:

- equal allocation: the same number of units from each stratum is selected
- proportional allocation: units from each stratum are selected so that the share of each stratum in relation to the sample frame size is equal to the share of selected units from this stratum in relation to the sample size (this means that more units are selected from a larger stratum)
- optimal allocation: based on the data for the auxiliary variable or according to the data from the previous reference period, the appropriate sample size in each stratum is calculated and the variability of the auxiliary or surveyed variable of the previous period is taken into account (a larger sample size is determined in strata with larger variability).

In all the mentioned allocations, care is taken that the number of selected units is not too small in any stratum. If, after allocation, the sample size in the stratum is too small, a minimum fixed size is determined (usually about ten units) or the whole stratum is selected in the sample if its number of units is less than this fixed number.

[6] Monte Carlo methods, or simulations, are a broad class of algorithms that rely on repeated random sampling for the purpose of obtaining numerical solutions to problems that are difficult to solve.
[7] Sample Design Performance Indicator (Deff).

Sample selection results in sample weight (in the case of a probabilistic sample) and a list of key units. Sample weights are numbers that, together with data obtained from sample units, are used in the calculation of estimates of population parameters. Key units are units that are more important than other units for the survey in question, given the expected impact on the final result. Therefore, they are treated differently both in the data collection phase and in the data editing phase.

Quality assurance instructions

If different surveys relate to the same basic target population and the same reference period, the same procedures should be used to establish a sample selection framework as this contributes significantly to greater consistency of statistical results.

Every effort should be made both in the process of identifying data sources and in the production process. Removing inappropriate and duplicate units can make a significant contribution to improving the quality of statistical results.

If the values of the sample selection framework variables are determined from several different data sources, the procedures for prioritising individual data sources must be carefully considered.

The quality of the procedures for developing the sample selection framework and the quality of the data sources used need to be assessed regularly and systematically on the basis of the feedback collected by statistical processes. If large deviations from still acceptable standards are detected, quality improvement procedures need to be implemented.

In the case of a probabilistic sample, stratification should be used to select units (unless there are good reasons not to use it) as this contributes significantly to the representativeness of the sample and at the same time to the greater accuracy of the statistical results. The selected stratification variables should be highly correlated with the surveyed variables or those that are determined by the publication areas. According to sampling theory, the most effective stratification is the one in which the units in the stratum are as similar as possible (with respect to the key variable), and the units from different strata are as different as possible.

In the case of a highly asymmetric distribution of target variables in the population, some strata should by all means be included in the sample (take-all stratum*)*. This is especially recommended for business surveys where large companies must be included in the sample.

In the probabilistic selection of the sample, it is important to use a computer-generated random mechanism. For standard sampling plans, the use of ready-made procedures is recommended. In periodic surveys, the accuracy of the results obtained on the basis of a random sample must be continuously checked in order to meet the required criteria and standards. If it has not been done already, the design of the model should be modified accordingly. As part of the planned changes, the possibility of using a more efficient model design should be explored first. Only if it is determined that sampling procedures cannot improve accuracy should options for increasing the sample size be considered.

It is recommended not to include too many key reporting units (maximum one third of the total sample) if the desired effect is to be achieved by identifying and using these units.

4.2. Preparing for data collection

Description

Preparation for data collection includes planning procedures, preparation for their monitoring, programme development and testing, and training of participants. Each method of data collection has its own specifics in terms of appearance and content of the questionnaire, way of completing the questionnaire, applied procedures and requirements for the reporting unit.

The course and length of preparations for data collection depend on several factors, the most important of which are the periodicity of statistical survey, the method of data collection and data collection techniques. In longitudinal studies with monthly periodicity, the preparation is usually performed once over a longer period, e.g. for a quarter, for half a year or for the whole year, while cross-sectional surveys do not require preparation for a longer collection period. Longitudinal surveys usually become a routine, so they can be more effective, but care must be taken that such a routine does not cause the problem of repetitive errors.

During the **self-completion procedure**, it is necessary to prepare all materials and an address book:

When collecting data using the PAPI method, it is necessary to prepare an appropriate number of copies of documentation (questionnaires, notices, methodology, etc.) depending on how many addresses the address book contains. Letters of notification can contribute to a higher response rate and increase the credibility of the survey on reporting units. Address documents are delivered by mail. At the same time, it is necessary to prepare response recording tools and programmes for manual data entry and/or scanning.

Documents to be returned by reporting agencies must be properly addressed and provided with appropriate identification to track responses. In the case of printed materials, it is desirable that there is a barcode that contains identification on each page of the material.

When collecting data by electronic questionnaire, reporting units are mostly informed by e-mail (and only exceptionally by ordinary mail). This collection method also requires special preparation of the questionnaire, integration with address books and its testing. The entry of preliminary data from the list of codes or registers has been simplified.

Interviews are the most expensive way to collect data. Therefore, it is necessary to plan it in as much detail as possible. In addition to the preparation of documentation as well as for self-completion, everything necessary for the participation of interviewers must be prepared, from the accompanying documentation, distribution of reporting units on the sample by interviewers and the programme of monitoring the work of interviewers. Surveillance of interviewers should be more frequent with this method of data collection than with self-completion. For the purposes of conducting the field survey, workshops are being prepared for interviewers, where training is provided on ways of accessing reporting units, the type of questionnaire and the methodology for filling it in. The workshop also presents practical examples of completing questionnaires and tests the skills and expertise of interviewers.

An address book must be developed for each statistical survey. The address book is used to describe the list of reporting and observed units with addresses and other necessary data. Such a list is used to distribute questionnaires, check the arrival of questionnaires, communicate with reporting units and process data. The source for compiling the address book for individual statistical surveys is the appropriate register: the Statistical Business Register of the Croatian Bureau of Statistics (business surveys), database of censuses of population, households and dwellings (surveys on persons and households) and the Statistical Register of Agricultural Holdings (agricultural surveys). The selection of observation units and reporting units can be determined by using sampling method, by direct selection from the appropriate register based on well-established rules or it can be formed based on the address book for the same survey in the previous period by adding or deleting units. The address book must also list the key units, i.e. the list of units that are the most important for the survey with regard to the impact on the final result and which are treated with higher priority, especially in the phase of data collection and editing.

In periodic statistical surveys, the address book must be updated regularly. Current changes in the register and feedback from statistical surveys should be taken into account. In the process of updating the address book, new units are added as needed and demographic changes are monitored. Detected changes are not adopted automatically, but the change is entered on a case-by-case basis. Regarding address books, rejected mail from the field (unknown, moved or incomplete address) must also be dealt with.

Quality assurance instructions

The method of data collection should be adapted to the objectives and methodology of the survey, and not the other way around. The choice of how to conduct the survey should be tailored to the content. For example, telephone surveys should not be too long and should not contain complex questions with a large number of possible answers. On the other hand, it is irrational to conduct short and simple field survey. It is necessary to agree on a schedule and work instructions. Before printing the final questionnaire, the holder of the survey is obliged to coordinate it with the participants by areas.

For the purposes of subsequent communications, the data of the reporting units must be clearly visible on all materials containing data taken from the said units. Uniformity and visual recognisability of the Croatian Bureau of Statistics are essential in all documents. The deadline for the survey should be determined depending on the sampling method and the duration of the survey.

Methods and tools for data collection need to be thoroughly checked before starting fieldwork. Thorough testing of both rapid entry programmes, as well as the data control programme and the survey laboratory questionnaire, needs to be checked in detail. Testing is conducted by the survey organiser and by the organisational units in which the activities are conducted. Programmes suitable for use are prepared only after the successful completion of testing. The speed of data entry and the skill of the interviewer when handling the phone call should be checked. Laptops for interviewers should be set up in accordance with adopted standards. Data protection against possible theft must also be taken into account.

The implementation documents need to define the dates and persons responsible for the work plan in a timely manner. The address book must contain a unique identification of the reporting and monitoring unit and a complete address. Before adding new units to the address book, the holder of the statistical survey must review the previous address book in order to avoid duplication of reporting units. The observed units are not deleted from the address book. If an individual observed unit is no longer active, it is assigned the appropriate status. If the data are presented from other data sources that can reduce the burden on the unit or improve the quality of reporting, it is necessary to consider whether the relevant data can be entered in the questionnaire before sending.

4.3. Collection of primary data

Description

Data collection is carried out using various data collection instruments. It includes initial contact with data providers and any activities related to further contact or reminder of data providers.

The Croatian Bureau of Statistics collects data in various formats. If received in electronic form, they should be suitable for aggregation in the input database, as opposed to data collected from printed questionnaires, which should first be converted to electronic form for later processing. Converting data to electronic form can be done in two ways: by manual entry or by scanning.

The appropriate method of data entry is determined by the format of the input field and the general format, and the content of the questionnaire. In special cases, the verification of the questionnaire can be done before entering, while in the case of a printed questionnaire, the original data should be kept in electronic form. The scope of control depends mainly on which data entry procedure will be used: scanning or manual data entry. If scanning is used, data records are stored in the appropriate fields and further encrypted.

Necessary programmes and procedures for manual data entry and scanning are prepared in the preparation phase of data collection. At the same time, it is necessary to test the procedures and programmes and then make all the corrections.

When data are collected by using combined reporting methods (e.g. printed, electronic questionnaires, CATI and CAPI questionnaires), or when data entry is performed by using combined procedures (e.g. scanner and manual entry), or part of the data is taken over from an administrative data source, then the data are combined into a single database. Furthermore, the data can be checked, edited and later compared with the population distribution in the database. This makes it easier to determine the reliability of the results.

Quality assurance instructions

The survey holder must provide clear and unambiguous instructions for checking the data before entering. The questionnaire must be comprehensible, transparent and properly designed, and reviewed during the preparation of the data set.

The data entry programme must be prepared accurately and in a timely manner. Prior to practical application, it must be thoroughly tested to avoid errors in the source data.

In order to reduce the number of errors when digitising data, electronic questionnaires are preferred when collecting data. When collecting data in printed form, the use of scanning is more appropriate.

If scanning or fast data entry without repetition are used, the questionnaire should be designed so that internal controls can be performed (e.g. a separate field for total values). The data on the questionnaires that are read on the scanner must be written clearly and legibly. Optical checks should be avoided before recording data from printed questionnaires.

Still, the data must be checked even after scanning because the scanner may misread the data from the form.

Example of organising a scanning process – Census 2011

1. Head of shift management

Within the obtained schedule, and at the request of the scanning manager, he is in charge of scanning a specific city/municipality. He records the task in his own records and deliver him a "Scanning Order".

2. Scanning manager

He takes care of the timely supply of each scanner with scanning material (each scanner works independently, i.e. scans a different city/municipality). According to the obtained "Scanning Order", he fills in the "Order for Retrieval of Material from the Archives", in which he obligatorily states the name of the scanner for which the material is "picked up" next to the name of the manager and hands it over to the archives manager.

3. Archives manager

He takes over the "Order for Retrieval of Material from the Archives" and checks in the records whether the complete requested material has been stored. He prints the "Storage Status List" of the requested city/municipality and hands it over to the scanning manager. He then instructs his employees to place the required material in the ascending order of the PK form on the marked shelves of the scanner, which is indicated in the "Order for Retrieval of Material from the Archives". Finally, he records the material taken out of the archive in his own records, and keeps the "Order for Retrieving Material from the Archives".

4. Scanning Manager

After the requested census material is placed on the shelves, he checks its completeness according to the "Storage Status List". After that, he determines the start of a city/municipality scan.

5. Preparation group

They take the boxes of PK forms from the shelf in due order and place them on a (specially supplied) preparation table (there are always four boxes with PK forms on the table at the same time). They open the boxes and check if the P-0 form is there, the identification of the municipality and the census district. After that, they carry the material that is ready for scanning and place it on the scanner inlet table. From the boxes with PK forms, they separate the control books and arrange them in ascending order in a separate folder for each municipality, and write, using the black marker, the identification (county, city/municipality, MB) as well as the range of ordinal numbers in PK forms (e.g. 1 – 0100). Up to 100 control books are inserted in each folder (depending on the size of the city/municipality), taking into account that all multi-part control books are inserted in the same folder. Also, sketches and descriptions of PK forms are inserted in separate folders with the code entered in the upper right corner of the PK forms (according to P-8). For sketches and descriptions that are missing, the "Record of Missing Sketches" is filled in, which is placed at the top of the folder.

6. Scanner operators

They perform scanning according to the procedure, paying attention to the order in which they stack the material at the exit. Upon completion of each individual PK form, they verify the number of documents read.

7. Preparation group

After scanning each PK form, they record the completion by rounding off the ordinal number in the PK forms in the "Storage Status List". They check whether the number of scanned documents is identical to the number entered on the P-0 form. In case of any difference, they enter the number of forms in the "Storage Status List" according to P-0, the number of scanned forms and the difference in relation to the scanner. The material is returned to the corresponding box with PK forms. On the box, they mark the end of the scan by writing an "S" mark with a red marker under the title 2011 Census. The box is returned to the same place on the shelf.

8. Scanning Manager

He informs the head of shift management about the completion of the city/municipality scan and fills in the "Order for Moving the Material into the Archives", which he hands over to the archive manager.

9. Archives manager

He acts in accordance with the obtained "Order for Moving the Material into the Archives" and instructs his employees to move the required material from the shelves into the archives. He also downloads and archives folders with separate control books. He records the note on the return of the material in his own records, and keeps the order.

There will be minor modifications to this process, depending on the operation of the data processing centre.

Data collection and communication with reporting units depend on the method of filling in, the target population, the observed phenomenon and the available resources. The time spent and the non-response of the reporting unit are, as a rule, inversely proportional. When collecting data, special attention should be paid as little data as possible are lost due to human or systemic errors of the interviewers or errors in data processing.

A high response rate of reporting units can be achieved by setting clear objectives, using proper tools and choosing the convenient way of collecting data. The consequence of conducting survey at unfavourable times for reporting units, e.g. during holidays or higher seasonal loads (preparation of annual financial statements), is a lower response rate. The response rate can also be influenced by the type of survey, e.g. health surveys or accounting surveys have a higher response rate, while housing, culture and innovation surveys usually have a lower response rate.

Non-response can be divided into two main categories: item non-response and unit non-response. We talk about the item non-response when some questions in the questionnaire, which is otherwise filled in, are left without an answer (the data vector is incomplete). This situation can occur when the respondent does not have enough knowledge to enter the answer to the question or it is a sensitive question that the respondent does

not want to answer, and it may be a contradictory answer that does not comply with data control rules. Unit non-response is mentioned when data for all variables are missing. The highest non-response rate occurs when the questionnaire is self-completed. Therefore, in sample surveys on households and natural persons, data are usually collected using the interview technique because interviewers know how to motivate people and thus reduce the non-response rate.

The duration of data collection also depends on how the questionnaire is completed. A telephone survey is usually faster than a face-to-face or postal survey. In postal surveys, it is difficult to estimate the duration of the collection process because it depends on the number of reminders delivered to those who do not respond and, above all, on the reaction of reporting units. The more time and data sources are spent, the higher the response rate will be. There is an opinion that in one survey only two of the following three requirements can be satisfied: lower cost, high response rate and little time spent. Monitoring of the reasons for non-response is carried out more often and in more detail in interviews than in self-completion.

In self-completion, the reporting unit interprets the questions and instructions by its own and provides answers in accordance with the understanding of the submitted instructions. This is why the data often contain many more errors than when other data collection methods are used: this method of data collection later requires more work on data editing procedures. Completed questionnaires or comments from reporting units can be obtained by regular mail, e-mail, electronic reporting or telephone. Records of responses and non-responses are monitored regularly. Reporting units that did not submit a response to the questionnaire are provided with a maximum of two written reminders. Sometimes it is necessary to conduct a telephone conversation with key reporting units and inform the reporting unit on the necessity and purpose-serving quality of filling in the questionnaire.

When data are collected by the interview method, the interviewer fills in a questionnaire on behalf of the reporting unit. He knows the questions and what answers are expected in the field, and in accordance with expectations, he fills in a statistical questionnaire. Since the interviewer is familiar with the questionnaire, he can fill in the questionnaire without any interruptions. Communication between the Croatian Bureau of Statistics and interviewers is carried out in the way similar to those used in the self-completion of questionnaires. The Croatian Bureau of Statistics takes care that an adequate number of interviewers is always available and that possible cancellations can be compensated as soon as possible. Each non-response is continuously monitored by sending interviewers to the field within the agreed deadlines. If the interviewer does not achieve a plan for a sufficient number of completed questionnaires, or if the completed questionnaires do not meet the appropriate quality standards, or if methodological errors are detected, first he will be warned and then the process of checking the questionnaire goes on. If errors or serious breaches are repeated, the interviewer is excluded from the data collection process and replaced by another interviewer.

During the telephone survey, the controller is in charge of continuously monitoring the work of the interviewer. He monitors the progress of the survey, its quality, warns the interviewers about errors and possible insufficient quality, and takes care of breaks during the work. Teaching interviewers, following and distributing additional instructions is simpler than other data collection methods. It is common practice for each new interviewer to be in a contact with a survey holder who is able to answer methodological questions from the survey.

Organisation and field work need to be appropriately adapted to the target population. Timetables, information letters, survey brochures, instructions, methods and tools for data collection, technical devices and other tools need to be carefully planned in order to reduce the reporting burden for reporting units.

The instructions written in the implementation documentation are followed when collecting and communicating with reporting units. The deadlines set in the work plan must be respected during the work. In communication with reporting units and interviewers, it is necessary to respect the established rules of the Croatian Bureau of Statistics. Problems need to be solved quickly and patiently.

For the interviewers' work related to data collection and communication with the reporting units, it is very important that the servers work flawlessly and smoothly (seven days a week, 24 hours a day). It is necessary

to constantly strive to find reporting units with which the Croatian Bureau of Statistics has not established any contact. Appropriate rejection reduction procedures should also be used. The number of rejections must be kept to a minimum. Segments of the population with a higher non-response rate need to be identified and the intensity of contacts with them increased. If the observed variables are markedly asymmetrically distributed in the population, it is necessary to ensure the response of the most dominant units.

Units need to report a minimum set of data so that may still be considered a complete answer. It is necessary to take care that reporting units do not bear any costs for statistical reporting purposes, especially when surveying persons or households. For example, the questionnaire must be accompanied by an envelope with postage paid, which must be completed and returned by post to the Croatian Bureau of Statistics.

The non-response rate to the questionnaire can also be reduced by using symbolic gifts. However, in doing so, one should act rationally in accordance with the principle of sound financial management.

Such an approach is used in the Croatian Bureau of Statistics only in exceptional cases, bearing in mind that if reporting units use gifts irrationally in the long run, it can have a negative effect. With the postal method, it is necessary to closely monitor whether the right reporting unit has received the right statistical questionnaire. If the reporting unit did not receive a particular questionnaire, the same questionnaire must be resubmitted.

The situation regarding the receipt of questionnaires should be regularly reported to the survey holder. There must be systematic control of the process, i.e. systematic monitoring of progress in data collection and control of data for which no answer to the question has been received. The aim is to achieve a reasonable ratio between the quality of data and the cost of organising and conducting a statistical survey. Important quality indicators are the following: response rate, processing errors, required number of completed forms submitted and non-response rate depending on the reasons for rejection.

While preparing appropriate procedures, the following facts must also be taken into account: the period of data collection, the duration of data collection and the expected workload for all participants, such as the length of the questionnaire, the complexity of the content of the questionnaire, the periodicity of the survey and the like.

In methodological sense, it is necessary to determine the maximum number of contacts with the reporting unit. Reporting units are usually provided with a maximum of two written reminders. For key units, the second or third reminder can also be in the form of a telephone call.

For key reporting units, answers to the questions asked are required to be collected. If key units do not provide any answers, they are asked to provide written reasons for not being able to provide data. For key units, response dates should be entered into records or displayed in the appropriate status.

At the beginning of the survey, the holder of the statistical survey should be involved in supervising the work of the interviewers. The Croatian Bureau of Statistics use modern technologies for transmission of field data. By establishing regular supervision of interviewers, better quality of business processes is achieved. Since the telephone book is the basic tool for the smooth running of work in the survey lab, it needs to be updated regularly.

For the field work of interviewers, it is optimal that laptops that are taken over to perform field work have the same functional and technical characteristics.

In order to facilitate and standardise the conduct of surveys and obtain human, financial and information resources needed, it is necessary to plan in advance the schedule of surveys to be conducted next year.

4.4. Retrieval of data from administrative and other secondary sources

Description

The download of administrative data is done on the basis of concluded agreements between institutions, while the transfer takes place through various media (optical media, network protocols, registry replication, etc.). When downloading data, the structure and format of data are controlled. This activity includes planning the use of administrative data sources, downloading data from other institutions for various purposes and organising cooperation with other institutions as well as transferring a certain phase of survey to another institution.

In this document, we use the term administrative data source for all data sources not collected in a statistical survey. These are data collected on the basis of administrative regulations and other secondary data sources, and can be used for statistical purposes as an auxiliary tool for conducting classical statistical surveys, as a primary data source or to supplement, edit and control data collected in statistical surveys. The advantage of using administrative data sources is that already available data are used, so there is no need to further burden the reporting units. In addition, the use of administrative data sources is cheaper compared to the classical implementation of statistical surveys. Given that administrative data sources are established for non-statistical purposes, it is necessary to focus on their shortcomings, such as the application of different definitions, lack of control over data quality and timeliness of data. The use of an administrative data source for data collection activity does not mean that it completely replaces all statistical procedures that take place in a classical statistical survey, so these types of activities relate to data editing, linking records, harmonisation with statistical identifiers, etc.

For the purposes of conducting some statistical surveys, the Croatian Bureau of Statistics cooperates with other institutions, whether for organisational or technical reasons. If the whole business process is observed, it may happen that institutions are involved to conduct a certain phase of survey, e.g. for electronic data collection.

When administrative data sources are used, a cooperation agreement needs to be concluded. The purpose of the data collection agreement is to determine the content of the administrative data source, the method of retrieval and the terms and conditions of use; the purpose of concluding a cooperation agreement and the division of tasks between the signatories of the agreement.

Depending on the type of data and the owner of the administrative data source from which the data are taken over, an agreement is concluded and a protocol for data exchange is agreed between the owner of the administrative data source and the user of the registry service interface (for example, SGA – Central Register of Spatial Units), or an agreement is signed and the Catalogue of Mutual Demands for Statistical Data from records (for example, Tax Administration). In the case of taking over data from the Financial Agency on an annual basis, the Protocol on the Submission and Taking Over of the Database from the financial statements is signed.

The collection of data in statistical surveys seeks to reduce the non-response rate as much as possible. In surveys where data are obtained from secondary data sources (administrative registers, records, etc.) the problem of non-response is usually less. However, other problems arise, especially those caused by methodological inconsistencies with statistical concepts.

Records of secondary or administrative data sources should be kept in one place, so that, in the case of additional needs, existing data sources can be checked first. The records must be kept regularly and continuously and the date of the last data collection must be recorded. If the desired data source has not yet been downloaded, depending on the content of the administrative data source, it is necessary to conclude an agreement, technical protocol and/or define a data catalogue.

Data collection is carried out in accordance with the provisions of the agreement and the technical protocol. Prior to the first retrieval of data, it is necessary to prepare and verify instructions, procedures and programmes for retrieving data and conducting formal checks, such as legibility of data, adequacy of records and their number, and verification of characters. The usual way to download data includes that the Croatian Bureau of Statistics is notified that the data have been prepared. This is followed by acceptance and technical verification of data and submission of data to the agreed location.

Data transfer can be organised using a variety of storage media, such as CDs, DVDs, removable disks, USB sticks, FTP protocols and the like. In the case of irregularities in the acceptance of errors in the data structure, feedback is provided to the data source.

Administrative data sources may, when used as the primary data source, contain auxiliary variables with errors. To avoid undesirable situations, it is necessary, as a precaution, to check the administrative data source before using the data or to detect errors during use and eliminate them in cooperation with the owner of the administrative data sources. Editing administrative data sources is especially important when connecting multiple data sources. In order to make integration as easy and efficient as possible, all resources used must first be edited to determine if there are any errors that could prevent data integration.

The most common types of checks performed are checking the values associated with the code lists, checking the range of values for numerical variables (e.g. negative values), checking the estimated number of records and checking the suitability of the reference periods to which the data relate. At this point, the correctness of the relationships between several variables in the same source can be checked. Most consistency checks are performed later in the statistical editing process, when the compatibility of variables from different sources, administrative and statistical data sources, is also checked.

Quality assurance instructions

All data sources upon receipt must be clearly and accurately documented as this will facilitate work at a later stage of the use of administrative data sources in statistical surveys.

For each variable associated with the code list, it is necessary to determine exactly which version of the code list is appropriate for data processing and dissemination. For all numerical variables, it is necessary to check the characteristics of the data and then determine the set of acceptable values. In most cases, it is necessary to decide whether zero or even negative values will be allowed.

All meaningfulness checks for individual variables must be carefully documented. If the value is limited for certain variables, the reason for such "limitation" must be stated. Feedback on detected errors should be provided to the database administrator if the error was identified solely during the data source check.

In the case of errors, procedures should be defined for their systematic resolution in a rational manner. The technical solution should be as general, simple and detailed as possible.

An agreement and/or catalogue of demands must be prepared and signed before data collection can begin. The following items should be listed in the catalogue of demands: the receiving office in the Croatian Bureau of Statistics, the person in charge of receiving in the Croatian Bureau of Statistics, the user office in the Croatian Bureau of Statistics, the responsible person for using the data file in the Croatian Bureau of Statistics, the responsible person of the data source owner and the responsible person – IT support of the holder of the administrative data source. All participants in the process must be familiar with the content of the data before the first download. The holder of the administrative data source must inform the data user in a timely manner and within the agreed deadline that the data are ready. The entry of data on downloads in the data record should take care that users are notified in a timely manner about changes, possible delays and other emergencies.

In order to reduce administrative barriers and simplify business, efforts are made to use administrative data sources as much as possible. Assuming that administrative data sources are of satisfactory quality, this method of data collection has an advantage over others. Although administrative data sources have advantages, one should take into account the disadvantages as well. All shortcomings need to be minimised and the administrative source needs to be adapted to statistical needs as there are differences in coverage, definitions, timeliness of data, quality of data sources, etc. Since legal provisions are the basis for establishing an administrative data source, it is necessary to thoroughly study the content and methodology in order to use that source for statistical purposes. The establishment of continuous cooperation with administrative data owners is important both in the first download of data and in the implementation of all content changes in administrative data. The use of administrative data sources is sometimes more demanding than the implementation of the usual statistical survey because it involves a different approach to data processing and editing. Pursuant to Article 50 of the Official Statistics Act (Official Gazette, No. 25/20), holders of administrative data sources and of data collected by using the observation and monitoring method are obliged to inform the Croatian Bureau of Statistics in a timely manner that they are going to introduce either collection of administrative data or data collected by the observation and monitoring method.

When opportunities allow, the Croatian Bureau of Statistics should participate in the preparation and introduction of a new administrative data source. When downloading data from an administrative source, it is necessary to know in detail the content of the data that will be downloaded by the Croatian Bureau of Statistics. Confidential data needs to be protected by switching to statistical identifiers and controlled allowing access to data. Depending on the content of the data, access to this type of data is managed by the data administrator. Before carrying out statistical data processing, the data must be adjusted to statistical needs and their quality must be checked.

4.5. Enter the collected data

Description

It involves entering the collected data and metadata into the process for further processing. It may include manual entry of data from printed questionnaires or using OCR, automatic download of data via a network questionnaire and applications (CAWI, CAPI and CATI), web scrapping or conversion of data files received from other institutions into another format.

Preparations for automatic encryption

If a text is entered on the questionnaire that represents the value of a feature, it is usually necessary to translate it into a numerical designation (code) of the corresponding classification. This process is called encryption. It is possible to do it "manually" in its entirety, in a way that a trained group of people reads the typed texts, look through (or knows) the classification and then enters the codes that they assume are appropriate.

Another way is to perform the process of reading the text, going through the classification and determining the appropriate code by using a computer. In that case, it is necessary to develop an algorithm that performs these tasks instead of people and write an appropriate programme. It is common for such applications to provide fully automatic encryption when the probability of a correct answer is very high, but also to provide the possibility of final choice to the expert if the probability of one choice is insufficient and/or equal to the probability of choosing another code. The second case is called computer-assisted encryption.

All automatic encryption programmes have some definition of similarity (proximity) and try to determine which of the classification texts is more similar to the typed text for which the code is requested. With small classifications (containing one word) it is very simple, but with complex and large classifications it is a rather sophisticated procedure. The quality of the obtained code, the recognition efficiency and the speed of operation are features that measure the overall performance of such algorithms and programmes. The inputs to

encryption programmes are, on the one hand, the texts for which the code needs to be determined, while, on the other hand, they are extended classifications of the features to be encrypted. For this purpose, the classifications have been expanded with new lines in which, in addition to the already existing codes, the expected answers are entered, which differ from the existing, official ones. This results in taking over classifications with multiple names with the same code. Such extended classifications are called thesauri.

Thesauri are extremely important for the overall success of automatic encryption and therefore need to be well expanded before starting work. Of course, you should only enter the really expected answers, and not go to extremes and enter extremely rare and little expected texts. This reduces both speed and recognition efficiency. It is desirable to fill and upgrade thesauri during the process of automatic encryption, as well as on the basis on the results of such work. If it is noticed that a text appears more than once and is not included in the thesaurus, it is necessary to include it. A quality automatic encryption application must also support a feature that is commonly called learning. In addition, by analysing the assigned codes, it can be noticed that some of the previously assigned unofficial values are not used at all, so they need to be removed from the thesaurus.

It is desirable to expand the thesauri with tests that include some expected abbreviations, synonyms, old names that are still used, descriptions that belong to a code, but are not listed in the official text, and the like. Of course, thesauri need to be entered on a computer in the appropriate format, where they are later maintained, supplemented and corrected.

Optical readers are used so that a large amount of data collected on questionnaires in printed form can be quickly read and transferred to digital media. The basic characteristics of optical readers are speed of operation and level of reliability. The speed of operation depends directly on the speed of loading the paper into the machine and on the average speed of reading characters. Some of these parameters may be affected:

- the number of questions and signs to read
- the level of reliability required for each character separately
- legibility of handwritten characters with respect to the desired standard
- number and complexity of controls
- the skill of the operator serving the optical reader (OCR).

Requiring a higher level of reliability also provides better reading quality, but at the same time slows down reading. In addition, signs that are not recognised, i.e. those that do not meet the level of reliability, are later sent to the operator for correction, and these corrections can seriously slow down the whole process. It is very important to maintain the same speed of the OCR and the corresponding manual correction.

Computer-assisted web interviewing (CAWI) is part of a methodology based on a questionnaire delivered to a respondent through a link in a panel or on a website. Compared to other methods of data collection, it is the most economical way of data collection because no interviewers, devices or additional tools are needed. Therefore, online surveys, also called online questionnaires, are one of the most common methods of data collection. When the Croatian Bureau of Statistics decides to apply the method of data collection using the CAWI method, all activities are focused on the design of the questionnaire because the response rate is directly related to the quality of the questionnaire.

The CAWI method may be appropriate for several reasons. If the questionnaire is well designed, it will automatically manage questions using logically set conditions, such as a display or a skip. The questionnaire can be structured to facilitate understanding and increase the response rate. Instructions can be inserted to help the respondent understand the questions asked and complete the online questionnaire. This method of data collection reduces the cost of interviewers, the purchase of electronic devices and the time required for data analysis because this data are available in real time in the administrator's database. There is also software that allows the preparation of real-time reports on online surveys and statistical data collected in this way. The main disadvantage of online surveys is the lack of interviewers, which in some cases are very useful because they can help and guide respondents by filling in questionnaires for a particular statistical survey. Other shortcomings that need to be taken into account are that not all respondents have an internet connection and it may happen that the online survey lower the attention capacity of respondents, thus increasing the chances of obtaining poor quality or incomplete statistical data.

Computer-assisted personal interviewing (CAPI) is a method of data collection in which the interviewer uses a tablet, mobile phone or computer to record the answers given during the interview.

The advantages of the CAPI method are that they make it easier to check logical settings, skips and checks of the meaningfulness during the interview. These procedures make the survey more efficient and help ensure a higher level of data quality. They also save time when it comes to data editing. CAPI is a good tool for tracking interviewers. It can automatically record start time, end time and GPS location of the interview, making it easier for controllers to verify that the interviewer actually conducted a particular interview, comparing their time and GPS data with the times of other interviews attended by a person in charge of process supervison and control. It should be taken into account that in some places there is a bad signal or no signal at all, and that in such cases the computer cannot read the GPS location. Due to the strong connection to the Wi-Fi network or data, the data collected in CAPI surveys are immediately transferred to electronic format. This allows the controllers located in the regional units to check a certain part of the data according to the instructions given to them by the methodologist. The shortcoming of the CAPI survey may relate to hard-to-reach areas because the CAPI relies mainly on electricity availability or data connectivity, which may not be feasible in some parts of the Republic of Croatia. If the data collected by the CAPI method are damaged before they are transferred to the server, they will be lost forever. Batteries and offline data storage can certainly facilitate the conduct of the CAPI survey in regions without available electricity or data connectivity. The interviewer is instructed that before going to such places he must check if the battery on the laptop is charged and after returning to the area where he has an internet connection he should send the data. Otherwise, all data are stored locally first and the interviewer must start sending/downloading the data himself. It has not been set as an automatic function yet. It is not easy to conduct an interview with some respondents due to their distrust, and this can result in unreliable data or in that a certain set of such respondents refuses to participate in the statistical survey. CAPI surveys require technologically competent interviewers or longer training periods for interviewers. This can be a problem with emergency or last-minute surveys. If an interviewer withdraws from the survey, his surveys are forwarded to another interviewer until a replacement is found.

Computer-assisted telephone interviewing (CATI) is a telephone data collection. The interviewer calls the number shown on the screen. Then, if the respondent wants to join the survey, the interviewer follows the script on the screen choosing the answers. The CATI tool will automatically move on to the next question following the logic of the questionnaire. At the end of the questionnaire, the interviewer displays the new respondent for a call. For example, the CATI system offers the following number that the interviewer can call.

The system manages contacts according to the rules set by the administrators. For example, busy contacts will be redialled after 15 minutes. The application can be set to intervals that suit the survey organiser. It is currently set up so that it can call a busy line up to six times in one day. Usually, at the beginning of the day, the person in charge of supervising and controlling the process selects the number of surveys he would like to do with the available interviewers, the so-called day batch system selects the default number of interviews from the loaded sample. If the contact answers the call but is not available to talk, the interviewer can also schedule an appointment. In this case, the CATI software will automatically re-display that contact on the scheduled date and time.

The advantages of the CATI method are obtaining accurate answers and reliable information about respondents, achieving surveys of almost the entire population, precise sample management and stability of answers, easy calculation of time, avoiding out-of-quota interviews, full control of survey progress, real-time feedback, presence of interviewers discourages withdrawal from the interview, interviews can last longer because the withdrawal rate is lower, full satisfaction of respondents is provided, sensitive topics are well accepted, a telephone interviewer can help understand the questionnaire and there are no non-responses. The respondent may answer certain questions with "I don't know" or "may refuse to answer".

Disadvantages of the CATI method are high costs for telephone interviewers, costs of PC stations (the cost is stated in the initial establishment of the CATI center), costs of counters (software counts calls and assigns statuses to them), software and hardware, the number of interviewers is directly proportional to the daily number of interviews. The disadvantage is the price of telephone lines and the difficulty of hiring new interviewers who then need to be trained to work in the CATI center and in every conducted survey.

Quality assurance instructions

In optical reading, the level of reliability implies that the probability of reading is one hundred percent, i.e. that one character is read correctly. The demand for greater reliability allows for better reading quality, but at the same time slows down reading. In addition, signs that are not recognised, i.e. those that have not satisfied the level of reliability, are subsequently sent to the operator for correction, and these corrections can seriously slow down the whole process. The level of reliability is set very high for all identification questions and for the first letter in each word. The lowest level is more demanding for all other numerical answers. Under the set conditions, the mean reading speed is calculated for the averagely completed questionnaire of each type.

The assessment of optical reading and the assessment of the combined impact of optical reading and automatic encryption are carried out on a sample of census data. A sample with a fraction of 0.006 was used to check data entry. Different sample ratios were applied, allowing an appropriate ratio for each type of sample unit within the population from which the sample is taken.

The sample is planned for evaluation of:

- deviations of the census material read by the optical reader from the one entered manually
- deviations of the census material read by the optical reader from the correctly or accurately entered one.

The first type of evaluation allows the evaluation or comparison of optical reading in relation to today's available technology used for data entry. The second rating allows measuring of reliability and assessment of meaningfulness. Any difference between the "input" characters and the corresponding characters from the "reading" set is marked as "suspicious" on a separate report. This difference could also have occurred due to an error in manual entry, which is why a special group created an "authentic" set of data and went through questionnaires that were marked as "suspicious". It is important to point out that the members of the group could not see the optically read sample data. This group checked for any suspicious signs and, if necessary, corrected them to match those on the questionnaire. The comparison and correction process was repeated until all signs marked as "suspicious" matched those in the questionnaire. The end result of this process was a set of accurately entered data that could be called an “authentic” sample.

After extensive work in obtaining an "authentic" sample, various discrepancies were noticed, i.e. deviations between the authentic sample and the one that was called "reading":

- - position-level difference (simply listing cases where the optical reader has recognised one thing, while something else is written at that point in the questionnaire)
- - field-level difference (where a field is a set of characters that correspond to an answer to a particular question).

The analysis of automatic encryption showed that it was performed at the expected high level of performance. The high quality of the automatic coding algorithm resulted in a high percentage of automatic variable encryption.

Statistical surveys carried out by using the CAWI method enable monitoring of data quality and reliability. While the survey is ongoing, it is advisable to check completed and partially completed questionnaires. This can be used to check whether complex (or long) questions can cause a higher withdrawal rate or to lead respondents to fill in a questionnaire hastily and incoherently. The quality of the questionnaire can also be assessed during the data analysis process. The level of satisfaction of the respondents is also important information. Inserting an assessment question at the end of the interview or comparing the number of completed questionnaires with

the number of respondents who did not complete the questionnaire allows monitoring and eliminating the shortcomings of the questionnaire, which is crucial for obtaining reliable data for high quality analysis. In order to establish effective web survey, it is necessary to establish a system that will enable the achievement of excellent results at low cost.

In statistical surveys conducted using the CAPI method, data quality is monitored throughout the data collection period. The quality information collected also includes data related to other accompanying processes, such as travel time, time spent on site and number of interviews conducted/in progress/rejected.

The CAPI data quality report was prepared and reviewed. The report included, for example, for each interviewer the number of interviews conducted, incomplete data by statistical survey items, distortions or cases where survey responses differed significantly for one interviewer compared to other interviewers within a particular spatial unit.

Statistical surveys conducted using the CATI method enable rapid coverage of the total target population, as well as individual approach to respondents and adaptability. The data collected by this method are constantly checked, which created the preconditions for the appropriate level of quality to be maintained and improved in accordance with the set statistical standards.

In order to determine the level of interview quality, the application of the CATI method has advantages because it allows monitoring of interviewers and, if necessary, allows for improvements in the business process. The CATI monitoring method provides useful information regarding the interviewer, operational management, training, and the development and design of data collection instruments.

5. Data processing

The data that are the basis for the calculation of statistical estimates in the implementation of statistical surveys can be downloaded in different ways or from different data sources. According to the method of data collection, they can be divided into primary and secondary sources. Primary sources are data collected primarily for the purpose of statistical survey, while secondary data are those that are primarily collected for some other purposes, such as administrative, and the Croatian Bureau of Statistics has taken them over for statistical purposes. If the survey uses data from several different data sources, they first need to be properly linked and then correctly statistically processed.

Statistical data processing refers to all procedures by which data collected in a statistical survey are processed and edited in a form suitable for publication and dissemination, applying appropriate statistical methodologies. The procedure also includes the processing of data at the micro level, such as the removal of errors and incomplete values, mathematical procedures on the population (weighting, aggregation) and the processing of already collected data, such as seasonal adjustment and data protection. In the course of statistical data processing, more or less advanced or demanding statistical methods are applied, the purpose of which is primarily to enable the calculation of accurate and unbiased statistical data. All available data sources should be systematically connected. Errors in the data should be identified and corrected. After that, it is necessary to define and apply an appropriate statistical model and define a mathematical procedure for estimating the population.

Thanks to the development of statistical methods and advanced information technologies, statistical data processing has improved the most out of all statistical survey implementation processes as it has become faster and more efficient. Automated subprocesses have been established, which enable fast and efficient procedures. Although automated procedures have their advantages, they also have disadvantages in the form of so-called "black boxes", when procedures are conducted without any possibility of the influence of statisticians. In order to mitigate the negative effects of such implementation, the processing of statistical data should be carefully planned and documented in detail. First of all, it should be possible to prepare the necessary metadata processes, such as providing the statistician with insight into automated procedures.

5.1. Integrate collected data

Description

It involves the integration of data from one or more sources during data processing. In statistical surveys using several different data sources, the same data sources should be combined and prepared in a format that allows for further statistical processing. If there are unique identifiers in the data sources, data aggregation is performed directly through them. Otherwise, it is necessary to define and prepare the combination of records through other parameters, the so-called indirect connection. In order to simplify the procedures, it is assumed that in the process of aggregation there is always a reference set of observation units that are also the starting point for integration processes. The data set resulting from the integration process should contain all units from the reference data source.

Direct connection. Linking records from different data sources takes place using unique identifiers (e.g. OIB, ID or IDS of the company), which are an integral part of both data sources. Records from an associated source that could not be linked to the reference data source using a unique identifier are registered in a separate list. The process of direct connection is also shown graphically.

Indirect connection. If there is no common unique identifier in the reference and associated data source, the connection is established by indirect linking procedures. It is done with the aid of other selected common variables that exist in both sources. The units of the reference data source in this case are "divided" into two parts.

The first part consists of units that are uniformly connected to the source using indirect connectors. For these units, the values of the indirect connectors in both sources also match exactly in the associated source, so there is only one such record.

In the second part, there are units that could not be uniquely associated with the source, since each record from the reference source is associated with multiple records from the associated source. In this case, the records in which the values of the indirect connectors in both sources are most similar are associated.

Similarity needs to be defined uniformly. It is common to use one of the already existing functions in different programming environments for this purpose for the numerical estimation of array matching.

Only one record needs to be selected for further statistical processing in the second data set. Whether this selection will be made manually or programmatically based on an agreed algorithm must be decided for each survey separately.

By combining different data sources, different values for the same variable are taken over. In such cases, it is necessary to determine the priority of data collection and add a new variable with an indicator for that variable, which for each unit shows which data source is determined as a priority in defining the reliable value of the variable.

Instructions for quality assurance

Before determining the procedure, the hodler of the survey should carefully examine the available administrative data sources and their applicability in a particular statistical survey. When determining which variables will be taken over from administrative data sources, it is necessary to examine whether the variables meet the statistical criteria set in the data source and whether data from administrative data sources will be available in a timely manner, i.e. early enough to publish and disseminate statistical results.

If several different data sources are used for the same variable, it is necessary to determine whether the data from these different sources are mutually coherent. The sampling methodologist is responsible for setting up the data model. However, the general methodologist and the holder of the statistical survey should be involved in the entire business process, because only in this way can the application of the model in the implementation of the statistical process be ensured.

If data from data sources that are sensitive in the sense of personal data are used, procedures must be organised to prevent abuse of access to such data as much as possible. Therefore, before the data integration process, the unique administrative identifiers must be replaced by alternative statistical identifiers, which are later used to carry out the statistical procedure.

If indirect linking procedures are used, they should be set up to meet full coverage as well as relevance and accuracy requirements. In the first place, manual procedures should be avoided, as this could jeopardize the timeliness and punctuality for publishing statistical results.

At the end of the data aggregation procedure, each observation unit in the database should have the corresponding unit status and each variable should have the corresponding variable status. If the values of a particular variable are derived from several different sources, care must be taken for each value from which the data source is taken.

For each variable associated with the code list, it is necessary to determine exactly which version of the code list is appropriate according to the validity of the data. This version of the code list is then used to control the value of the variable. For all numerical variables it is necessary to carefully examine their characteristics and then determine the set of acceptable values. In most cases, it is necessary to decide whether zero or even negative values are allowed.

All data verification used in this part of the procedure need to be documented. If the set value of certain variables is limited, it is necessary to describe the reasons for the set limit. The set data verification should be documented in detail.

5.2. Controlling, editing and correcting data

Description

Control includes data verification according to given check rules, check of aggregated data or group of units, extreme values, outliers and critical values. Editing and correcting involves automatically editing the data or activating a warning that you need to manually review and correct the data. Check, editing and correction of data can be performed repeatedly until the data reach a satisfactory level of quality. Depending on the level of applied data, data editing procedures can be roughly divided into micro-level data editing and macro-level data editing. When editing data at the micro level, procedures are performed at the level of data of individual units. There are also several procedures for editing micro-level data, depending on how the data were obtained.

When conducting a field or telephone survey, a complete or partial set of logical checks is applied. Upon completion of the survey, the data are combined with the available secondary data sources into a common file. This is followed by additional verification of the data, and all identified errors are corrected by manual or automatic correction.

In the case of self-completion (printed or electronic questionnaire), logical checks are performed after the data entry is completed. Checks of the entire data set are performed simultaneously, by automatic corrections and by generating error statistics. Detected errors are corrected by manual or automatic correction procedures.

If a combination of several different data sources (e.g. statistical and administrative) is used in the survey, each data source should first be edited separately. The combined data set must then be checked. It should be emphasised that errors detected in the consolidated data file need to be eliminated on each individual data source (and not in the consolidated file) because this is the only way to achieve long-term sustainability in the next data collection and processing.

Since data editing is a time-consuming process, it needs to be automated as much as possible. Computer programmes should be designed to identify and correct errors at the same time.

Automatic corrections can be roughly divided into two groups: deterministic and probabilistic. The first corrections are determined by simple deterministic rules that can be written in the form of logical and arithmetic expressions (e.g. IF X > 10 AND Y <> 1 THEN Y = 1). Other corrections are determined on the basis of probabilistic procedures, usually on the basis of a minimal change in the reported data (Fellegi-Holt approach[8]).

Instructions for achieving quality

Logical checks must be defined and tested in a pilot programme. The practice of setting up a large number of checks, with too strict criteria, in order to "purify" the data collected in the survey as much as possible should be avoided. Such a practice leads to the so-called overvaluation, and the consequence of such regulation is that we get too many units whose values need to be rechecked.

It is necessary to investigate whether automatic data editing procedures can be introduced, as this could significantly reduce costs, shorten the time of statistical survey and increase efficiency. Before such procedures are introduced into the regular procedure, it should be examined in detail whether they are meaningful and feasible in terms of content and technology.

[8] It refers to the assumptions and editing and imputation goals set out by Fellegi and Holth in their 1976 paper in the Journal of American Statistical Association.

All data editing procedures must be documented in detail and properly. In periodic surveys, it is necessary to enable the results of data editing to be used to improve quality. Data editing should be organised in such a way as to enable the sustainability of procedures. In practical terms, this means that data editing procedures using the same input data must be provided and that the same methods always give the same end result.

A "history" should be kept of all changes to the data. In other words, the original data must not be substituted with corrections during the editing phase. However, a new version of the data should be prepared if the data are corrected. Each corrected piece of data should contain associated metadata that provide information about where the data are in the process and why the data were corrected.

5.3. Imputing and weighting

Description

The term data imputation refers to all procedures in which incomplete or inaccurate values determined in the process of data editing are replaced by statistical estimates. It is necessary to distinguish between procedures in which values are replaced by statistical estimates and procedures in which corrected values are obtained by re-establishing contact with reporting units. The insertion of data is discussed only in the first case.

Data imputation procedures should help to improve statistical estimates at the overall level. To achieve this goal, data insertion methods must be chosen carefully and prudently. These methods must be chosen so that the imputed values are as close as possible to valid, credible and reliable values, in order to ensure consistent data (according to a given set of logical checks) and to maintain as much as possible the basic distribution of the obtained data.

Imputation methods are divided into:

- deductive methods – imputation methods in which the decision on the value of imputation is made on the basis of known information about the respondent

- deterministic methods – imputation methods that always impute the same value to units with the same specific characteristics

- stochastic methods – imputation methods that, as a rule, impute different values to units with the same specific characteristics.

Different methods of data insertion are used, and below are displayed the most commonly used ones:

Logical imputation method (deductive method) – it enters the value that logically follows from the data, it is a piece of data that could not be collected for the subject unit. If, for example, the date of birth of an individual is known, but not his age, the missing age can be calculated for a certain point in time.

Average value imputation method (deterministic method) – it replaces the missing value with the average value of the data of the units whose data we have. Average values are usually not calculated from the whole data set, but from the data for the area in which the unit for which the data are entered is located (e.g. data for respondents of a certain age).

Internal donor method (hot deck; stochastic method) – the value taken from the other unit (donor) for which we have data is inserted. The donor can be precisely determined or randomly selected based on the function of auxiliary and other studied variables. Depending on the implementation of the method, the value of the donor is either taken over or adjusted (e.g. for the ratio of the value of the auxiliary variable of the unit for which the data is inserted and the donor).

External donor method (stochastic method) – it inserts a value that we take from an external source, e.g. from the same survey for the previous reference period (for periodic surveys), from another survey or from administrative data sources. As with the internal donor method, the value can be easily taken over or adjusted (e.g. for the auxiliary variable growth coefficient).

The stock method, that is, data structure (deterministic method) is used when the sum is known and when it is necessary to insert individual components of that sum. It is necessary to insert the value of each component by multiplying the sum by some fraction. This share can be calculated in several ways (depending on the method of implementation), but it is important that the sum of the shares of all components is equal to 1.

Regression method – values calculated from the appropriate regression model are entered into it. The simplest regression model is the linear regression model. If, for example, two auxiliary variables (X, Z) are used after the calculation of the value of the variable Y, the mathematical equation of such a model is as follows: $\mathrm{Y} = \alpha \cdot \mathrm{X} + \beta \cdot \mathrm{Y} + \varepsilon$.

Without a random component (residual), the ε regression method is a deterministic imputation method, and if a random component is added to the model, we get a stochastic method.

We can say that, by **deterministic methods**, the inserted value is calculated by a certain analytical procedure in which the appropriate deterministic function is used, while the value inserted **by stochastic methods** is calculated by a procedure using the probability mechanism.

In order for statistical survey providers to have a better insight into the impact of data insertion procedures used on statistical results, it is necessary to enable the calculation of appropriate quality indicators in the process. Two key quality indicators for the data insertion process are the share of data inserted and the impact of data insertion/input on statistics.

The inserted data share indicator is calculated for key variables, as the ratio of the number of units for which data are inserted and related to a particular variable and the number of all units that should contain data for this variable.

The impact of data insertion on the statistics indicator is calculated for key statistics, as the (relative) difference between the calculated statistics before the data insertion procedure and the statistics calculated according to this procedure.

In most surveys, data on desired population characteristics are obtained from only a part of that population (even if it is a census, i.e. a complete observation, it is common that not all units provide answers); this means that it is necessary to adjust the procedures for calculating statistics and take into account the fact that there are no data for variables of interest to the whole population. In the case of a probability sample, formulas are used to calculate statistical estimates. One of the most common ways to reduce the entire population is weight gain. When weighting, each piece of collected data is multiplied by some positive factor greater than 1 or equal to 1. These factors are called weights and are denoted by w_k, where the index *k* indicates that this weight belongs to the *k-th* reporting unit, if the reporting units are denoted by 1, 2 ... n.

If we apply probability patterns, the weights are actually determined by the sampling design. As already mentioned, the probability of selection in the sample other than zero has been already determined for each unit from the sampling frame by the sampling design, which we denote by π_k. The most commonly used weights, which are determined for the units in the sample, are reciprocal values of the probability of choice (if the sum of the population is estimated by an estimator, then in this case it is called the Horvitz-Thompson estimator), i.e.

$$\mathrm{w_k} = \frac{1}{\pi_\mathrm{k}}$$

Such weighting calculations are very convenient because they are simple, weights are known before the sampling procedure (assuming responses are collected from all units included in the sample), and probabilistic sampling theory allows the formulas used to calculate estimation accuracy to be simple, at least for most commonly used estimators (average, sum, share). Such a choice of weights at the same time allows for an unbiased assessment.

In practice, attempts are being made to improve this simplest procedure for calculating weights (and thus increase the accuracy of estimates) by using the so-called ratio estimator, in which we multiply the weights w_k by a certain factor (called proportional correction or calibration correction), which is calculated based on data for an auxiliary variable that is well correlated (linear relationship whose graph passes through the origin of the coordinate system) with the variable of interest. A variable whose data are known before conducting the survey (i.e. they exist within the sampling frame) or whose value for the total population is known, and individual data were obtained on the sample, is called the auxiliary variable. If the auxiliary variable (marked as x) is sufficiently correlated (at least 0.5) with the variable of interest, the above factor is calculated by dividing the total of the auxiliary variable for the whole population (t_x) by the weighted sum of the auxiliary variable values for the sample units (x_k), where the weighs of the reciprocal values of the probability of choice equal $\frac{1}{\pi_k}$

$$\mathrm{g-weight} = \frac{t_x}{\sum_{k \in s} \frac{1}{\pi_k} x_k}$$

The weight of the ratio estimator is

$$w_{k,\mathrm{ratio}} = \mathrm{g-weight} \cdot w_k$$

Ratio estimator weights are very often used in business surveys because these surveys use auxiliary variables from administrative data sources and other sources (number of employees, company turnover, salaries, investments, etc.).

Surveys in which individuals and households are observed often use a similar technique called calibration. The goal of calibration is to change sample weights so that, when used to weight data for an auxiliary variable, the exact value of the auxiliary variable for the population is obtained. The procedure for calculating calibration weights is exactly the same as for ratio estimator weights. The intention is often to perform calibration on several auxiliary variables (e.g., sex and age groups). A special example of calibration methods is poststratification. In this procedure, the values of the population of auxiliary variables are also known for some subpopulations that we did not plan with the sample design (e.g. because we did not have enough data within the sampling frame). In the poststratification, the weights for each such subpopulation are recalculated separately, so that the estimates for the variable of interest are equal to its exact value in all subpopulations.

If the survey is not conducted on a probability sample, then models are used to determine weights. The biggest problem with this is that it is generally difficult to estimate the accuracy of such estimates when using the model.

Finally, there is an example of a lack of data due to unit non-response. In practice, when conducting individual statistical surveys, the units covered by the statistical survey never fully respond to the questionnaire. Thus, there is a need for additional weighting of the units that submitted a response because in this case it is necessary to first recalculate the weights of the units in the sample that responded to the sample and then the sample to the total population of the sampling frame. In order to address this challenge, it is necessary to carefully analyse data that are incomplete due to unit non-response to determine if there are differences between responses and non-responses. The most common procedures for correcting sample weights due to respondents' non-response are the assumption of a two-stage sampling plan (model) and the use of calibration methods. In the case of a two-stage sampling plan, it is assumed that in the first stage the actual sample of units was selected, while in the second stage we have responding units and possible inappropriate units. The probability of selecting the unit that responded is the product of multiplying the probabilities of the first and second stage, while the weight is the reciprocal value of this product. Of course, these weights are used only on the data taken over from the responses, assuming that only statistics on the areas of the relevant units are of interest.

Quality assurance instructions

Data that are missing as a result of item non-response should always be replaced by an appropriate statistical estimate. Maintaining incomplete values in the final data set can, on the one hand, lead to bias in the results and, on the other hand, make it more difficult to conduct more complex statistical analyses.

In the case the data are incomplete due to unit non-response, insertion methods should be used only for a small number of variables or if there are good reasons that the use of these methods can significantly reduce bias. Otherwise, the non-response weight calculation should be used for data that are incomplete due to unit non-response.

Before selecting data insertion methods, precise and sufficiently extensive simulation studies should be performed to assess the selection of the most acceptable data insertion method. The following should be performed by conducting simulation studies:

- develop a mechanism for creating data that are missing and that will best match the assumed reality of the survey conducted
- when testing the methods, use all auxiliary data available during statistical data processing
- test the estimation method that best preserves the statistical properties of results at the macro level, rather than the method that best predicts values at the micro level.

Each piece of inserted data must be marked in the statistical process with the appropriate status of the variable as this indicates which insertion method was used. At each stage of the procedure, it should be possible to separate the collected data from the data resulting from statistical estimates.

Existing basic software solutions should be used as much as possible when preparing software to perform data insertion procedures.

It is necessary to ensure the calculation of quality indicators that provide insight into automated insertion procedures. Indicators at the micro and macro levels should also be calculated. Particular attention should be paid to the proportion of data inserted and the impact of data insertion on final results.

The relevance of the applied methods should be constantly checked in the analysis. If the analysis shows that the applied methods are not appropriate (any more), the procedure needs to be adjusted.

If possible, the survey is conducted on a probability sample because only on this basis are weights determined that provide credible data on the accuracy of estimates. When weighting, special attention should be paid to outliers, i.e. to units that have a very large (or very small) value of the analysed variable compared to other units in the sample. If it is reasonable to assume that such a unit is also isolated at the level of the whole population (or the corresponding stratum), its weight is corrected either to a value smaller than 1 or to a value that equals 1.

After calculating the final weight values, they need to be summed for check purposes because their sum must be at least approximately equal to the number of units in the population (or areas of the population).

When using a ratio estimator or any other calibration method, it is recommended that the corrected weights do not differ much from the sample weight.

If the non-response rate is very high, it is recommended to conduct additional survey on the sample of unanswered units and to recalculate the weights of the sample based on the analysis of this sample.

If auxiliary information is not available for non-response analysis, then the final weights in which the response is taken into account are calculated using the assumption of a two-stage sampling plan. If ancillary information is available that correlates well with the variable of interest, then the non-response can be analysed. The results of the analysis help to determine the groups in which the units behave similarly, and it is these groups on which the assumption of a two-stage sampling design is made or the influence of non-response is removed by an appropriate method.

In business surveys, the Croatian Bureau of Statistics has a status list for all units conducted on a sample that can be grouped according to the response status, the non-response status and the status for ineligible units. At the end of the survey, the general methodologist should get a list that includes all sample units and in which each unit has the appropriate status.

If the observed unit is added to the address book after the selection of units, this must be recorded, as well as the reasons for it.

If a unit is reported to belong to a group that is not part of the stratum in which it was selected, and the group is still part of the observed population (e.g. another stratum), then the weights are usually calculated at the stratum level in which the unit was selected, while it is statistically published at the level of the group in which the existence of such a unit is registered.

5.4. Production of derived variables

Description

The production of derived variables and statistical units that are not explicitly collected during data collection is done by applying arithmetic formulas to one or more existing variables. New statistical units can be derived by aggregating or dividing data for collection units or according to different estimation methods (e.g. household creation, when the units in the data collection are persons or enterprises, or units are legal entities).

Deflation in the statistical process refers to the process by which the impact of price changes in a given period is excluded from value data. At the theoretical level, it can always be imagined that value data consist of two components: quantity (or volume) and price. When calculating the ratio of value data at two time points, the calculated ratio includes the change in quantity and price. If we want to obtain the most accurate estimate of the change in quantity based on the measured change in value, the impact of the change in prices must be removed from the value ratio (index). The index calculated from the “original” individual data is called the nominal index, and the index calculated from the data from which we excluded the impact of price changes is called the real index.

There are two basic approaches to performing the deflation process. The first is the process of deflation at the micro level, and the second is the process of deflation at the macro level. In micro-level deflation, each piece of data at the individual level is first divided by the appropriate deflator, and thus (in terms of price changes) the data are converted into the same (fictitious) time point. For example, if the consumer price index is used on a fixed basis, the average of the year, as a deflator in 2005, then each recalculated individual value piece of data represents a value that would (with then valid prices) be valid at a fictitious time in the average year.

In the case of macro-level deflation, price "removal" procedures are carried out on already calculated aggregates or indices. It differs from the micro-level calculation mainly in that a deflator is needed here for the same period as the index being deflated. If the index in the current month is deflated compared to the previous month, a deflator (i.e. the corresponding price index) is needed for the current month compared to the previous month, but if the index is deflated in a fixed base year, a deflator in a fixed base is needed (e.g. 2005 average). In the process of deflation at the micro level, only one time series of deflators is needed, while in deflation at the macro level, as many time types of deflators are needed as there are time series that we want to deflate. This is also the biggest advantage of micro-level deflation. The advantage of macro-level deflation is a greater process stability because macro-level deflation is less demanding to process due to the smaller amount of data being processed.

Quality assurance instructions

In deciding whether to use micro- or macro-level deflationary procedures, all the advantages and disadvantages of both approaches must be taken into account. First of all, it is necessary to choose an approach that will enable a simpler and more rational placement of the procedure in the overall statistical procedure that best suits the observed phenomenon.

When determining the procedures for deflator preparation, the first step is to check which price indices are the most appropriate as basic data for deflator preparation. In the second step, it is necessary to check whether the price index is already a suitable deflator or whether it should be combined with other structural data. The aim is to bring the calculated deflator closer to the phenomenon of real price movements measured on the basis of value data.

The simplest and, if possible, automated generation of price indices should be ensured. Restricted access also needs to be provided.

The suitability of the deflator used must be constantly checked. The information on it can be very useful for the area of national accounts.

5.5. Aggregate calculation

Description

Aggregated data and total results are calculated from microdata (e.g. aggregation of data according to economic, social, geographical and other classifications). Data with common characteristics are summarised, average and dispersion measures are determined, and the weights obtained in the subprocess of imputation, weighting and estimation of totals are applied. If a random sampling method is applied, standard errors are calculated.

The term calculation of statistical estimates, or aggregation of data, means the part of the statistical process in which estimates are calculated from the final microdata, which are also called statistics. Since in most statistical surveys not the whole population is observed, but only a part of it (a random sample), the statistical calculation also includes the calculation of the value of the population or the calculation of population estimates. The process or function by which population estimates are calculated from sample data is called the estimator.

In general, there are several different types of aggregates and estimators, but in the practice of the Croatian Bureau of Statistics some of the most important aggregates predominate. We will give a description of the most commonly used estimators. It is assumed that the corresponding weight of the population (w_i) is already calculated for each unit of the sample and that the weighting process is done respecting the sample design. Further, it is assumed that $\{y_i, i = 1,..,n\}$ are values of the target variable Y measured on the sample. We denote the population estimate by $\hat{Y}$.

Estimation of total population. Assessment of the total sum of the variable Y on population is calculated by using the formula

$$\hat{Y} = \sum_{i=1}^{n} w_i \cdot y_i$$

Population average. Assessment of the average of the variable Y for the population is calculated by using the formula

$$\hat{\bar{Y}} = \frac{\sum w_i y_i}{\sum w_i}$$

Number of units with certain characteristics. Let D_{and} be a variable that takes on two values for each observed unit: value 1, if the unit possess observed characteristics, and value 0, if the unit does not possess these characteristics

$$\hat{Y} = \sum w_i D_i$$

Proportion of units with certain characteristics. The share of units with certain characteristics in the population is estimated by using the formula

$$\hat{p} = \frac{\sum w_i D_i}{\sum w_i}$$

The ratio of the total population. The measure of the total population for two variables (X, Y) is estimated by using the formula

$$\hat{R} = \frac{\sum w_i y_i}{\sum w_i x_i}$$

If the population value estimate for the selected statistics is calculated on the basis of data obtained from the sample, such an estimate also contains a sampling error.

In the case of a probability sample, this error must be estimated and displayed in an appropriate way when the results are published. The procedures for calculating the sampling procedure error can be quite complex. They are therefore determined by the type of estimator and the sample design used. Roughly speaking, sampling error estimation procedures can be divided into three groups:

- application of direct formulas for estimating standard errors of linear estimators and use of "approximate formulas" of Taylor linearisation for non-linear estimators (analytical approach)
- estimation of standard error based on subsample selection and then use of appropriate formulas for such application (re-sampling)
- application of a regression model for estimating standard estimation errors on subgroups of the considered population.

There are also several ways to present estimated sampling errors in publications. In the practice of the Croatian Bureau of Statistics, two methods are used: the sample error is explicitly (value-wise) published together with the estimated statistics, or less accurate estimates are marked with special labels.

Quality assurance instructions

Estimates that are at least approximately impartial should be selected when calculating population estimates based on the sample.

It is certainly necessary to calculate estimates and sampling errors of these estimates in the same statistical process. The survey holder, together with the estimated statistics, gets a sampling error and assesses how reliable the estimated statistics are.

The sample design used must always be taken into account when preparing the procedure for calculating the sampling error. When preparing instructions for presenting sampling error, it is necessary to take into account the standards adopted by the Croatian Bureau of Statistics. Sampling errors need to be analysed regularly, mainly due to estimates calculated with insufficient precision when sampling error is too large. If the analyses show that the calculated estimates are not accurate enough, measures should be introduced to reduce sampling errors. Such measures may result either in an increase in the sample if available resources allow or in the definition of a more efficient estimator using auxiliary variables.

5.6. Creating final data files

Description

The results of all subprocesses in this phase are combined, resulting in a final data file used as input for the analysis phase. Sometimes this is a transitional rather than a final file, especially for very important statistical surveys for which there is a need to produce preliminary and final estimates.

The notion of editing data at the macro level, in the narrowest sense of the word, means identifying errors on already collected data. Such procedures in the process of conducting statistical survey are placed in the phase of data editing at the micro level. As a rule, only error detection occurs at the macro level, while error correction should always be performed at the micro level. There are several methods for editing data at the macro level, and below is an overview of the methods:

Aggregate check method. The basic idea of the method is to first check the weighted and aggregated data (e.g. at the four-digit level of the NKD), and then to check, at the micro level, all units from the groups that are questionable at the macro level. The definition of macro-level checks is mainly based on the distribution of ratios and differences according to the results from the previous period.

Gradual extraction method. The basic idea on which this method is based is that with the help of an appropriate computer application, the control and editing of data is limited only to those units whose impact on the final assessment in a particular subgroup is not negligible. The procedure using the appropriate application shows the units whose impact on the aggregates is not negligible. The values of these units are then checked in the subsequent procedures.

Graphical methods. The purpose of applying graphical methods is to graphically present the distribution of weighted data values that need to be checked. A graphical presentation of the distribution and key distribution parameters (e.g. quartiles) can facilitate the determination of extreme value limits. Graphical methods are mainly used as a "complementary" tool when implementing other data editing methods.

In the tabulation process, structured aggregate data are prepared for different types of publications: for electronic publications, the Croatian Bureau of Statistics database, printed publications, tables that meet the requirements of standard users and for tables intended for international reporting. Prepared tables can be used in other sub-processes, except in control tables: in macro-level editing, in statistical data protection, in suitability analysis and in validation of results. It is important that the publicity tables are the result of the application of pre-prepared computer procedures that can be performed automatically or in a way controlled by the survey holder. This increases the ability to organise and store all calculated estimators in a single

macrobase, reduces the risk of time constraints that jeopardize the timeliness of publishing results, and largely eliminates the influence of the human factor on the accuracy of results.

Tables must be designed optimally according to the content, data preparation technology, technical capabilities of the tool and accepted data publishing standards. The process of preparing tables can be carried out independently, but for the most part it makes sense to combine it with some other processes, especially with statistical data protection and calculation of statistical estimates.

The sub-process defines in detail the results that we have roughly determined in the process of analysing the necessary requirements.

In this sub-process it is generally necessary to take into account the following: the type of observation units and characteristics that determine the population, the geographical location of the units included in the population and the reference period for which important population characteristics apply.

In order to realise the above, it is necessary to use harmonised terms, variables and classifications as much as possible, and in some cases, the above definitions must be adjusted to the specific needs of the survey. All definitions, variables and classifications should be documented, as well as any deviations from the standard.

In addition, a detailed list of all variables to be covered by statistical surveys needs to be prepared. For each variable it is necessary to indicate the following: data source, i.e. whether the data source is a statistical questionnaire or an administrative data source; reference period for data entry; information on when the requested data source is available.

Quality assurance instructions

Group checks should be carefully defined. In particular, situations of excessive and inefficient management of data at the micro level should be avoided due to too strict checks at the macro level.

The feasibility of checks should be examined before use in the regular procedure. Data from past statistical surveys or data from related statistical surveys can be used for test data. The use of graphical methods is recommended because visual presentation is usually the most effective way of perceiving “suspicious” values. Macro-level data editing should only be used to determine inaccurate or questionable values.

All data corrections should be performed at the individual record level. Procedures should be as automated as possible as this improves the efficiency of the whole statistical process.

Standard instructions should be used to create the tables. If code lists are used when creating tables, the standard code lists located on the classification server (KLASUS) should always be used. Tables for public disclosure should be formed on the basis of procedures prepared in advance.

For the preparation of internal ad hoc and control tables, the holders of statistical surveys should use standard interactive analytical tools. Since this procedure combines different processes and includes different organisational units, it is necessary to establish good coordination and properly document work procedures in a standardised way.

Standard concepts, variables and classifications are used to set definitions. All terms, variables and classifications used in the statistical survey are prescribed in detail, while all possible deviations from standard terms are described separately. In addition, the differences between the potential phenomenon being monitored and the phenomenon important to users of the statistics, e.g. the difference between the target[9] and observed population[10], is described.

[9] Target population – a set of units to whom the survey refers and to which the obtained results are generalised.
[10] Observed population – list of units of the basic set from which the sample is selected, e.g. registers of business entities.

In determining the classifications, the classifications stored in the KLASUS classification server are used, such as national classifications and classifications of international organisations such as Eurostat, the United Nations, the OECD, the World Bank and the International Labour Organisation.

Classifications should be designed to allow reporting units to easily classify the phenomenon while allowing the publishing and dissemination of data at a sufficiently detailed level without any restrictions referring to the provisions of confidentiality and accuracy of data. By applying hierarchical classifications, i.e. publishing data at higher hierarchical levels, the stated condition can be met. Regarding the assessment of the quality of results, it is necessary to have at least basic quality indicators available that are the result of the implementation of statistical surveys.

5.7. Production and updating of statistical registers and databases

Description

Statistical registers are continuously or regularly updated sets of objects of a certain population. They contain the information needed to identify and access a particular population of units as well as other features needed to conduct statistical population surveys. Statistical registers most often contain the current and historical state of the population and the causes, consequences and sources of changes in the data on the populations of statistical units. They integrate data from multiple statistical and administrative sources, linking them using common identifiers, and storing them in structured databases.

Statistical registers have a coordinating role in the implementation of statistical surveys, and their timeliness and quality have a great impact on the quality of surveys and statistical products.

Among the many functions of statistical registers, the following are particularly important: 1) identification and construction of statistical units, 2) unambiguous recording of links between units, 3) population records, 4) records of the main characteristics of units; recording the classification of units, 5) developing a sample selection framework, i.e. establishing a common framework for organising and coordinating statistical surveys by ensuring a harmonised sampling framework according to stratification characteristics, 6) providing a basis for aggregating sample results and 7) monitoring creations and closures and other demographic events of units.

Statistical registers are the backbone of the organisation of statistics, and some (such as the SPR) are based on legal documents.

The Croatian Bureau of Statistics has established statistical registers for the purposes of official statistics, using data from administrative data sources, censuses, statistical surveys and data collected by the observation and monitoring method.

The established statistical registers are SPR, SRPG and PSR.

SPR is a register that contains data on business entities (legal entities and natural persons), their parts and groupings, which are institutionally and formally involved in the production and financial processes of the national economy. Business entities, as administrative units, are transformed into statistical units suitable for statistical monitoring and analysis in order to, among other things, achieve comparability with other EU Member States. The most important statistical unit is the enterprise as the smallest combination of legal units that makes up the organisational unit for the production of goods and services. The enterprise is the end result of a profiling method that analyses the legal, operational and accounting structure of groups of enterprises at national and global level in order to create statistical units for the most efficient collection of statistical data.

The purpose of the SPR is to enable efficient collection of data on business entities, their processing and analysis, carried out by organisational units of the Croatian Bureau of Statistics, aimed at monitoring business entities and other producers of official statistics.

SPR is the main source of data for business demographic statistics as it observes business-related demographic events.

The main sources of data for updating the SPR are administrative and statistical sources, such as the European Register of Enterprise Groups and the statistical surveys of the Croatian Bureau of Statistics.

The SRPG is a register containing data on units engaged in agricultural production: legal entities and their parts, craftsmen and family farms. The SRPG consists of a basic and a statistical part. The basic part contains basic data on the agricultural holding (identification data, data on the owner/name of the business entity, address data, etc.).

The statistical part contains data from the field of agricultural statistics (agricultural areas by categories, number of livestock, etc.). The purpose of the SRPG is to ensure that survey is conducted by applying a reliable for the sample selection framework for statistical surveys in the field of agriculture, listing and updating address data for various surveys in agricultural statistics and better coverage of agricultural units in agricultural statistics. The main sources of data for updating the SRPG are administrative sources, SPR and regular statistical surveys in the field of agriculture.

PSR is a register containing data on spatial units for statistics of the 1st, 2nd and 3rd level of the National Classification of Spatial Units for Statistics, counties, cities/municipalities, local self-government units, settlements, statistical circles, census districts, streets/squares, house numbers, households and more. It was established as the official basis for the collection, recording, presentation, exchange and connection of various types of spatial data in the field of official statistics of the Republic of Croatia. The data contained in the PSR serve as a technical basis for official statistics activities.

The PSR data are kept in alphanumeric and graphical form. PSR is a source of data for statistical registers and statistical activities that use data on spatial units. The data are updated regularly. Methods of data collection can be online access, electronic media and printed form. The main source of administrative data is the State Geodetic Administration, and the main source of statistical data is the surveys of the Croatian Bureau of Statistics.

KLASUS is a tool for all users of classifications that offers the ability to view and search classifications by name and code in one place, allows the user to display individual elements of classifications, their names, explanations or indices, if any, as well as a correspondence tables between different versions of classifications. It also provides the download of multi-format classifications with all levels and elements of classification. Classifications are grouped within "families" that include classifications related in a certain way, while "shortcuts" allow direct access to the most commonly searched for classifications.

This application allows tabulation and viewing of data described in the central metadata database, while in the Survey Processor application allows tabulation of these data. Users can create simple tables themselves and save them in a number of different formats, and they can also save aggregated data as new sources of data for tabulation. Tables prepared in the Survey Processor can also be used in the Warehouse Browser, whose results are created in Excel based on queries directly from the data.

The statistical data warehouse includes tabulation-ready data that can be divided into microdata, macrodata and ready-made data.

Macro data contain cubes. A cube is a synonym for aggregated data or macrodata obtained by processing microdata, and it can also be generated from other macrodata. Aggregate values in cubes are most often obtained by summing quantitative variables in microdata by categories belonging to selected qualitative variables. The corresponding categories of qualitative variables are called cube dimensions.

Tables (ready-made data) contain tables of aggregated microdata, but they contain more aggregated levels of data, unlike a cube, which contains only one single level. This means that summary rows for some dimensions can also be displayed in the table.

Quality assurance instructions

Statistical registers should be relevant to the needs of their data users. They should contain relevant units and variables and enable the selection and generation of relevant populations to carry out the sampling procedure for statistical surveys. Accuracy is an essential feature of the quality of statistical registers, and belongs to the most easily measurable principles of quality.

Timeliness and punctuality, as well as coherence and comparability, are also important quality principles for most users of statistical registers. Ease of access to data from the registers is considered a component of quality as well as the interpretation of information stored in the register. As far as comparability is concerned, there are two aspects of comparability, namely: comparability in space and comparability in time. Thus, for example, comparability for statistical business registers in the EU is prescribed by the Regulation on business statistics. Therefore, in this sense, one of the quality measures is the level of compliance with applicable regulations. This type of comparability is very important because all statistical registers are vital for many statistical areas. Comparable registers facilitate the process of harmonising statistics derived from such registers. Comparability of units and classifications used in statistical business registers is also one of the set goals of European statistical legislation. Comparability over time may be less important for many users, but there is still a need for insight into data and units as well as for comparability of total data over different periods.

On the example of the SPR, compliance can be observed in terms of internal compliance and compliance with other registers. Internal connectivity refers to the consistent recording of data in the registry, such as the consistent use of update and profiling rules. Connectivity with other registries is best maintained in the storage and use of unique identifiers. This allows comparison of data and units between registers, e.g. within the Republic of Croatia: OIB, ID Number or MBS in SPR allows relatively easy connection with administrative registers (e.g. Administrative Business Register, Taxpayer Register, Court Register, Trade Register, etc.). To connect the SPR with the EuroGroup Register, a unique international identifier is used, the so-called LEID ID. The use of a unique identifier in all business registers (administrative and statistical) is one way to achieve better connectivity. Even when this connectivity is achieved, it has certain limitations because the definitions of the units stored in the registers sometimes differ.

The statistical register can be said to be complete if it includes all units from the target population and all the necessary variables. In reality, it is impossible to achieve full coverage due to the impossibility of timely delivery of data. Therefore, completeness should still be the goal, and its measurement will be shown in quality indicators.

6. Analysis

Statistical data analysis is performed using various tools and techniques. The aim of the conducted analyses is to explain situations and events, to single out certain rules that are characteristic of the observed phenomenon and to briefly interpret the obtained results. The data are analysed at the macro level, and when such an analysis is not precise enough, then the analysis is also performed at the micro level.

The analysis is conducted to confirm the meaningfulness of data or to reveal their possible shortcomings. By eliminating shortcomings, better data quality is achieved. If the analysis shows systemic deficiencies, the calculated data are used to improve quality, that is, to supplement the business process or change the methodology. For the correct interpretation of the data, a systematic analysis is performed and practical cases are presented on concrete examples. This way of working can determine certain statistical laws that occur systematically in the observed area.

At the macro level, data relating to a specific time point or a longer period can be analysed. If significant discrepancies are found in the macro-level analysis and cannot be explained at that level, a micro-level analysis is performed because macro-level results are derived from the microdata.

Data analysis also includes an analysis of the impact of procedures carried out in statistical data processing. For example, the analysis checks the impact of non-response and data insertion on the final results and checks the impact of the selected data insertion procedure, then the differences between provisional and final data and the like. All the above information is the basis for the preparation of data quality descriptions, and also the basis for supplementing and improving the survey methodology. For the analysis and verification of data, other data sources and information that are directly or indirectly related to the subject area are used, among other things. Before interpreting a statistical phenomenon by analysing the data, it is necessary to determine what the data show. At this stage of the business process, the statistician must ask himself what the data shows and answer the questions asked by analysing and interpreting the data.

6.1. Statistical analysis of results

Description

The collected data are transformed into statistical results. Statistical analysis of results includes the production of additional measures, such as indices, trends or time-adjusted series, as well as the indication of qualitative characteristics. In order to increase the value and create the preconditions for the analysis of statistical data, the preparation of maps, GIS results and geostatistical services can be included in this sub-process.

Data time series refer to time sequences of data, in our case collected statistical data for certain survey areas (e.g. industrial production, retail trade turnover, labor cost indices, etc.). The main part of the analysis of time series is related to seasonal adjustment and trend determination, and the smaller part is related to forecasting. Seasonal adjustment removes the influence of seasonal and calendar components from time series, when these components are pronounced and important. The values thus obtained are called seasonally adjusted values or values with excluded seasonal and calendar effects. It is useful to perform this procedure whenever it is necessary to compare data from different periods of the same time series or data for the same period of the same time series from different countries because they are usually subject to seasonal oscillations, number of working or trading days and other influences. When comparing data for the same period of different years in the same time series, checks whether calendar effects are present in the time series are conducted and then they are excluded because the season for the same period in the year is approximately the same (e.g. comparison of April 2010 with April 2009). Data adjusted in this way are considered data with excluded calendar effects or data adjusted to the effect of the number of working or trading days.

The impact of the number of working or trading days

The impact of the holidays

The impact of Easter

The impact of the leap year

For gross, non-adjusted time series (X_t), a model of seasonal adjustment is made, from which the seasonal component (S_t),

irregular or random component (I_t) and trend cycle component (TC_t) are then determined.

$X_t = TC_t + S_t + I_t$

The trend cycle component contains a long-term tendency of the development of the phenomenon, while the irregular (random) component contains random effects, i.e. residual deviations from the trend after removal of other components. Models are generally revised once a year, usually with the first or last data in a calendar year. Then the holder of the statistical survey submits to the methodologist for time series all files with model parameters and data for check, i.e. annual review or revision of the model.

There are the following types of outliers, i.e. atypical values:

- **additive outliers** (abnormal values in isolated points of the series)
- **temporary changes** (series of outliers with decreasing effects on the batch level)
- **level shifts** (a series of innovation outliers with a constant long-term effect on the level series, where the innovation outlier means atypical values in the innovation series)
- so-called **ramp**, which describes a smooth, linear, or quadratic transition between two time points, as opposed to a sudden change associated with a level shift
- **temporary level shifts**, in which the level shift has a short-term rather than a long-term effect.

If the outlier appears at the end of the time series, then it is a temporary atypical value and the methodologist fixes it only when he receives at least the next three pieces of data in the time series because only then can he determine the type of the outlier. In this case, each time when new data are added, the statistician should submit the time series for check to the time series methodologist until the outlier is fixed.

The length of the time series that is time-adjusted must be at least three years for monthly time series and at least four years for quarterly time series. The existence of calendar influences (one or two regressors) is determined in time series that contain data for at least five years, and the existence of several regressors (six or seven) only in time series that contain data for at least seven years.

Quality assurance instructions

The survey holder should carefully review and analyse the time series before submitting them to the methodologist for modelling, since this can improve the quality of the time series models.

The holder of the statistical survey must inform the time series methodologist of all changes related to the time series, such as changes in sampling, changes in methodology, changes on the market and the like.

The time-series methodologist should consult with the holder of the statistical survey on the importance of the seasonal adjustment results (existence of calendar influences, causes of atypical values). Comparisons and reviews of seasonally adjusted series can be found in publications, e.g. on an annual basis (example: April 2010 compared to April 2009) as well as data on the calendar adjustment, i.e. the impact of the number of working days in individual time series. The trend cycle component can only be represented graphically, and attention should be paid to the problem of endpoints.

Long time series, which are longer than 12 years, sometimes need to be shortened, e.g. to seven years, in order to improve the quality of the model. Namely, if the peculiarities of the season change over such a long period, the results of seasonal adjustment may have poor quality. Data not included in the seasonal adjustment model, e.g. older than seven years, are only available as original data.

It is important that the holder of the survey does not forget about the annual verification of models and their revision if any and that once a year he submits all time series models and data to the time series methodologist for check.

6.2. Quality control of results

Description

The quality of the produced results is checked in accordance with the general quality framework and expectations.

Quality analysis of results is a process in which the relevance, accuracy and comparability, availability and clarity of data in time and space, and compliance and comparability with existing internal and external reference data sources are checked.

Quality analysis or validation of results is performed after the macro-level editing procedure and may include the following procedures:

- checking the consistency of results if it has not already been incorporated into the macro-level editing process. For example, whether the value of production is greater than value added.
- the consistency of the results with the results from the previous reference periods is checked, and this is especially true for those surveys whose primary purpose is not to measure changes over time.
- it is necessary to make an internal check of the results at the level of the Croatian Bureau of Statistics and periodically check the relevance and consistency of the results with external experts.

Quality assurance instructions

Prior to the final validation of the results, all data available in connection with the implementation of the data processing process must be re-analysed, especially the results prepared at the macro level.

When choosing methods and criteria for assessing the consistency and comparability of results, it is necessary to take into account the characteristics of the basic data and the status of the prepared results, i.e. whether these are provisional or final results.

When choosing methods and criteria for assessing the consistency and comparability of results, it is necessary to define the procedures that are carried out before each publication of results, which are carried out occasionally together with the periodicity of their implementation.

Verification procedures should be adapted to the period of implementation of the statistical survey, the target population, the method of data collection and the type of data source, which may be primary or secondary.

The results of the analysis should be used in quality improvement programming documents.

6.3. Detailed analysis and interpretation of data for publishing

Description

What is assessed is how successfully the statistics reflect the initial expectations by observing the statistics with the help of various tools and media. It is ensured that statistics reach the required level of quality and readiness for use.

The data must be relevant and applicable, while the conclusions should be supported by data resulting from the statistical process. The interpretation of the data should take into account the methods of data collection, i.e. whether it is full coverage, conducting a sample survey or using an administrative database as well as additional information closely related to data collection, such as non-response rate. All information should be available to users with relevant metadata, from methodological explanations, code books, indications of data sources, etc., as they can help in interpreting the accompanying comments, tabular or graphical representations. Particular attention should be paid to descriptions of data gaps, such as deviations from the target population and the applied population. The principle of data confidentiality must also be taken into account when interpreting data. For most publications, such data are listed in the final part of the publication, while the results are presented in the initial part of the publication.

Current topics, emergencies and the interpretation of data and related phenomena should be tailored to the data user, taking into account what type of information is important to the user.

The interpretation of the data should be adapted to the basic characteristics of the user group, whether they are the general public or experts in a particular field. The statistician should be familiar with the wishes of the user and be able to present the data so that they are in line with user expectations, which means that the data should be understandable, interesting and useful. The general public is interested in the most important or most interesting general information expressed in statistical data, in contrast to experts in a particular field who mainly use detailed data for the purpose of further data analysis.

Quality assurance instructions

The interpretation of the results is adapted to the target population and the medium in which the data are subsequently published. It must be impartial, objective, accurate, clear and understandable. The interpretation of short-term statistics must be different from the interpretation of structural statistics. The use of reference points has a strong influence on the interpretation of data, so it is necessary to use reference points, especially in time comparisons, when the data are stable and unbiased. When interpreting the results in the form of indices and other relative numbers, a representative comparison period must be chosen that will allow the data user to effectively interpret the evolution of the phenomenon. When presenting results in the form of indices and other relative numbers, care must be taken to accurately interpret the change in the phenomenon when expressed as a percentage.

6.4. Protection of confidential data

Description

Statistical data protection is a set of methods that reduce the risk of disclosure of data collected from reporting units (persons, households, enterprises, etc.). It includes:

- statistical protection of tables (aggregated data)
- statistical protection of microdata (individual data on reporting units).

The official statistics of the Croatian Bureau of Statistics are organised on the basis of Regulation (EU) No. 223/2009 on European statistics and the Official Statistics Act.

There are two types of statistical data protection methods:

- perturbative methods – modify data to such an extent that the risk of recognition is minimised (e.g. rounding procedure)
- non-perturbative methods – they do not change the data, but cover them up (e.g. by the method of defective values).

Some methods can only be used to protect tables (e.g. domination rule), while other methods can only be used to protect microdata (eg microaggregation).

Statistical data protection prevents the disclosure of sensitive data of all reporting units that appear in table cells. In statistical data protection, two groups of tables are distinguished: frequency tables (the number of reporting units belonging to this cell appears in each cell of the table) and value tables (in each cell there is a sum of values for certain variables collected by reporting units).

In the statistical protection of tables, attention should be paid to the hierarchy of variables and to the relationships between tables and variables.

In recent years, we have increased the availability of available databases (e.g. census data were not previously available, nor it was possible to link databases of various surveys, including external databases), and there is a visible increase in the number of requests for confidential statistical data (microdata, i.e. statistical units that allow indirect identification) for the purpose of conducting statistical analyses for scientific purposes. Namely, in relation to aggregated data, microdata enable more detailed and precise analyses necessary for drawing conclusions envisaged by the goals of survey projects. Microdata protection is important due to the large number of variables contained in the microdata file because certain combinations of variables are rare, while the risk of statistical unit disclosure is high. If microdata is publicly available in the form of a file for public use, the level of protection will very high and data loss will be much higher. If microdata is used for the purpose of conducting statistical analyses for scientific purposes based on a signed contract, the level of protection will be lower because the likelihood of misuse is lower.

Quality assurance instructions

Input data must be prepared in the appropriate form and format. The relationships between the tables and/or the variables to be protected need to be examined in detail. All published and transmitted tables that contain or will contain the same data (e.g. tables for national publishing and tables published by Eurostat) must be taken into account.

Each organisational unit of the Croatian Bureau of Statistics determines the rules and methods of protection of confidential statistical data in accordance with the provisions of the Guidelines concerning Protection Rules and Methods in Dissemination of Official Statistics of the Croatian Bureau of Statistics (Official Gazette, No. 78/21).

It is recommended that the variable or its classes not to be defined in more detail than necessary, since this can have a strong impact on information loss.

The number of variables in the microdata file must not exceed the number required to perform quality statistical analyses for scientific purposes. The Guidelines concerning Protection Rules and Methods in Dissemination of Official Statistics regulates the procedures and methods of applying physical, technical and organisational measures to ensure physical, technical and logical protection of confidential statistical data from unauthorised access, alteration, loss, removal or destruction and other issues important for security and protection of confidential statistical data collected for the production of official statistics within the scope of the Croatian Bureau of Statistics. Measures for the protection of confidential statistical data include rules and procedures for the protection of identification data on statistical units in the course of their collection, processing and delivery, and in particular include:

- protection of premises in which statistical data are stored, which includes control of entry and exit of employees
- protection of access to information technology equipment (hereinafter: IT equipment) and network servers, devices and media used to store confidential statistical data in electronic form
- protection of access to confidential statistical data by use of security clearances and authorisations
- recording events related to access, use and disclosure of confidential statistical data
- keeping records of permits and authorisations of employees who have access to confidential statistical data in the performance of their duties
- the obligation to sign a declaration on statistical confidentiality
- other issues related to security and protection of confidential statistical data.

Measures for the protection of confidential statistical data must be applied by all employees of the Croatian Bureau of Statistics as well as other natural persons or legal entities entrusted with the performance of certain tasks or duties related to official statistics activities on the basis of a contract or authorisation.

If the reporting units agree to publish data that enable their identification, such reporting units are granted a special status, which reduces the loss of data in the tables in which the data are published that relate to that reporting unit.

7. Dissemination

Publication of statistical data and information includes the dissemination of statistical data to the public or to all users, using various media and types of publications.

The results of statistical surveys should be published in a timely manner and in compliance with the publishing deadlines. Data should be relevant, accurate, consistent, comparable, accessible and clear.

The most important medium for publishing data from the Croatian Bureau of Statistics is the website https://www.dzs.gov.hr/. The data should be published so that they are available to all users at the same time and in the same way. Data and metadata are available to users, including methodological descriptions and information on the quality of the data collected. An important way of publishing data is to inform users about planned data releases and about revisions and corrections of inaccurate data.

The Croatian Bureau of Statistics publishes data in the following publications, cartographic portal and databases:

STATISTICAL INFORMATION

This pocket-sized 105-page publication contains a summary of annual data for the Republic of Croatia and an overview by county. The data are given in three-year series and are a valuable selection of basic information about the Republic of Croatia.

CROATIA IN FIGURES

This publication contains the most important data on economic and social developments for the Republic of Croatia presented in five-year series. Tables and graphs are located within 20 chapters. It is intended for the widest circle of users.

FIRST RELEASE

This is short and flash statistical information that is published according to the periodicity of statistical surveys (monthly, quarterly, semi-annual, annual, biennial and triennial). The first release is published a few days after the completion of the processing of the results of a particular statistical survey.

STATISTICAL REPORTS

They present comprehensive results of individual survey or of several surveys from the same field. Each report contains methodological explanations in which the sources of data collection, definitions and explanations as well as comparability, tabular presentation of data in series for the last year and presentation of published publications are given. Data are presented for the Republic of Croatia in total, and in some reports by counties, cities and municipalities.

METHODOLOGICAL GUIDELINES

Methodological guidelines contain data on sources and methods of data collection and on the scope and definitions of statistical surveys. They are intended for all those who conduct statistical surveys and for users to better understand the data.

STUDIES AND ANALYSIS

This series consists of authors' works that are the result of the application of statistical methods and analyses from individual statistical surveys. The authors are employees of the Croatian Bureau of Statistics and/or external associates.

SPECIAL EDITIONS

These editions are especially important for the further development of statistics, and are mainly intended for the professional circle of users. Issues important for the development of methodology and statistical standards comparable to the statistical standards of the UN, Eurostat and other international organisations are published periodically. Some of these editions are translations.

PUBLISHING PROGRAMME

This catalogue of about 120 pages is a guide through the titles of all publications of the Croatian Bureau of Statistics planned for publication during the current and until June next year. The ordinal number and title of most publications are accompanied by a brief content description, deadline and periodicity of publishing, data presentation level, language(s) and format, and the medium in which the publication is published. After the part of the Publishing Programme that announces the publishing deadlines there is a list of previously published publications by statistical areas. The prices of publications and services of the Croatian Bureau of Statistics are listed at the end of the Publishing Programme. The Publishing Programme is free of charge.

CALENDAR OF STATISTICAL DATA ISSUES

This publication is directly related to the Publishing Programme. About 65 pages of the Calendar of Statistical Data Issues containg the exact date, day and week of publishing first releases, the period to which the data relate and the level of presentation. For other statistical publications, only the months in which they will be published are listed. It is intended for all users of publications of the Croatian Bureau of Statistics.

CARTOGRAPHIC PORTAL

The Croatian Bureau of Statistics has developed the GeoSTAT application, which enables the cartographic presentation of statistical data in such a way that the selected spatial level is combined in great detail with the selected statistical data.

Specifically, users can monitor demographic data from the 2011 Census, even at the micro level of cells of 1 km^2. Further on, the data of the Register of Business Entities (all legal organisational forms) and data from the field of tourism (accommodation capacities, tourist arrivals and overnight stays) can also be monitored at this micro level. This facilitates the monitoring of available statistical indicators, except in the standard form, through first releases and publications and interactive maps on which all interested parties can spatially monitor the movements of individual statistical indicators. Also, data from the fields of agriculture, national accounts, employment, education, industry, construction, trade, at-risk-of- poverty distribution and business demographics are available on the GeoSTAT portal. In addition, all data on the GeoSTAT portal are also visible in the Metadata Catalogue of the Croatian Bureau of Statistics, which was developed as part of the National Spatial Data Infrastructure. The Metadata Catalogue was developed according to the requirements of INSPIRE.

DATABASE

Sectoral classification of institutional units

Institutional units in the Republic of Croatia are classified into sectors and subsectors in accordance with European regulations for the purpose of compiling official statistics.

The European System of National and Regional Accounts (ESA 2010) is the latest accounting framework for achieving international harmonisation of national accounts, which, among other things, prescribes a methodology for classifying institutional units into sectors and subsectors.

ESA 2010 was published on 26 June 2013 in the Official Journal of the European Union as Annex A to Regulation No. 549/2013.

Institutional units classified in accordance with the ESA 2010 methodology can be searched and downloaded in CSV format by clicking on the link https://web.dzs.hr/sektorizacija.htm.

Data on business entities (NKD 2007.)

Every business entity entered in the Register of Business Entities can see the basic data kept about it and about a part of the business entity on the website of the Croatian Bureau of Statistics with the help of its registration number.

STS database

Business Statistics Database or STS Database (Short-Term Statistics) is a relational database of consolidated data on indices in the field of short-term business statistics. The STS application allows users to search all the data in one place in the database. All data stored in the STS database can be viewed in tabular and graphical form in the application.

PC AXIS database[11]

- Agriculture, hunting, forestry and fishing
- Industry
- Foreign trade in goods
- Retail trade
- Tourism
- Transport and communications
- Environment
- Employment and wages
- Prices
- Personal consumption and poverty indicators
- Structural business statistics
- Criminal justice
- Population
- Census 2011
- Agricultural Census 2003
- Agricultural Census 2020
- Settlements and population of the Republic of Croatia, 1857 – 2001
- Subnational statistics

The Croatian Bureau of Statistics provides the following services to all users of statistical data:

- solving user requests
- providing basic information on data available on the Eurostat website via the Eurostat portal – European Statistical Data Support (ESDS) EUROSTAT at https://ec.europa.eu/eurostat/web/main/search/-/search/estatsearchportlet_WAR_estatsearchportlet_INSTANCE_bHVzuvn1SZ8J?p_auth=bQ6G7SQg&text=data+support
- use of library materials in the reading room.

An important task of the Croatian Bureau of Statistics is to store statistical data for further use. The Croatian Bureau of Statistics maintains aggregate statistical data in both electronic and printed form. The storage of aggregated data in electronic form takes place as part of the implementation of the policy of backup and archiving of electronic data and in the form of monthly archiving of the website of the Croatian Bureau of Statistics https://dzs.gov.hr/en.

All publications published by the Croatian Bureau of Statistics are kept in it in at least one copy. Printed copies of publications are submitted in two copies to the NSK archives. Data on publications are entered into the catalogue of publications via the NSK digital system.

The Croatian Bureau of Statistics allows research agencies and independent researchers to use confidential statistical data to conduct statistical analyses exclusively for scientific purposes.

The storage of statistical microdata in electronic form takes place as part of the implementation of the policy of backing up and archiving electronic data.

[11] PC AXIS was initially developed for the 1990 Census in Sweden. Further development is led by an international reference group in charge of PC AXIS with participants from other countries licensed for PC-AXIS used by national statistical offices. Close co-operation between the Swedish and Danish statistical offices is particularly well known.

7.1. Design and production of dissemination products

Description

Publishing of statistical data and information complies with standardised procedures in different technologies. Standardised procedures are based on pre-prescribed structures, formats and metadata and are used in the preparation of tables in the statistical processing phase. All procedures are subject to the same, standardised working procedures in accordance with the principles of availability and clarity of data, taking into account timeliness. Well-prepared, structured and organised documentation, internal exchange of knowledge, well -established standard communication channels and archiving of documentation and working procedures are important when preparing new content and regularly updating output data.

For each form of publishing statistical data and information, the output must first be updated. It can be data in the form of databases, or tables with the final consolidated data ready for publication.

When it comes to the quality of audited data, it is necessary to follow the instructions of the Croatian Bureau of Statistics on the revision of statistical data. If a revision of the statistical data is prepared, the statistical data need to be updated and published depending on the statistical area in accordance with the set deadlines.

The updating of output data varies depending on the type of publishing media or subprocess type:

- adding a new time dimension to database files of the Croatian Bureau of Statistics
- adding a new time dimension to the macrobasis of the Croatian Bureau of Statistics for foreign trade in goods data
- data prepared by the Croatian Bureau of Statistics for transmission to Eurostat
- preparing revised data for interactive tools.

Quality assurance instructions

When updating the results, standardised procedures should be complied with and implemented in a transparent manner. It would be good to automate processes that repeat at equal intervals.

Information on data transfer must be available (in the Calendar of Statistical Data Issues or directly/in person). Documentation must be edited, published, accessible and up to date.

As several people from different fields are involved in this business process, internal coordination must be documented.

Published methodological guidelines that are no longer current should be replaced by a new version, while the old version should be archived.

All rules relating to the revision of statistical data must be complied with.

7.2. Managing publication of dissemination products

Description

It includes informing specific groups, such as media representatives or state administration bodies, as well as dealing with possible bans before publishing and providing data to subscribers.

It also includes managing access to confidential statistical data to conduct statistical analyses exclusively for scientific purposes by research agencies and independent researchers.

The Statistical System of the Republic of Croatia produces impartial statistical data on social and economic processes, providing the factual basis necessary for monitoring and analysing the state of the Croatian economy and directing policies related to the development of the Croatian society and economy and EU policies. By

continuously promoting publishing of data, the Croatian Bureau of Statistics informs users about the importance of statistics and encourages reporting units to deliver data without hindrance.

Each publication of statistical data must be announced in accordance with the publication plan (the Calendar of Statistical Data Issues). The Calendar of Statistical Data Issues is prepared at the end of the year, 18 months in advance, depending on the completion of the processing of individual statistical surveys and the possibilities of producing a particular medium of publication.

Once a year, the Publishing Programme is published (in Croatian and English) so that data users are fully informed about the media for publishing statistical data of the Croatian Bureau of Statistics. The Programme is free of charge and can be found on the website of the Croatian Bureau of Statistics. At the beginning of the Publishing Programme, users can find information about the types of publications that are published.

Along with the list of publishing media, a short description is given which contains the type of publishing media, ordinal number and title, data presentation level, deadline and periodicity of publishing as well as the language(s) in which it is published. At the end of the Publishing Programme, users can find detailed information on how to obtain statistical information and data, what are the possibilities of purchase, complaint, payment, delivery and information on the prices of publications and services.

Only data already published can be forwarded to international organisations or Eurostat. If confidential data are also provided to Eurostat, confidential cells must be marked, and the same content at a higher level of aggregation must be previously or simultaneously published on the website of the Croatian Bureau of Statistics.

Data publishing procedures vary depending on the type of publishing medium: publishing news on the web page; publishing printed publications; publishing data in the database of the Croatian Bureau of Statistics; transmission of data to Eurostat.

Error correction is a procedure related to publishing of statistical data. The purpose of correcting errors in published statistical data and information is to provide accurate and high-quality statistics and information to users.

Detailed procedures are prescribed in the guidelines for correcting errors in published statistical data and information.

A good example from practice is given in Methodological Guidelines No. 66, Monthly Survey on Persons in Employment and Wages (RAD-1 form) on page 53 of Chapter 5.5. Guidelines for correcting errors at the link https://podaci.dzs.hr/media/qf4jd3xw/metod_66.pdf.

Contents to be published should be prepared for the needs of various user profiles. The presentation of data should be applicable, understandable, clear, interesting and prepared bilingually in Croatian and English.

General principles for presenting data vary depending on the type and medium of publishing.

When presenting data, it should be kept in mind that not all users are equally statistically literate and do not have the same prior knowledge regarding the use of statistics and information. The professional public mainly uses detailed data for the purpose of further data analysis, so it is advisable to publish such data in databases in formats that allow further processing (electronic form). The general public is interested in the most important or most interesting general statistics and information published in a clear and understandable way. The author of the data presentation must consider who the target audience is and what they want to know, and accordingly the statistics should be presented in an understandable, interesting and useful way. Published data should contain commentary and visualisation of the data in an understandable and clear manner. The comment should be applicable, short, simple, clear, understandable and interesting. Data visualisation should be accompanied by simple tables and graphs. When presenting data, it is necessary to follow the recommendations given in the Guidelines for Formatting and Standardising Publications[12].

[12] Upute za uobličavanje i ujednačavanje publikacija, CBS, Zagreb, 2012. (link: Upute za uobličavanje).

Tables are an integral part of statistical publishing media because they are useful for displaying larger amounts of data. With the help of tables, it is possible to present any phenomenon. They are suitable for displaying absolute data and relative numbers (coefficients, indices, averages, structures). Summary tables are suitable for publishing in printed publications and news on the website, while detailed data are published in tables in the database of the Croatian Bureau of Statistics.

Graphs are graphical presentations of statistical phenomena in the form of columns, lines, pies, etc. The characteristics of phenomena and data determine which type of graph is most suitable for displaying data. Graphs can display information more efficiently and comprehensively than comments, while a table as a medium has the ability to display a larger amount of data in the form of absolute and/or relative numbers.

Graphically presented statistical data are more understandable and clear in relation to their presentation in a table, while greater visibility of the graphical presentation and the strength of the first visual impression of the characteristics of the observed phenomenon are its advantages.

Maps are the most appropriate tool for visualising phenomena in the spatial dimension. They allow the comparison of spatial units of different sizes and display a large amount of data in a vivid way. Statistical maps usually show relative numbers, especially indicators per 1 000 inhabitants and calculations at different levels of spatial units.

Preparations for the presentation of statistical results differ depending on the medium of publication and in what form the results will be presented, so we distinguish the following:

- preparation of information for publication on the website
- preparation of a printed publication authored by a holder of statistical survey
- preparation of a printed publication prepared by the editor of the publication
- preparation of new content in the database of the Croatian Bureau of Statistics.

Access to confidential statistical data for the implementation of statistical analyses for scientific purposes may be provided only on the basis of a written request of the user, in a manner and under conditions clearly defined by the Ordinance on the Conditions and Terms of Using Confidential Statistical Data for Scientific Purposes (Official Gazette, No. 137/13). Coordination of resolving requests for access to confidential statistical data and access to data should be defined by internal procedures and rules that will enable quality management of the entire process.

Quality assurance instructions

According to the date of publication, which is indicated in publishing programmes and calendars of statistical data issues, publications are available at exactly 11 am, in electronic and printed form, thus following the Code, which requires timely dissemination, or the exact time of publication.

In order to meet this condition, it is necessary to design the entire statistical survey procedure so that the data are published in a timely manner. Regardless of the fact that the statistics are accurate and detailed, they will not be useful if they are not published in a timely manner and within the deadlines for publication.

The data must be available to all users in the same way. Users should also have as many simple options as possible to make additional requests and order data and services.

Ways of publishing data should be clear to users. All relevant metadata, such as methodological explanations, basic quality indicators, questionnaires, etc. should be available in accordance with the structure of the publishing programme on the website in the area of quality. Also, it is necessary to explain to users in a simple way the procedures of error correction and data revision.

Documentation on rules, instructions and recommendations for the presentation and publishing of data needs to be prepared and made available. The way in which statistical data are presented and writing of unambiguous comments, as well as the production of tables, graphs and maps in the media, should be in line with good practice.

The author, the holder of the statistical survey, the editor of the publication and the organisational unit in charge of disseminating statistical data and related methodologies should participate in the preparation of data for publishing.

The templates are designed according to the content, data preparation technology, technical capabilities and tools of the Croatian Bureau of Statistics and the adopted publishing standards.

The following principles should be followed when presenting data: relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, as well as accessibility and clarity of data. When preparing comments on changes in the phenomenon, the observation period and the comparison period must be indicated. Translators and proofreaders should be engaged in the preparation of texts for publication.

It is necessary to clearly define the terms and methods of access under which confidential statistical data can be used to conduct statistical analyses for scientific purposes, and to define internal procedures and rules for resolving such requests.

7.3. Promotion of dissemination products

Description

The field of dissemination and communication of official statistics is extremely important for bringing statistical data closer to users in a simpler and more understandable way, i.e. by popularising official statistics.

It is important to make the promotion of statistical products prepared during the production process available through different communication channels and in different ways to a wide range of users.

It includes, inter alia, the use of social networks as an important communication channel for disseminating statistical data, holding user councils with specific target groups of users to get first-hand useful information about the needs of statistical users; holding educational workshops for students on whether statistics can be interesting and what is its importance in modern times.

Statistical literacy can be promoted by organising statistical competitions to stimulate the interest of high school students and teachers in statistics. It can also show high school students what possibilities the application of acquired statistical knowledge can open by creating interactive statistical portals which, in addition to theoretical part and explanations, contain examples of solved assignments as well as assignments for individual exercise of acquired statistical knowledge. Moreover, it includes interactive presentations and the production of brochures that explain in a simple way frequently used, but insufficiently understood statistical concepts.

In addition, important components are a functional and modernised website as the main source of statistical data and a channel for dissemination of statistical products and communication with media representatives. Interesting articles prepared in advance on the occasion of marking various international, world and European days and sending them to media representatives may result in a number of media releases mentioning the Croatian Bureau of Statistics and its scope of work.

Instructions for quality assurance

In addition to achieving a higher degree of customer satisfaction with the use of modern information technologies and reducing the need for printed publications and statistical forms and instructions, further rationalisation of operating costs is achieved.

It is necessary to actively use social networks as an important communication channel for the dissemination of statistical data (Facebook, Twitter, Instagram, YouTube), enabling proactive communication with existing and future users and strengthening the image and reputation of the institution. Proactive work should be done on development activities aimed at popularising official statistics and bringing statistical data closer to users in a simpler and more understandable way.

It is necessary to continuously modernise communication channels, primarily the website of the Croatian Bureau of Statistics, with an increased coverage of published statistical data that are available in formats tailored to user needs.

Intensive attention should continue to be paid to working with specific target groups of users, holding user councils and educational workshops.

Promotional activities should be carried out continuously in order to increase the response rate of respondents to surveys conducted by the Croatian Bureau of Statistics. using all communication channels (press releases, social networks, etc.), but also to modernise leaflets, announcement letters and correspondence sent to reporting units.

7.4. Customer Relationship Management

Description

Taking into account the protection of statistical confidentiality, users should be provided with timely access to statistical data resulting from the implementation of statistical surveys. Customer support is provided by e-mail, telephone, personal contact and through social networks.

Customer support includes the provision of statistical data and information, advice on access to data as well as assistance in finding statistical data and information on the website of the Croatian Bureau of Statistics.

Publicly available statistical data and information are available to all users in accordance with the terms of use. Preparation of data at the request of the user is charged according to the valid price list of the Croatian Bureau of Statistics.

Records of written requests for data, including requests for access to confidential statistical data for scientific purposes, shall be kept in the organisational unit responsible for user relations. It is used to make an analysis of user requirements, based on which the structure of users is determined, ways of communicating with them and which are the most sought-after statistical data by statistical areas.

Users are also provided with access to confidential statistical data for the implementation of statistical analyses exclusively for scientific purposes, but under special conditions and on the basis of a concluded contract. Access to confidential statistical data for scientific purposes is determined by the Ordinance on the Conditions and Terms of Using Confidential Statistical Data for Scientific Purposes[13].

International reporting and compliance with official statistics includes, inter alia, the transfer of data from other bodies authorised by the Official Statistics Act and the Programme of Statistical Activities of the Republic of Croatia to perform official statistics, the so-called producers of official statistics.

Ways of customer support can vary depending on the type and form of requests, as follows:

- written request for data
- telephone inquiry regarding data
- access to confidential statistical data for scientific purposes
- international reporting
- user access to library materials.

[13] Link: http://www.propisi.hr/print.php?id=12692

Quality assurance instructions

The information provided to users should be accurate, relevant and unambiguous. Users should receive information in a timely manner, and in cases of delayed delivery should be notified in advance. All statistical information is free of charge, while statistical data processing services at the users' request are charged, i.e. the costs incurred are borne by the user who requested such a service, in accordance with the Price List of Publications and Services of the Croatian Bureau of Statistics.

The Croatian Bureau of Statistics is entitled to reimbursement of actual material costs incurred by providing information to the user of the right to access information and reuse information, as well as to reimbursement of costs of delivery of requested information, which is charged in accordance with the Criteria for Determining the Amount of Reimbursement and Manner of Paying for the Reimbursement.

Information on the availability of statistical data is published on the website of the Croatian Bureau of Statistics and in publications. The organisational unit in charge of customer relations continuously provides customer support during working hours.

Users should have free access to online data, printed publications and library documentation during business hours.

Basic information on access to confidential statistical data should be published on the website of the Croatian Bureau of Statistics.

Concern for quality in international reporting is based on standardised work instructions, transparency, proper documentation, working documents published on the intranet and in the internal archives of the Croatian Bureau of Statistics.

8. Evaluation

The priority is the quality of statistical surveys and data. In order to ensure a common framework for the quality of statistics, the Code of Practice was prepared, which was first adopted by the SPC in 2005. The Code was revised by the ESSC in 2011 and 2017, and was last revised on 16 November 2017. The revised Code has 16 principles concerning the institutional environment, statistical processes and statistical output.

The aim of the Code is to ensure that statistics produced within the ESS are relevant, timely and accurate as well as in line with the principles of professional independence, impartiality and objectivity. A set of best practice indicators and standards for each principle provides guidance and reference for evaluating the application of the Code. With the adoption of the Code, the statistical offices of the Member States of the European Union and Eurostat have committed themselves to implementing activities that will ensure high quality statistics.

The collection of data on the quality of statistical data takes place systematically during the implementation of all statistical processes. The prepared documentation on the implementation of statistical surveys at all stages helps to identify possible systemic errors in individual subprocesses. With the help of such data it is possible to assess the quality of statistical data and critically evaluate their calculation. In addition, they are important for users as they gain additional insight into the process of data collection, processing and dissemination. Publishing data on their quality is a transparent way of informing users about various aspects of statistics. Based on the collected data, an improvement plan is prepared as needed and changes are introduced in the implementation procedure in accordance with the set plan.

8.1. Gathering information for evaluation

Description

The process of conducting statistical survey includes subprocesses that may affect the quality of statistical results. In order to better assess the quality of statistical results, as much data as possible on each subprocess should be collected.

Data collected for quality assessment can be divided into two groups. The first group includes information derived directly from statistical processes and subprocesses, while the second group consists of data taken from the results of statistical surveys. Particular attention should be paid here to customer satisfaction, information on new or changed requirements or the inapplicability of published results to customer needs.

With regard to information collected directly from the statistical process, special attention should be paid to quantitative information, i.e. information that we call quality indicators. Two typical examples of such indicators are sample error and non-response rate. These indicators are important because they enable an empirical and analytical approach to monitoring and ensuring the quality of statistical products and processes. The calculation of these indicators needs to be included in the statistical process, while at a later stage it is necessary to ensure that the values of the indicators are regularly analysed and published in quality reports.

Quality assurance instructions

Both statistical processes and all possible sources of information that could be used to assess quality should be carefully analysed. The list of standard quality indicators should be consulted when determining the quality indicators to be calculated and subsequently included in the quality reports. If necessary, indicators specific to a particular statistical area (e.g. mirror statistics for a foreign trade of goods or tourism domains) should be added to the list.

In preparing instructions for the calculation of standard quality indicators, the link https://podaci.dzs.hr/media/1cmjru1a/metod_73.pdf should be strictly adhered to. All available data sources should be used to retrieve information and data from users. In order for the information to be useful for further analysis, it needs to be adjusted and included in the quality reports. It is also necessary to ensure that all information is available as soon as possible, as this is the only way to make its use useful and effective.

8.2. Evaluation of results

Description

Evaluation material can be derived from any phase or subprocess. The process of preparing documentation for statistical survey includes a detailed description of statistical activity, starting with the description of terms, definitions, applied methods, production process, information system and all the way to work instructions. The quality of the statistical survey will be set depending on the extent the documentation for the implementation of the statistical survey is well prepared. Statistical survey documentation is an important tool in communication between different participants in the process of organising and conducting statistical surveys as well as between producers and users of statistical data. Survey documentation is part of the metadata.

In general, documentation can be divided into documentation for users of statistical survey results and documentation for the implementation of statistical surveys. The first type of documentation describes and documents statistical results and is usually published in public. The second type of documentation describes the statistical procedures and processes used during statistical processes and subprocesses. This documentation is created during the organisation and implementation of individual processes and subprocesses for statistical surveys and is mainly intended for internal use.

The purpose of user documentation is to help users better understand what data and statistical methods represent. Quality preparation of user documentation enables users to more easily understand, find and process data.

Examples of the content of user documentation are survey questionnaires, presentation leaflets, methodological explanations and reports on the quality of statistical surveys (https://dzs.gov.hr/highlighted-themes/quality/quality-reporting/quality-reports-by-statistical-domains/892).

The purpose of the documentation for the holders of statistical surveys is to follow the principles of relevance, accuracy, timeliness and punctuality, coherence and comparability, accessibility and clarity of data. The documentation for the holders of statistical surveys should describe in detail the individual procedures used in conducting the survey (e.g. determination of the target population, sample selection, design of the questionnaire, editing of data, publishing of data, etc.). Descriptions should include a report on what was done at each step and an explanation of why a particular implementation method was chosen. All this is useful information for the development and improvement of processes and subprocesses of well-established statistical surveys and/or for planning new surveys. Sometimes some texts from such documentation are interesting to external data users, so it is sometimes necessary to publish it, depending on the need.

The Croatian Bureau of Statistics keeps statistical microdata in electronic form for further use in the Croatian Bureau of Statistics and for use in research and analytical purposes. The archiving of statistical microdata in electronic form takes place as part of the implementation of the policy of backup and storage of electronic data.

The backup is divided into a database backup and a backup of different servers. All databases of the Croatian Bureau of Statistics that contain data from statistical surveys and metadata are permanent values and subject to a unique backup system. Backups are made only for data that are improved daily. Data are stored in a secure media storage space.

Statistical microdata are stored together with files and metadata in electronic form on servers intended for it.

The Croatian Bureau of Statistics archives aggregate statistical data for further use in both electronic and printed form. Archiving of aggregated data in electronic form takes place as part of the implementation of the policy of backup and storage of electronic data and in the form of monthly archiving of the web portal of the Croatian Bureau of Statistics https://dzs.gov.hr/en.

Aggregate statistical data in printed form may be published in publications issued by the Croatian Bureau of Statistics or in publications taken over by the Croatian Bureau of Statistics from other institutions. All publications are kept in the library archives of the Croatian Bureau of Statistics. Publications published by the Croatian Bureau of Statistics are archived in at least one copy.

Quality assurance instructions

Survey documentation must be accurate, comprehensive and understandable to the target audience for which it is intended. Standard forms and templates should be used when preparing the documentation. The content of individual processes should follow the standard structure as much as possible. To avoid ambiguities and misunderstandings, user documentation should be prepared in simple language.

The documentation for users should include clarity and repeatability of the procedures and processes applied, in particular descriptions of any deviations from standard procedures. The content and level of detail of the documentation should be tailored to the target audience.

In case of identified errors, the documentation of all processes should contain descriptions of the procedures to be followed and the names of the persons or organisational units to be contacted in case of ambiguity.

Backup and storage should be performed in accordance with the instructions for backing up and storing.

Business process owners are responsible in their organisational units for the systematic supervision of information risk management and for the harmonisation of necessary activities with the prescribed procedures. Checks of these activities are performed once a year or as needed.

8.3. Improvement Action Plan

Description

The improvement action plan should include an evaluation of the production process with proposals for its improvement based on the quality report. The data collected for the quality assessment need to be analysed and a report prepared containing the most identified critical points in the process or in the broader context of the implementation of the statistical survey. Based on the conducted analysis of the process, it is necessary to submit proposals for the introduction of improvements in the implementation of statistical surveys in the next programming period.

In statistical surveys whose organisation and implementation take place from time to time, special attention should be paid to value indicators of quality that need to be compared with previous periods. If larger differences are found compared to the values from previous periods, it is necessary to analyse the reasons for such differences and prepare an action plan that will improve the quality of the statistical area whose quality is being measured.

Based on the collected and prepared information, a quality report is prepared, which includes descriptions of all dimensions of quality, together with value indicators of quality.

Special attention should be paid to the analysis of information downloaded from data users. The summary report should contain descriptions of all identified unfulfilled user requirements and anticipate which of the unfulfilled requirements could be met when conducting the statistical survey in the next programming period.

The collected quality information and its analysis is used to prepare an improvement plan. If the main shortcomings in the collected information on the course of organising and conducting the survey are identified, an action plan for improvement should be prepared and, based on that plan, the necessary improvements should be introduced into the procedure.

Quality assurance instructions

The data collected should be analysed in a professional and impartial manner. Deficiencies need to be identified in the analysis. The results of the quality analysis should be presented in simple language that will be understandable even to someone who is not directly involved in the survey in question.

In the analysis of quality indicators, special attention should be paid to the comparison with quality indicators from previous periods or with other statistical surveys. The quality report is prepared and published after the publication of statistics. All collected user responses should be reviewed and analysed and given a critical review. It is important in the analysis phase to highlight requirements that may be included in the improvement plans for the next programming period. A comprehensive overview of the improvements that can be made to the survey process should be an integral part of the programme plan.

LITERATURE

1. Cox, B. G. et al. (1995). Business Survey Methods. New York: Wiley.
2. Sarndal, C.E., Svensson, B., Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Werlag.
3. Cochran, W. G. (1963). Sampling Techniques. London: John Wiley and Sons, Inc.
4. Design your questions right. How to develop, test, evaluate and improve questionnaires. (2004). Stockholm: Statistics Sweden. Accessed on 19 June 2012 on the website: https://share.scb.se/ov9993/data/publikationer/statistik/_publikationer/ov9999_2004a01_br_x97op0402.pdf
5. Dillman, D. A. (2003). Mail and Internet Surveys: The Tailored Design Method. London: John Wiley and Sons, Inc.
6. Fellegi, I.P., Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association. 71(353), 17-35.
7. Granquist, L. (1991). Macro Editing – A Review of Some Methods for Rationalizing the Editing of Survey Data. Statistical Journal, 8, 137-145.
8. Hidiroglou, M.A., J.M. Berthelot, (1986). Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, 12, 73-83.
9. Cochran, W. G. (1963). Sampling Techniques. London: John Wiley & Sons, Inc.
10. Levy, P. S. Lemeshow, S. (2004). Sampling of Populations, Methods and Applications. New York: Wiley.
11. Presser, S., Rothgeb, J. M., Couper, M., Lesser, J., Martin, E., Martin, J., Singer, E. (2004). Methods for Testing and Evaluating Survey Questionnaires. New York: Wiley.
12. Quality Guidelines for Official Statistics. (2007). Helsinki: Statistics Finland. Accessed on 19 June 2012 on the website: http://www.stat.fi/org/periaatteet/qg_2ed_en.pdf
13. Sarndal, C. E. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.
14. Sirken, M. G., Herrmann, D. J., Schechter, S., Schwarz, N., Tanur, J. M., Tourangeau, R. (1999). Cognition and Survey Research. London: John Wiley & Sons, Inc.
15. Snijkers, G. (1999). Cognitive Laboratory Experiences: On Pre-testing Computerised Questionnaires and Data Quality. Deen Haag: Statistics Netherlands.
16. Snijkers, G.J.M.E. (2002). Cognitive Laboratory Experiences: On Pre-testing Computerised Questionnaires and Data Quality. Utrecht: University Utrecht.
17. Thompson, S. K. (1992). Sampling. London: John Wiley & Sons, Inc.
18. Survey Methods and Practices. (2003). Ottawa: Statistics Canada.
19. Quality Guidelines. (2003). Ottawa: Statistics Canada.

20. Quality Guidelines for Official Statistics. (2007). Helsinki: Statistics Finland.
21. Tourangeau, R., Rips, L. J., Rasinski, K. (2000). The Psychology of Survey Response. Cambridge: Cambridge University Press.
22. Wallgren A., Wallgren B. (2007). Register-based Statistics; Administrative Data for Statistical Purposes. London: John Wiley & sons.
23. Cochran, W.G. (1977): Sampling Techniques. London: John Wiley & Sons, Inc.
24. ESS guidelines on seasonal adjustment, 2015 edition, Eurostat
25. Handbook on Seasonal Adjustment, 2018 edition, Eurostat
26. Seasonal Adjustment | CROS (europa.eu)

https://www.idsurvey.com/en/methods-web-survey-cawi/

https://dimewiki.worldbank.org/wiki/Computer-Assisted_Personal_Interviews_(CAPI)

https://www.idsurvey.com/en/

https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=252469