What are official statistics and how are they produced?

What are official statistics?

Official statistics refer to public information which is produced for the benefit of the society and is funded by the state budget under the official or European Union statistical programme. Official statistics are equally accessible to everyone and enable the consumers to make the necessary decisions in their private or business lives. Official statistics comply with international classifications and methodologies and meet the principles of impartiality, reliability, relevance, cost-effectiveness, confidentiality and clarity. European Union official statistics are regulated by the quality criteria established in the European Statistics Code of Practice.

There are two producers of official statistics in Estonia – Statistics Estonia and Eesti Pank (central bank of Estonia).

At first there is a question

Humans have always been curious; this constitutes the basis for the development of the mind and rational behaviour. A great number of questions start with the words "how many". Nowadays such questions are often addressed to Statistics Estonia.

- How many children will start school in county N next year?
- How many households with partners who are not officially married are there?
- How many hours do people in Estonia sleep per night?
- What kind of production is exported/imported in Estonia and how much?
- How much do people earn each month, on average, in Valga county, Ida-Viru county, in Estonia as a whole?

In the case of a question of great public interest, it needs to be determined where to find the answer. Very often the answer is available free of charge in Statistics Estonia's Statistical Database. The answers to all of the above-mentioned questions can also be found there. Statistics Estonia regularly collects feedback from the consumers about the information the society needs. The main consumers of statistics include state authorities, professional associations, research and educational institutions, and local governments. If a question is vital for a great number of people and stakeholders and there are no data on it, the issue needs to be examined and the corresponding data need to be collected – an already existing statistical action needs to be expanded or an entirely new statistical action needs to be launched.

Survey preparation (phase 1 of the Generic Statistical Business Process Model: specify needs; phases 2 and 3: design and build)

In the European Union, there are a number of surveys which are carried out more or less simultaneously on the basis of a common programme and methodology. Such surveys allow discovering common trends and developments in people's lives and determining the idiosyncrasies of European countries, comparing and ranking the countries. Estonia also organises several surveys which are held in the European Union on a uniform basis, e.g. the Labour Force Survey, Information Technology in Enterprises, and the Social Survey.

In addition to periodic surveys, new ones are also started. For example, in co-operation with the Ministry of Social Affairs, Statistics Estonia organises the Working Life Survey. The different sides of working life are examined from the point of view of both employers and employees; satisfaction with the job and the working conditions are also studied.

Nowadays it often happens that the information that is of interest has already been partially or completely collected into a state database. Upon the production of statistics, the aim is to use the already existing data, including the information in state databases, to the greatest extent possible. The data of the Population Register, Commercial Register, Estonian Education Information System, State Register of Construction Works and several other databases are constantly used for statistical purposes.

Using the information generated in automatic processes (e.g. mobile positioning data, social network data, satellite images) in the production of statistics is attracting more and more interest. Due to their large volume, such datasets are called big data. At the moment, one of the examples of using big data is tourism statistics which are calculated based on mobile positioning data and are available on the website of Eesti Pank.

If the examined characteristics include estimations (e.g. satisfaction) or information which is not available in any of the databases, people or economic units will inevitably have to be interviewed.

Survey preparation typically consists of the following closely related stages:

- formulating the aims;
- determining the sources – to what extent the existing (register) data or the data of previous or upcoming surveys can be used;
- developing the questionnaire;
- determining the method of data collection;
- sample formulation;
- developing a data processing program;
- determining the form of output;
- assessing the schedule and resources (money and labour force);

The data to be acquired (survey questions) need to be covered in the questionnaire and this needs to be taken into account in both data processing and output design. One of the most limiting factors in survey planning, however, is resources – each question which is added to the questionnaire and each unit entered in the sample increases the cost of the survey and the response burden.

Questionnaire compilation

Questionnaire compilation is one of the most labour-intensive stages of survey preparation. The most important principles underlying the compilation of questionnaires are as follows:

- There need to be as many questions as necessary and as few questions as possible.
- Each question is to be based on a survey question and reflected in the output, otherwise the question is superfluous.
- If the questionnaire is based on one developed in another country, it needs to be checked whether each question fits into our environment.
- The wording of the questions and multiple choice answers needs to be simple and clear (as close to actual language use as possible, but still correct).
- Adding a number of questions with no pre-set answers in the questionnaire is not practical because processing and analysing such questions is very labour-intensive; however, inserting these questions often provides important additional information.
- The accuracy and reliability of survey results do not depend much on the length of the scale used (5, 7, 10, 100 graduations), therefore it is preferable to use as short a scale as possible. It is also advisable to use only one scale in the questionnaire.

Choosing the method of data collection

Upon choosing the method of data collection, it needs to be taken into account who the data are collected from and what kind of information is sought.

The ways of interviewing people have improved over time, but so far the **classical face-to-face interviews** have not been cast aside. Nowadays they are just conducted with the help of a laptop – the interviewer enters the answers into the computer during the interview. Since face-to-face interviews are time-consuming and expensive, an effort is often made to get by without organising them. What is becoming increasingly popular are

telephone interviews, in the case of which the interviewer and the interviewee do not meet, and **online surveys**, which entail completing the questionnaire form independently online. The best results are generally achieved through a combination of several methods. For example, the participants in the European Health Interview Survey had the option of filling out an online questionnaire within one month. Those who did not choose to do so were visited by an interviewer in order to conduct a face-to-face interview.

Most of the surveys among economic units are conducted as online interviews because this method of data collection is the most convenient one for the respondents. Often the questionnaire presented to economic units contains questions which can be answered after looking through the information systems or paper documents of the enterprise, or the relevant knowledge is shared between several employees. The flexible use of time is one of the biggest advantages of online interviews over other ways of data collection.

Total population and sample (phase 4 of Generic Statistical Business Process Model: collect)

Upon defining the objective of the survey, the total population, i.e. the number of persons or items which are the focus of estimations, is determined very precisely. In a sample survey, all of the items in the total population are not studied, but only a certain share is selected – a sample, which will serve as the basis for estimations regarding the total population. Thus, each person or economic unit in the sample presents a vast number of similar persons or economic units. This fact places great responsibility on those forming the sample because if they fail to respond, the person using the survey results may reach wrong conclusions or may be unable to arrive at one altogether. Drawing the sample relies on the theory of survey sampling, which is based on the probability theory. The fact that this method yields reliable results was ascertained by statisticians a hundred years ago.

Before sampling, the size or volume of the sample is determined. The volume of the sample is determined taking into account the aims of the survey – if the subsamples are too small, it is likely that the posed question cannot be answered with sufficient certainty and using the existing resources.

When the preparations for the survey are complete, data collection may start. Data collection is the most expensive and time-consuming stage of the production of statistics. If data collection is unsuccessful because the survey participants fail to respond or their responses are illogical, the entire survey is going to fail because without reliable data there will be no reliable results. In order to reduce non-response, letters of reminder are sent to survey participants; in the case of technical problems, Statistics Estonia's Contact Centre is ready to help. To avoid errors in data entry, checks have been added to the questionnaires to draw the respondent's attention to contradictory information and possible mistakes already during the completion of the questionnaire.

Data processing (phase 5 of Generic Statistical Business Process Model: process)

The collected data are still not ready to be used for analysing and drawing conclusions. Preparing the data for analysis has become significantly faster thanks to developments in technology and data collection programs having been supplemented with checks which do not allow entering contradictory answers (e.g. a 17-year-old cannot have tertiary education). Data coding can also be done electronically to a large extent.

All this does not mean that electronically collected data are always correct to the minutest detail – each person who has had experience with data knows that errors may occur in texts as well as datasets. Such hidden errors are discovered upon the application of more complex checks and comparing with other sources, but it also happens that an error goes unnoticed or is detected a while later. A large share of the data processing work is nowadays done by computers, but all the checks and models which the computers run on the data still have to be developed and programmed by people. Thus, the increased capacity of data processing has simplified our work on the one hand (computers are able to process bigger volumes of data and do it faster); on the other hand, it has also provided the opportunity of creating even more numerous and complex checks and models (made the work of a person engaging in data processing more complicated and exciting).

Imputation

One major problem that occurs with datasets is data gaps, which disturb data processing especially if the application of more complex models is desired. In order to replace missing data with substituted values, several methods of **imputation** are used with the main idea being that the missing value of a certain item is replaced with a value of the same characteristic for items that are as similar as possible. It must be noted here that although imputation does not basically add any new information to the dataset, it makes the dataset more workable.

Expanding data to total population

Although an analysis is based on sample data, inferences must be made about the total population, i.e. the target group under study. For example, the answers of the participants in the Estonian Labour Force Survey form the basis for conclusions regarding the entire working-age population of Estonia (15–74-year-olds), and conclusions about all units which were economically active in the reference period are based on the responses of those having submitted the wage report of economic units. In order to generalise sample data to the total population, a weight needs to be calculated for each sample item. The weight of a sample item shows how many similar items in the total population the sample item represents. The weight is calculated based on the sampling method used and it may be adjusted according to the information received in data collection and processing. The weight of a sample item is always one or a number bigger than one. In surveys on economic units, the weight of large units is often one, because they stand out in the Estonian context and there are no similar units in the total population.

Calculation of statistics (phase 6 of Generic Statistical Business Process Model: analyse)

The information gained through data collection needs to be converted into data which allow answering the questions posed at the start of the survey. The collected data are quite often used for calculating average values (average gross monthly wages) or aggregates (the total number of unemployed persons) but also several indices. Price indices show the development of prices over time, e.g. the export price index reflects the change in the prices of exported goods. The most complex calculation, which combines the statistics of various domains, is the calculation of the gross domestic product (GDP).

Quality indicators

The quality of the published statistics is assessed based on five key principles: relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, availability and clarity.

There are several quality indicators for measuring accuracy and reliability, with the most important ones being standard error and the coefficient of variation. The aggregates and average values calculated based on sample data are estimates of the actual aggregates and average values, which are generally unknown, unless we interview all items of the total population. The similarity of the estimate to the actual value is indicated by the estimate's standard error and/or coefficient of variation. The smaller they are, the more exact the estimate. Other indicators reflecting the quality of statistics include the response rate, imputation rate, errors due to under- and over-coverage.

Presenting survey results. Tables

When the data have been gathered, we must return to the beginning and start looking for answers to the questions posed.

In electronic form, one table usually features a number of several characteristics, which is convenient in the sense that a suitable characteristic, pair or set of characteristics can be found as needed. If, for example, a table presents the data on a person's age, sex, ethnic nationality and county of birth, it is possible to separately view the distribution of each characteristic, each pair of characteristics (e.g. sex and ethnic nationality), each triad of characteristics (e.g. sex, age and the county of birth), and finally the joint distribution of all four characteristics, a total of 15 tables. One base table with multiple characteristics can be supplemented with even more characteristics; this has been done, for example, in the case of census data, but the number of respondents sets its limits. The more characteristics in one table, the more table cells and the fewer items in each cell, and it also happens that some cells are left completely empty. At the same time, a table with multiple characteristics allows making several fast inquiries, which are not enabled by smaller tables.

Upon the publication of statistics, Statistics Estonia must ensure that the information published in the tables would not allow identifying any economic units, persons or other items. There are several mathematical methods for preventing identification. The easiest way is to "hide" the content of the cell which allows identification and that of the related cells, but this involves a loss of information. Information loss can be avoided by using the rounding method, for example. This was done when the data of the 2011 Population and Housing Census was published. In the case of rounding, the content of cells allowing identification is not hidden, but the actual content is replaced with a rounded value.

Models

A large share of survey results is presented as distributions and tables, but sometimes the data are analysed further, with more complex statistical methods being used. One method increasingly employed in statistical actions is the use of **models**. Models help to find relations and reasons. For example, it can be asked which characteristics/circumstances influence the size of income. There are quite a few probable factors (explanatory variables): sex, age, place of residence, level of education, etc. Running a regression analysis helps to determine which variables actually help to explain the size of income. Logistic regression is a type of analysis used for predicting or estimating probabilities; for example, for estimating the probability that a person belongs to the population of permanent residents.

There are also rather complex models used for analysing **time series**. A look at the changes in the average wages over the course of a year reveals that, each year, average wages are lower in the 1st and 3rd quarter than in the 2nd and 4th quarter; this peculiarity (caused by summer holidays, for example) is called seasonality and it is something that needs to be taken into account when analysing time series.

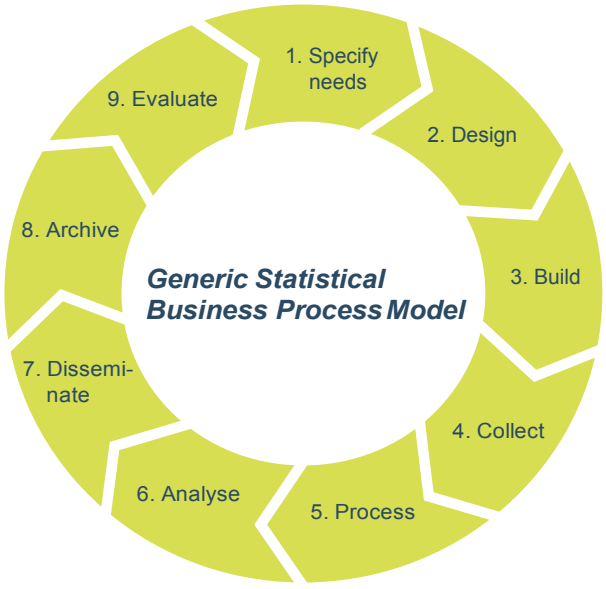
Publishing results (phase 7 of Generic Statistical Business Process Model: disseminate)

Era of paper publications

The publication of official statistics means making the statistics available to the public. Carrying out analyses is worthwhile only if the results reach the consumers. As recently as just a few decades ago, paper was the main carrier of statistical information – the tables and graphs as well as the analyses explaining and commenting on their content were all printed on paper. What was also definitely one of the means of dissemination was oral speech – scientific or popular lectures, which in the last half-century have been supported by electronic media.

Statistics on computers and in your pocket

The number of ways to disseminate analysis results has grown significantly these days. Electronic dissemination tops the list without precedent. Electronic tables and figures have the advantage of being interactive, which means that consumers can both select tables and design graphs according to their interests and needs. Texts which are presented electronically have greatly reduced the number of printed materials, but still have not completely taken the place of books and journals, which have instead become more content-rich and compact – paper is no longer used for printing only tables, as was done just 20 years ago. A very important supplement to spatial statistics has been the option of presenting multi-layer interactive maps. A contemporary person is followed by statistics everywhere: the data are available on smart phones via the application "Estonian Statistics" and help people to orientate themselves in the economy and society just like the GPS helps you to orientate yourself in physical space.



Fact sheet

Data collection

- Each year, Statistics Estonia organises approximately 200 statistical actions, two-thirds of which are completely or partially based on administrative data (registers).
- Survey samples comprise about 90,000 respondents per year, with approximately a half of the respondents being economic units and the other half – persons. Submitting survey data is compulsory for economic units and voluntary for persons.
- Statistics Estonia collects about 420,000 questionnaires per year, i.e. an average of 1,700 questionnaires per working day.
- There are approximately 150 different questionnaires, 140 of which can be submitted in the online environment eSTAT and the remaining 10 in hard copy or by means of a face-to-face interview.
- Nearly a half of the eSTAT questionnaires are pre-filled based on previously collected questionnaires, annual reports, declarations filed with the Estonian Tax and Customs board, and other such data.
- In order to notify respondents of their reporting obligation and remind them of submitting their questionnaires, Statistics Estonia sends an average of 450,000 e-mails to economic units each year. As a result, a half of the questionnaires are sent in by the deadline and another 25% by the end of the data collection period. By the end of data collection, about 25% of the questionnaires remain unsubmitted.
- 75% of the questionnaires are submitted by economic units via the electronic data submission channel eSTAT. A half of the remaining 25% of the questionnaires is submitted by e-mail or other means and a half in hard copy.
- An average of 800 respondents log on to eSTAT per day, and up to 2,000 respondents near the submission deadline of questionnaires with a large sample.
- If they run into problems, the respondents can address Statistics Estonia's Contact Centre. Each year, the centre answers approximately 65,000 calls and letters, with the average being 250 each working day. The respondents contact the centre the most often in February when the daily number of customer contacts exceeds 500.
- Nearly 65% of the persons included in the sample choose to complete the questionnaires. Face-to-face interviews are the prevalent method of data provision while telephone interviews are held as well, some surveys can be completed online.
- Questionnaires provide data for approximately 132,000 variables, 80% with surveys on economic units and 20% with personal surveys.
- Respondents spend a total of 43,000 days per annum on the compilation and submission of reports on economic units – this requires 170 full-time employees considering the number of working days.
- The amount of resources used for data collection and processing have decreased year after year, with approximately 100 employees engaging in these activities. An average of 2 hours is spent on collecting and processing the data of each report and questionnaire in Statistics Estonia.

Dissemination of statistics

All of the information published by Statistics Estonia is available to those interested free of charge on the website www.stat.ee. The number of website visitors has been growing each year. The average number of visitors is 400,000 per year and 13,000 per week. 86% of the website visitors are from Estonia. 6% of the users visit the web-site on their mobile phone or tablet.

The Statistical Database features approximately 4,000 tables, which get more than 900,000 views per year, i.e. 2,500 views each day.

Statistics Estonia issues about ten statistical publications per year and they can be either bought in printed form or downloaded from the website free of charge. Each year, approximately 9,000 copies are disseminated and the publications downloaded from the website free of charge total more than 20,000.

40–60 posts are published on Statistics Estonia's blog per year and the blog is visited more than 106,000 times each year.

Since 2013, the consumers of statistics can also use the smart application "Estonian Statistics" and, since 2014, the statistics map application, which provides geo-referenced statistics on maps.

Each year, Statistics Estonia receives over 3,500 requests for statistical information (data inquiries, orders).

In order to present its products and services, Statistics Estonia participates in seminars and conferences and organises consumer trainings.