Data protection and the result checking process

These guidelines describe the data protection of the data files used for research and especially the outputs (tables, figures, statistical models, etc.) produced from them, as well as the output checking process.

The data protection guidelines apply to the use of data files both in research projects and in microsimulation. However, the output checking process in microsimulation differs from other research projects.

If data protection issues raise questions, the researcher can contact the Research services (tutkijapalvelut@stat.fi / mikrosimulointi@stat.fi).

Data protection of research data files and outputs

Both Statistics Finland and researchers are responsible for the data protection of research data files. Statistics Finland is responsible for the data protection of the data files prior to their release for research use and for the data security of the remote access environment. Researchers must for their part ensure data protection during the research use of their data files and when publishing research outputs.

The researchers are responsible for the implementation of data protection in the research results they publish. These guidelines were prepared to support responsible activity in data protection matters, and they aim to prevent both accidental and intentional data protection breaches. The Research services use a checking process for research results, which also supports caution in data protection matters. The checking process is described in more detail in the section Output checking process.

Why data protection?

Ensuring the data protection of research data files is a precondition for Statistics Finland to release data files collected for statistical purposes for research use and for microsimulation. Statistics Finland’s right to collect register and survey data files for statistical purposes is guaranteed by the Statistics Act, as well as the right to release these data for research use. On the other hand, the Act also obliges to ensure appropriate protection of the data, and some of the data files contain very sensitive data.

The processing of personal data is guided by the General Data Protection Regulation (EU) 2016/679 of the European Union and the supplementary national Data Protection Act (1050/2018). According to the obligation of secrecy related to the user licence granted to the research project, the researcher must ensure that the research results do not contain unit-level data, that is, data concerning an individual person or enterprise or the possibility of their disclosure. When processing data files released for research use or working in a remote access environment, the researcher must also ensure, in accordance with the obligation of secrecy, that the data are not revealed or released to a party that does not have a user licence for it.

Data protection guidelines for different output types

Below are instructions and rules for ensuring the data protection of various output types. Because there are many different research projects and different output types, the data protection of each data file and output cannot be assessed and instructed separately. It is therefore necessary to lay down general “rules of thumb”.

In respect of some of Statistics Finland’s data files and data files released for research use by other authorities, data protection guidelines may differ from the rules presented below. These deviating rules can be found in the data file descriptions in the Taika-catalogue and if necessary, they are also recorded in the user licence decision.

If researchers are unsure about the data protection of some output or wish to ask for clarification of the given instructions, they should contact the Research services (tutkijapalvelut@stat.fi).

Outputs produced from the sample data

It is common for all output types that the disclosure risk of the observations underlying the results is usually lower if the used data file is only a (random) sample of the total data. However, the lower disclosure risk possibly caused by sampling does not mean that the protection rules presented in these guidelines could be automatically relaxed for outputs produced from sample data.

If the risk of disclosure of an observation has to be assessed case-specifically in some output (e.g. if the distribution parameter is revealing or whether the observation behind an individual image point can be identified from the image), then the use of the sample has an effect on the assessment.

It should be noted that the total data do not always mean only all persons resident in Finland or all enterprises operating in Finland. From the perspective of data protection and when assessing the disclosure risk of observations, all enterprises in a certain industry, for example, form the total data rather than the sample data, because it can generally be concluded whether an individual enterprise belongs to a “sample” that is limited to cover all enterprises in a particular industry.

Frequency and magnitude tables

Aggregated data in table format may contain personal data, enterprise data or both. For example, both the person and enterprise levels must be protected in employee-employer data files. Data describing occupational groups may also indirectly contain business data if (nearly) all persons belonging to a specific occupational group are employed by only one (monopoly) enterprise, in which case the data protection of that enterprise must be ensured.

The main rule in the protection of tables containing both enterprise and personal data is threshold value 3. In other words, data from a table cell or observation group may only be published if they are based on at least three (unweighted) observations. Data on the number of observations may not be published for small cells or groups either.

The use of the threshold value rule as the main rule is the easiest way to calculate the disclosure risk of individual observations possibly directed to the tables to be published. Identifying an observation as the only or rare case in its group may increase the risk of identifying the observation and/or reveal additional information about it (e.g. by means of other tables and outputs concerning the same data file or topic). For this reason, the threshold value should be applied even if only numerical data are published in the table.

In addition to the above-mentioned main rule, the following rules must be taken into account:

- In addition to the threshold value, the dominance rule[1] (1.75) must be used in recent enterprise data (under 15 months from the reference period). The dominance rule can also be required for use in older enterprise data or other data files (e.g. the Incomes Register). The dominance rule requires protection, for example, for data where the turnover or wages and salaries sum of one enterprise or some other variable in the data file form over 75 per cent of the turnover/wages and salaries sum of enterprises, etc. in a certain industry (sector, region, etc.) or if, for example, over 75 per cent of employees in a certain occupational group work in a specific enterprise.
- When protecting establishment-specific data, enterprise level protection must also be ensured whenever possible, so each cell must contain establishments of at least three different enterprises. The same must apply to group-enterprise relationships.
- In commodity data (products, raw materials and supplies from the statistics on industrial production) the number of enterprises is confidential for all production headings.
- With regard to personal data, data protection must be ensured especially carefully if the table contains sensitive personal data [2] (incl. specific personal data groups listed in the EU General Data Protection Regulation [3]). It may be necessary to use a higher threshold value to protect sensitive data.
- Own-account workers included in business data are subject to the same protection rules as other business data even though they are, as a rule, persons.
- The publication of coordinate or grid data is subject to special conditions: No complete grid data may be published. Data based on grids can be taken out of the remote access system and published when there are at least 10 observation units in the grid. However, data based on grids can only be published summed up to a larger regional level, as ratios, or otherwise processed so that persons and household dwelling units cannot be identified.

[1] According to the dominance rule (n,k) , the cell value of a table cannot be published if its n biggest observations form at least k per cent of the total value of the cell.
[2] Sensitive data here include data describing a person's race or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health, sexual behaviour or orientation, criminal convictions and offences and related security measures, cause of death, language, nationality, origin or country of birth, income, debts, property, rare occupation or other socio-economic group, social welfare needs or received social welfare services, social welfare support measures or other benefits, or data on treatment or other comparable measures directed at the person.
[3] Regulation (EC 2016/679 of the European Parliament and of the Council, Article 9.)

Different distribution parameters

The minimum and maximum are usually connected to one observation, so they cannot usually be published. For example, the largest enterprise in an industry can usually be identified, so the connected maximum turnover data cannot be published. If the minimum or maximum cannot be combined with an individual observation and the threshold value is realised, they can be published.

Distribution parameters (excl. minimum and maximum), such as deciles, form an exception of tables where the number of observations left in between the distribution parameters correspond with the cell frequencies. If these numbers exceed threshold value 3 applied in the tables, the distribution parameters can be published.

Modes can be published if (nearly) all observations do not get the same value.

The average, other ratios and the highest sub-items of distribution parameters (e.g. variation) can be published if at least three observations have been used in their calculation.

When publishing shares, threshold value 3 must be realised for all groups forming shares. In other words, if you want to publish, for example, that women's proportion of the total population is 58 per cent, that 58 per cent, as well as men's 42 per cent, must include at least three persons. Therefore, it is not enough that there are at least three women and men in total in the whole population.

Other numerical output types

Index point figures, correlation multipliers and test aggregates (t, F, X^2, etc.) can usually be published if at least 10 observations have been used in their calculation.

The regression model can be published as a whole if the model is based on enough observations (at least 10) and the model does not depict a time series based on observations of one enterprise or person. Individual multipliers of the model can usually always be published.

The numerical results of more complex statistical models can usually be always published if the model is based on enough observations and the model does not describe a time series based on observations of individual enterprises or persons.

Images

Like numerical outputs, images shall not reveal data of a single observation. Images drawn based on the data file are permitted if a single image point or a part of the image does not reveal the underlying individual observation.

Bar charts and other images used to present a classified data file are typically accepted for publication as long as each category has enough observations. The information of such an image can usually also be presented in table format and the same data protection rules can be applied to it as to other tabulated data files (see above Frequency and magnitude tables).

Distributions, histograms or cumulative distribution functions that have been adjusted from distribution charts or are presented at a sufficiently crude scale are allowed. Some distribution charts may contain data on deviating observations or outliers, which may, case by case, reveal data on an individual observation. Software drawing functions often automatically mark deviating observations in box plots, for example. Images identifying deviating observations are not suitable for publication unless the researcher can well justify that the deviating observations recorded in the image cannot be identified.

Scatter plots are typically used to show the values of two continuous variables. As a rule, the dots of the scatter plots describe the data of an individual observation, which means that they are trickier in terms of data protection than the previous image types. When assessing the data protection of scatter plots, special attention should be paid to the nature of the data file, for example, in relation to the size of the sample, the sensitivity of the data and occurrence of deviating observations. The scatter plot does not meet data protection requirements if data on the largest enterprise in an individual industry can be directly seen or easily inferred from it.

Protection methods

Data concerning an individual person or enterprise must not be identified from files or tables exported from the remote access system. Data with a disclosure risk must be protected by planning the contents of outputs to be acceptable for data protection, for example by using sufficiently rough classifications.

Cells with a disclosure risk in the table can be protected by changing the structure of the table, by suppressing individual cell values or whole rows or by changing cell values, for example, by rounding or replacing the original cell value by an approximate random number. When selecting the protection method of a table, efforts should be made to find a method that sufficiently protects the table but retains the features important for its intended use as well as possible.

Changing the structure of the table means controlling the number of variables or changing the classification. By changing the classification, the cells with a risk of disclosure are removed from the table by combining the categories contained in them to other categories in the table. In practice, changing the classification usually means that the whole classification becomes less detailed.

Suppression includes primary suppression of cells with a risk of disclosure and secondary suppression. Secondary suppression ensures that the values of primarily suppressed cells cannot be disclosed by means of table row or column totals. Suppression can also be made specifically for each row. If only a small number of statistical units belong to a particular row total of the table (fewer than the used threshold value), the row is suppressed in total without regard to the number of statistical units in its different cells.

Further information about the data protection methods can be found in the materials of the online course on research data in remote access use (in Finnish only) mentioned at the end of this page.

Output checking process

Research services use a checking process for research results concerning data files used remotely. If the researchers are unsure about the data protection of the

output, they should contact the Research services already before the output is checked or exported from the system.

The role of the checking procedure is to support the researcher's responsible activity in data protection matters concerning research results. Regardless of the checking process, the researchers are responsible for the data protection of their research results. The checking process enables the Research services to monitor the realisation of data protection in the results of research data files and to notice the need to offer additional guidance on data protection matters.

In practice, the checking process functions differently in remote use of research projects and remote use of the microsimulation model. However, in both cases the researcher must ensure that the outputs that they ask to be sent out from the remote access system do not contain unit-level data files or the possibility to reveal data concerning an individual observation.

Consideration should be applied when the output is sent for checking or out from the remote access system. Only outputs intended to be published should be sent for checking or out from the system. In other words, the transfer of so-called intermediate results, and in particular of (large) log-type files, must be avoided. The contents of the tables and graphs to be checked should be in the same format as they will be published. Outputs that could not be published due to data protection cannot be sent out from the system.

Outputs must be documented carefully so that the data content of the outputs is clear to the person checking them. The outputs must show the numbers of observations used in calculating tables, images, key figures, etc. If data deviating from the data protection guidelines are to be sent out from the output checking, the realisation of data protection in outputs should be well justified.

The file format of outputs, especially images, must also be such that there is no risk of disclosure of unit-level data. Image formats suitable for checking include:

- Bitmap formats
- PNG (Portable Networks Graphics)
- BMP (Bitmap)
- JPEG (Joint Photographic Experts Group)
- TIFF (Tagged Image File Format)
- Vector formats
- EPS (Encapsulated PostScript)
- PS (PostScript)
- PDF (Portable Document Format)
- SVG (Scalable Vector Graphics)
- WMF/EMF (Windows Metafile)

In Stata software, the above-described image formats can be created with the graph export command. In SPSS software, the image format can be selected in the Export output function. In R software, information about the drawing function is available with the command help(grDevices). Certain image types, like Stata's

gph files, as a rule save the material used for drawing the image, which means that they are not necessarily suitable for checking and sending out.

Online course on research data in remote access use (in Finnish only)

A online course on research data in remote access use is available Statistics Finland's website as part of the eCourse in statistics (https://tilastokoulu.stat.fi/verkkokoulu_v2.xql?page_type=ketusivu&course_id=tkoulu_tutki).

The online course provides further information about remote access use of research data files and the SISU microsimulation model and protection of data files. The course also contains examples of the protection of tabular data files.

It is recommended that especially researchers using the remote access system for the first time familiarise themselves not only with the present instructions but also with the materials of the online course on research data in remote access use.