Data protection of delivered data

These guidelines describe the data protection of the data files that have been physically released to be used for research. They also describe the data protection for the outputs (tables, figures, statistical models, etc.) produced from them. Physically released data files are used in researchers research environment and not over the remote access system FIONA.

These guidelines have been drawn to support responsible action in data protection matters and aim to prevent both unintentional and intentional data breaches.

If data protection issues raise questions, contact Research services (tutkijapalvelut@stat.fi).

The responsibility of Statistics Finland and the researcher when it comes to data protection of the data files and the outputs

Both Statistics Finland and researchers are responsible for the data protection of research datafiles. Statistics Finland is responsible for the data protection of the data files prior to their release for research us.

Researchers must for their part ensure data protection during the research use of their data files, during storage in accordance with the obtained license, as well as when publishing research outputs. When the license expires, the researcher must take care of the destroying the delivered data files and the copies and intermediate files formed from it. The destruction must be completed by the end of the license period.

The researchers are responsible for the implementation of data protection in the research results they publish.

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. 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 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”.

Outputs produced from 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.

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 belonging to a certain age-group, for example, form a total data rather than a sample data, because it can generally be concluded whether an individual belongs to a “sample” that is limited to cover all individuals of a certain age.

Frequency and magnitude tables

Personal data presented in table format should be aggregated and the risk for indirect disclosure should be prevented. The main rule in the protection of tables containing 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.

With regard to personal data, data protection must be ensured especially carefully if the table contains sensitive personal data[1] (incl. specific personal data groups listed in the EU General Data Protection Regulation[2]).It may be necessary to use a higher threshold value to protect sensitive data.

Different distribution parameters

The minimum and maximum are usually connected to one observation, so they cannot usually be published. If the minimum or maximum cannot be combined

[1] 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..

[2] Regulation (EC 2016/679 of the European Parliament and of the Council, Article 9.)

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 66 per cent, that 66 per cent, as well as men's 33 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 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 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 scatterplots 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 e.g. patients with a rare disease in a location can be directly recognized or easily inferred from it.

Protection methods

Data concerning an individual person or enterprise must not be identified from files or tables that are to be published. 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 foreach 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).