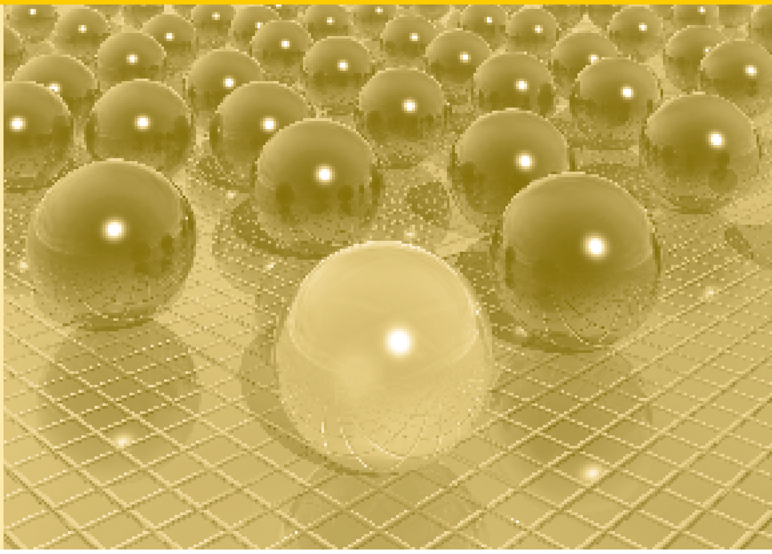
Regulations on the analysis of microdata

in the Research Data Centres of the Federal Statistical Office and the Statistical Offices of the Federal States (RDC)



Effective January 14th 2022

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Imprint

Publisher:
Research Data Centres of the Federal Statistical Office and the Statistical Offices of the Federal States

Production and editing:
Information und Technik Nordrhein-Westfalen
Mauerstraße 51
40476 Düsseldorf
Germany
Phone: 0211 9449-01
Fax: 0211 9449-8000
E-Mail: poststelle@it.nrw.de
Internet: www.it.nrw

Photographic rights
Cover photo: © Fotolia

Content

Please contact the RDC staff if the insert page is missing.

Dear data user,

With the **"Regulations on the analysis of microdata in the Research Data Centres of the Federal Statistical Office and the Statistical Offices of the Federal States (RDC)"** presented in this brochure, we provide you with important information that will help you with the analysis of RDC data. By following these regulations, you support a timely provision of your results and you avoid limitations through results that have to be blocked as part of the confidentiality check.

Background:

Federal Statistics Act requires confidentiality

According to the **requirements of the Federal Statistics Act** (Bundesstatistikgesetz – BstatG) on statistical confidentiality, the RDC are obliged to not release results that allow conclusions on individuals when making microdata available for scientific use. In consequence, all released results have to be absolutely anonymous. You find further background information on this in chapter 1 **"Basics of statistical confidentiality".**

RDC ensure confidentiality

Ensuring absolute anonymity is associated with substantial effort for the RDC. Every result that allows a conclusion on an individual has to be blocked. The main challenge in this is to ensure the **overall consistency** of all blocks in all results of your project. This prevents blocks from being uncovered by retroactive accounting.

> This brochure describes regulations that enable you and the RDC staff to handle the process of ensuring confidentiality easier and faster. **Thus, we ask you to follow the rules described below.**

Result economy

Furthermore, we ask you to consider the criterion of **result economy**. Have only results released that you actually need for your publication or thesis. Results of exploratory or preliminary analyses should not be submitted for release. Doing that, you avoid the consequence that your final results cannot be released because the subtraction between final and preliminary results generates confidentiality cases. If possible, we recommend the use of safe centres where you can see preliminary results. The advantage of a safe centre over remote execution lies in the fact that all results generated via remote execution have to be checked for statistical confidentiality before you can view them.

Which rules are there?

Requirements on the program code …

Chapter 2 **"Basic requirements concerning the program code"** deals, among others, with the documentation and structure of the program code and with the reproducibility of results. A comprehensible program code is important for the RDC staff because it is the only chance to correctly understand the creation of results and to correctly perform the confidentiality check.

… and on the output

Chapter 3 **"List of criteria on the admission of output"** contains a documentation of basic confidentiality rules of German official statistics. If these criteria are fulfilled, primary confidentiality cases can be avoided already during the analysis. This guarantees the absolute anonymity of the results. The **insert page**[1] gives information on the confidentiality rules that are applicable for a given statistic.

[1] BStatG of 22 January 1987, last changed by article 1 of the law of 21 July 2016

What are the advantages of following the rules?

Contractual commitment

The rules are committing as they are fixed in all contracts.

Defective output as a reason for rejection

If you do not fulfil the **basic requirements concerning the program code** (chapter 2), the RDC staff will not check the output you created at the safe centre resp. reject the program code you sent for remote execution. We then ask you to revise your program code in accordance to the rules.

Confidentiality issues as a reason for rejection

If you do not fulfil the **list of criteria on the admission of output** (chapter 3) there will probably be cases that have to be kept confidential. In that case, your supporting RDC can decide if your results will be checked regardless or if the output check is rejected (so called reservation rule).

If an output is rejected you have the chance to modify it to eliminate all confidentiality issues. If that is not possible for content-related reasons you can (in consultation with your supporting RDC) prioritise your results. If that does not lead to an agreement on the further strategy your output cannot be released.

Viewing results from remote execution at a safe centre

By the way: When using remote execution you can view your results at a safe centre without extra charge. That way you can find cases with confidentiality issues in advance and modify your program code accordingly or prioritise your output. Your program code is then made available at the safe centre.

What are the advantages of these rules?

- You get your results faster.
- Your results are less affected by blocked cases.
- You avoid the risk of not having your final results released. Final results may have to be blocked if the comparison with likewise preliminary results leads to confidentiality issues.
- You fulfil the criteria of scientific research when your results are transparent and reproducible because of a faultless and well documented program code.
- You consolidate the validity of your results when they are based on a sufficient number of cases.
- You make it possible for the RDC to use its resources available for confidentiality checks in an efficient way that is fair to all data users.

Please contact the RDC staff if you are insecure about the generation of your results or have questions regarding the regulations.

1. Basics of statistical confidentiality

1.1 Legal foundations

Statistical confidentiality is statutory

Statistical confidentiality is a central pillar of official statistics work. The following is stated in section 16 (1) of the Federal Statistics Law (BStatG)[2]:

Section 16 Confidentiality

(1) Individual data on personal and material circumstances provided for federal statistics shall not be disclosed by officeholders and persons specially sworn in for public service who are entrusted with the production of federal statistics, unless otherwise stipulated by a special legal provision. The obligation of confidentiality shall continue to apply after their activity has ended. […][3]

Statistical confidentiality protects respondents

Statistical confidentiality is closely associated with the obligation to provide information according to section 15 BStatG. It allows official statistics by law to claim information from the respondents. However, this obligation to provide information is an interference with the fundamental right of informational self-determination which declares that every individual can decide for themselves what personal information they pass on. To dissolve this contradiction the legislator has obliged official statistics by law to keep the respondents information confidential. That is why statistical confidentiality secures a fundamental right and is one of the most important tasks of official statistics

Statistical confidentiality safeguards quality of data collection

Furthermore, statistical confidentiality is an important factor for a trusting relationship between official statistics data collectors and respondents. The respondents have to be able to rely on the fact that no information on them or their conditions is made public and that their responses do not have negative consequences for them. It can only then be expected that the respondents are willing to give true and complete answers. That willingness to give correct and comprehensive answers is, in return, essential for the significance and quality of the collected data. Thus, statistical confidentiality also safeguards the functionality of the system of official statistics.

To enable the statistical offices to provide or release statistical results and data, the legislator has included exception rules that show solutions to do so without risking data security. Section 16 (1) item 3 BStatG states the following:

> *The obligation of confidentiality shall not apply to …*
> *3. individual data which have been merged with the individual data of other respondents by the Federal Statistical Office or the statistical offices of the Federal States and are presented as statistical results,*

Section 16 (1) item 3 BStatG states the exception rule that enables the statistical offices to release publications by combing results so that individual cases cannot be identified any longer. This exception rule is also the foundation for the release of results that are produced by researchers via the RDC.

[2] BStatG of 22 January 1987, last changed by article 2 of the law of 14 June 2021

[3] English working translation; only the German version is authentic.

1.2 Methodological foundations

No conclusions on individual cases allowed

According to the legal requirements, statistical disclosure control has to ensure that the published results do not allow conclusions on individual cases (e.g. persons, enterprises, companies, institutions etc.). This applies for results generated by the statistical offices as well as by RDC users.

Protection of individual cases through blocked cells

The RDC regulations are stated in chapter 3. In most cases, the RDC ensure statistical confidentiality by blocking cells. The method of blocking cells includes that not only primarily confidential values are to be blocked (those are cells that directly allow the re-identification of a respondents data) but also all values that allow the recalculation of blocked cells. This secondary and, possibly, cross-tabular confidentiality implies that cells that contain originally uncritical values have to be blocked as well. To ensure the secondary confidentiality, the RDC even have to consider results released in publications and released results you generated at an earlier stage of your project. To ensure the cross-tabular confidentiality the RDC of the Federation and the Federal States are obliged to match your results against standard publications and central analyses.

Blocked cells constrain (future) analyses

In consequence, every blocked cell in released results may lead to restrictions for the release of future results. We thus recommend you to decide on a strategy for your analyses (concerning the definition of variables and the applicable filtering) at an early stage of your project to avoid the appearance of confidentially critical difference groups (see criterion 2.9 and criterion 3.2.4). Additionally, you should take care to base all your results on a sufficient number of cases. The easiest way to accomplish this is to summarise characteristics when facing small numbers.

2. Basic requirements concerning the program code

The basic requirements concerning the program code are binding criteria for writing program code for the analysis of official statistics using remote execution or safe centres. The basic requirements were implemented to make your program code comprehensible for yourself and third persons and to produce a clear result file. The basic requirements are specified in the following chapters. A sample program code can be found on www.forschungsdatenzentrum.de. It is an obligatory template and provides the basis of your program code.

Violation of a criterion listed below may lead to the rejection of the execution of your program code (remote execution) resp. of the check for confidentiality for the output created with the program code (safe centre).

2.1 Clarity of the program code

Comprehensible structure

The program code has to be written with a clear structure. All program steps have to be comprehensible. Single program sections (header, data preparation, analyses, etc.) are to be marked clearly and separated visually.

All specifications that have to be adjusted by the RDC to execute the program code in the RDC environment (path specification, name of the dataset, output management) are to be stated only once and in the header of the program code (see sample program code).

A master program code is to be used if the preparation and the analysis of the data take place in different program codes. The path specifications are then to be stated only in the master program code. All program codes are to be described in a short and accurate way and to be started in an automated process. When using the program R, the use of a master code is always mandatory (see sample program code).

Further requirements for the visual and contentual design of the program code:

- Consistent notation of commands and terms (among others case sensitivity) and of other objects used in the code (e. g. labelling of relations or missing values)
- Consistent and self-explaining variable names (including the statement of the original variable)
- Readable (and consistent) abbreviations for commands
- Indention of loops
- Visual design of topically connected program sections (e. g. by using equal spaces)
- Comprehensible numbering of single results

The criterion is met if …

… the specifications that have to be adjusted by the RDC are written in the header of the program code

… AND the program code has a clear structure and consistent and unique names.

2.2 Detailed commentary

All steps for preparation and analysis of the data have to be commented reasonably and in detail. Their content has to be described.

Complete header

For that, a header has to be set at the beginning of every program code. This header includes the project (project title and number), contact data, layout of the program code (including a placement in the program context) and the RDC products (statistic and year), variables and macros addressed in the program code (see sample program code).

Expressive commentation

Additionally, all sections of the program code, commands resp. analyses and used macros have to be marked with comprehensible and unique comments. Relations to previous (and possibly future) analyses have to be stated. Especially alterations to earlier program codes concerning filtering, definition of subgroups, etc. have to be marked. Alterations that might have an effect on the statistical confidentiality have to be traceable.

Steps for preparation and analyses of the data include, among others:

- Generation of new variables based on the originally provided variables
- Changes to variables
- Pooling and merging of datasets
- Filtering
- Generation of statistical results
- Generation of graphical results/diagrams and the related check tables (see criterion 2.8)
- **If permitted and after consultation with the RDC:** Merging of external variables (see criterion 2.12)

The criterion is met if …

… there is a complete and up-to-date program code header,

… every step of the analysis is commented comprehensively and in detail

… AND relations to previous (comparable) preparations and analyses are marked.

2.3 Uniqueness of variable and value labels

Variable and value labels have to be assigned uniquely. If a new variable is created or if an existing variable is adjusted then all related labels (especially the variable label) have to be assigned newly and to be stated in the header of the program code. It is then to be paid attention to using descriptive names. Variable names that are identical to (possibly abbreviated) program orders are to be avoided. The values of newly generated or adjusted (categorical) variables have to be labelled.

The criterion is met if …

… variables contain the same information in all steps of the generation of results

… AND all values of categorical variables are labelled.

2.4 Reproducibility of the output

Program code reproduces output correctly

An output that was generated at a safe centre and that is to be checked for release has to be identically reproducible by the associated program code. An output will be rejected if manual adjustments were made in the generation of said output. The logging may not be switched off at any time, neither at the safe centre nor when using remote execution.

Program code seamlessly documents all work steps

A program code and the output produced by it have to be compatible. For every output that is to be released, the associated program code has to be available for the confidentiality check. The program codes have to seamlessly document the way from the original data to the output that is to be released. If more than one program code is involved in generating the output then a master program code has to be employed (see criterion 2.1). The names of the output datasets have to have a unique relation to the program code they're based on.

The criterion is met if …

… the program code seamlessly and completely reproduces the output that is to be released

… AND there is a complete log.

2.5 Specification of the output formats

Formats for tables

All tabular and analytical results are to be saved in a processable format so the RDC is able to conduct to the confidentiality check. The formats of the statistical analyses software packages and the excel format are suitable for this.

Formats for graphical output

In contrast, all graphical results have to be saved in a non-processable format to prevent underlying values or numbers of cases to be released. PDF, JPG, PNG or TIFF files are examples for suitable formats.

The criterion is met if …

… the correct output formats were used for all generated results.

2.6 Marking of output to be released and output for the confidentiality check

Marking of the output category and the relations

The RDC distinguish two categories of output: First, there is output that is to be checked by the RDC staff and released for publication. Second, there is output that is generated only for the conduction of the check for confidentiality (see the following criteria 2.7 to 2.10). Both output categories and their relations have to be unambiguously marked. The marking can be done in different ways and has to be agreed upon with the supporting RDC. Possibilities are, for example:

- Consistent, consecutive and identical numbering or name affixes
- Output below each other (output to be released and output for the confidentiality check are written below each other in one output file; output for the confidentiality check is then written directly below the associated output that is to be released)

  OR

  Output is saved in two different data files with identical layout (one file which only contains output that is to be released and one file that only contains output for the confidentiality check; the numeration shows which results belong to each other).

The criterion is met if …

… a way of unambiguously marking the output categories was agreed upon in advance and is implemented.

2.7 Output of the underlying numbers of cases and marking of the relations

Output of unweighted numbers of cases

For all output that is to be released, the underlying unweighted number of cases is to be stated. The output of the number of cases serves the check for potential risks for confidentiality. It is to be ensured that the results that are to be released and the results for the confidentiality check as well as the according relations are marked (see criterion 6).

The underlying unweighted numbers of cases are to be stated for:

- Statistical key figures (e. g. measures of central tendency, measures of variation, quantiles, ratios)
- Multivariate analysis
- Weighted results (e. g. weighted number of cases tables)
- Graphical results (see criterion 2.8)
- Value tables (see criterion 2.10)

The criterion is met if …

… the underlying unweighted number of cases is stated for every output that is to be released

… AND the results resp. relations are marked unambiguously.

2.8 Output of underlying numbers of cases and values for diagrams and graphical results

If diagrams and graphics are to be released then additional tables with the depicted values and the underlying unweighted numbers of cases have to be stated and to be marked unambiguously for the confidentiality check (see criterion 2.6). Consultation with the supporting RDC may be appropriate.

The criterion is met if …

… tables with the depicted values and the underlying unweighted numbers of cases are stated for all graphical results and diagrams that are to be released

… AND the results resp. relations are marked unambiguously.

2.9 Output of difference groups and marking of relations

Output of remaining subgroups

If results for one or more associated and not overlapping subgroup(s) are created in addition to results for the whole population then the results for the remaining (possibly summarised) subgroup(s) always have to be stated as well. Missing values should preferably be stated separately to avoid difference problems in the following analyses (see criterion 2.6).

In case of overlapping subgroups the number of cases has to be stated for every intersection. If, for example, there is an analysis for the age groups ‘below 18’ and ‘18 to 24’ and later on the same analysis is conducted for the age groups ‘below 16’ and ’16 to 24’ then the number of cases for the age group ’16 to 17’ has to be stated as well.

Output of difference groups

The output of these difference groups serves the check of results for confidentiality risks. It has to be ensured that results that are to be released, results for the check for confidentiality and the according relations are marked unambiguously (see criterion 2.6).

The criterion is met if …

… the according differences are generated for all results that are to be released

… AND the tables resp. relations are marked unambiguously.

2.10 Output of certain values for the check for dominance and marking of relations

Output of the highest two individual values

If value tables (sums) are created using economic or tax statistics then the number of cases and the highest two individual values have to be stated for the check for confidentiality. These values have to be stated in a table that is generated only for the check for confidentiality and the according relations have to be marked unambiguously (see criterion 2.6). The output of these values serves the necessary check for dominance (see criterion 3.1.3).

The criterion is met if …

… the two highest values that are included in the calculation of a statistical key figure, the according number of cases and the sum are stated

… AND the tables resp. relations are marked unambiguously.

2.11 Non-redundancy of statistical results

Avoidance of identical analyses

In the course of a project, identical statistical results may only be marked for release once. This serves the reduction of the amount of effort for output checking in the RDC. If results have to be released again in duly substantiated exceptional cases, an exact reference to the according earlier analysis has to be made (see criterion 2.2).

The criterion is fulfilled if …

… a statistical result is only marked for release once in the course of a project

… OR substantiation is given for the repeated release and a reference to the earlier analysis is made.

2.12 Merging of external variables

External variables as a separate dataset

The merging of external variables has to be reconciled with the RDC in advance – if possible already when handing in the request for the use of data. When requesting the merge of data, a description of the process is to be given. The RDC provide a form for this that has to be used. It can be found on **https://www.forschungsdatenzentrum.de/en/request**. The merging can be done by either the RDC staff or the data user and is liable to cost in both cases. After the RDC has agreed to the merging, the merge procedure is clarified the process is laid down in a contract, the external data is provided to the RDC in a separate and suitable data file. If necessary, a separate and well commented program code is provided by the data user. All external variables that are documented in this program code have to be associated unambiguously to the agreed-upon project description. The provided program code has to meet the basic requirements concerning the program code.

The criterion is met if …

… only variables that were agreed upon with the RDC in advance are merged

… AND (if the data user is merging the data) all required program steps and used variables are written in a separate and well documented program code.

2.13 Use of ado files (Stata) or R packages

Ado files and R packages can be used at the safe centres as well as via remote execution. For both ways of access, access to the internet is neither allowed nor possible. Thus, all necessary files must first be sent to the supervising RDC location as a compressed file (.zip-, 7z.-archive) or be made available via download-link or be downloadable with a separate program code. The supervising RDC location has to agree to the way of transmission. The supervising RDC location will then provide the agreed files in your project folder. Please note that all files are checked by the RDC. If necessary, explanatory details will be requested from the data user. Please schedule at least one working day before a visit at the safe centre or the provision via remote execution. This also necessitates an agreement with the supervising RDC location.

The criterion is met if…

… all necessary ado files or R packages can be downloaded by the RDC through a separate program code

… OR all necessary ado files or R packages including all dependent files are provided in time as zip folder

… AND the actual program codes for data preparation and analysis do not need access to the internet.

3. List of criteria on the admission of output

The legal obligation to keep individual cases statistically confidential is stated in chapter 1.1. It is also valid for results that were generated within the context of scientific projects in the RDC using remote execution or a safe centre. Before these results can be released, the RDC staff has to check them for absolute anonymity. This check follows a set of fixed rules.

In this chapter, the applicable confidentiality rules of German official statistics are documented in detail. They are valid for results based on original variables as well as self-generated variables. However, not every rule is applicable for every statistic. **The insert page**[4] shows the rules that are to be used for a given statistic

We ask you to apply the confidentiality rules applicable for your analyses in the best possible way before handing your results to the RDC staff for the confidentiality check and release.

The RDC can reject the confidentiality check and release of your results if their absolute anonymity can’t be guaranteed because of included primarily confidential cases. The RDC will then inform you about available possibilities.

3.1 Rules for numbers of cases and value tables

3.1.1 Rule of minimum number of cases

Protection of unique combinations of characteristics

A value is to be kept confidential if only one or two cases contribute to it. This is also applicable for more advanced analyses (regressions, test procedures, etc.).

The rule of minimum number of cases protects rare and unique combinations of characteristics that might otherwise lead to a re-identification

.

The criterion is met if …

… in tables at least three cases contribute to a table cell

… RESP. each value that is visible in a graphical result is based on at least three cases

… RESP. every value output within a more advances analysis is based on at least three values.

[4] https://www.forschungsdatenzentrum.de/sites/default/files/RDC_regulations-microdata_overview.pdf

3.1.2 Rule of marginal value

A table cell is to be kept confidential if the frequency of an inner table field is distinguished by only 1 from the frequency of the corresponding marginal field.[5]

Protection if the group membership is obvious.

The rule of marginal value prevents the assignment of characteristics to individual survey units or groups.

Example: *The income groups (categories: 'below € 2,000' and '€ 2,000 and more') and the region are under consideration at the personal level. For region X with 25 persons all 25 persons belong to the same income group:*

Region	Income < 2.000 €	Income ≥ 2.000 €	Sum
Region X	25	0	25

Third persons with knowledge of a person's regional belonging can unambiguously identify the income group.

The confidentiality risk also exists if only one person belongs to another income group:

Region	Income < 2.000 €	Income ≥ 2.000 €	Sum
Region X	24	1	25

The person with the higher income group knows that all other persons in this region have an income of less than € 2,000.

The rule of marginal value does not apply if there is only one characteristic that is logically possible for a group. For example, the result that nobody in the age group 0-10 years is employed is not a confidentiality case.

The criterion is met if …

… in tables the frequency of an inner table field is distinguished by at least 2 from the frequency of the corresponding marginal field.

[5] If the rule of minimum number of cases has to be met at the same time, the necessary difference increases accordingly.

3.1.3 Rules of dominance

Protection from approximate disclosure

In German official statistics, the rules of dominance are applied to protect the declared values of the one or two survey units that contribute most to the total value. They ensure that a person with prior knowledge of the value of one of the two largest individual values cannot estimate the other individual value. Even approximate disclosure is to be prevented. According to the rules of dominance, a value is to be kept confidential if the contribution of the largest individual value or the two largest individual values is larger than a set proportion of the overall value.

> ***Example:*** *The income of companies in a given economic sector and a certain region are under consideration. It is known that there are only two large and eight very small companies. The two large companies have incomes of € 74 m. and € 65 m. The eight small companies have a total income of € 1 m. The second largest company knows its own income and is aware that all small companies together take only an insignificant part in the total income. If it subtracts its own income (€ 65 m.) from the total income (€ 140 m.) it can estimate the income of the largest company with € 75 m. and thus approximately disclose this value.*

The thresholds that are used for the rules of dominance may only be used internally and may not be published. The publication of the parameter value might otherwise lower the anonymization's security level in released table cells which could cause the risk of the approximate disclosure of individual values.

The criterion is met if …

… the output of sums includes the statement of the two largest individual values and there are no dominance cases in the results.

3.2 Rules for more advanced analyses

3.2.1 Output of individual values

No output of individual values

Individual values are (with very little exceptions in certain statistics) to be kept confidential. The prohibition of the output of individual values includes, among others, results from the following analyses:

- Listings of individual values (e. g. ‘list’ command)
- Minima and maxima

  Exceptions to this rule may be, for example, minima and maxima of specifically generated auxiliary variables (e. g. concentration measures) and of dummy variables. In these cases, it has to be checked if the release might cause a disclosure risk. It has to be marked if minima and maxima are stated only for the confidentiality check.
- Residuals

The criterion is met if …

… the output contains no individual values.

3.2.2 Quantiles

Minimum number of cases for every quantile segment

A result is to be kept confidential if less than 3 cases contribute to a quantile segment.

Example: *The 50% quantile separates the statistical units in two segments. Every segment has to be based on at least three units, so the minimum number of cases for this analysis is six.*

In particular, the following minimum numbers of cases are valid for the output of quantiles:

50%-quantile	→ N ≥ 6
25%- resp. 75%-quantile	→ N ≥ 12
10%- resp. 90%-quantile	→ N ≥ 30
5%- resp. 95%-quantile	→ N ≥ 60
1%- resp. 99%-quantile	→ N ≥ 300

The criterion is met if …

… the minimum number of cases is met for each quantile analysis.

3.2.3 Output of Graphical results

Minimum number of cases for graphical results

Graphical results are to be kept confidential if there is at least one confidential value among the underlying numbers of cases or values. The rules that define if a case has to be kept confidential can be found in chapter 3.1 of this brochure.

For the check for confidentiality, the underlying numbers of cases or values are to be stated for every graphical result

The criterion is met if …

… the numbers of cases or values that underlie a graphical result do not contain any confidential cases.

3.2.4 Analysis of subgroups

Minimum number of cases for subgroups and remaining populations

Results for subgroups are to be kept confidential if knowledge of the results of these subgroups and prior knowledge of the total population allow conclusions on the remaining population that are relevant for confidentiality. That's why every examined subgroup and (if existent) the not examined remaining population has to fulfil the confidentiality rules (see chapter 3.1).

> ***Example:*** *An output contains the number of companies in a region (n=27) and the number of all companies in this region with a maximum income of € 1 m. (n=26). Because these numbers differ by only one, it can be concluded that there is exactly one company in this region with an income of more than € 1 m.*

If analyses for subgroups are conducted then the results for all subgroups in the population have to be stated. Results for subgroups that are of no interest for the analysis can be summarised. Missing cases should also be stated separately to avoid difference problems in future analyses.

The criterion is met if …

… results for all examined subgroups and the remaining population are stated

… AND they fulfil the confidentiality rules.

3.3 Additional special rules for certain analyses

Additional rules for specific analyses

Depending on the kind of analysis there may be additional rules that have to be fulfilled. There may, for example, be special rules for a given statistic concerning the analysis of geo-referenced data or for certain multivariate methods. As the RDC enable a very large array of analyses not every specific rule can be explained here. If you plan to conduct analyses that are not listed above, please contact your supporting RDC. It can inform you about additional rules for certain analyses.

3.4 Additional rules for certain statistics

Additional rules for specific statistics

Some statistics differ from the confidentiality rules described above and use other methods or parameters to ensure confidentiality. There are, for example, some statistics which do not use the methods described above but ensure confidentiality by implementing data transformations (e.g. rounding procedures). There are also statistics that use, for example, a higher minimum number of cases or do not need to be kept confidential at all. For these statistics, there might be other criteria than the ones described above.

The statistic-specific rules of confidentiality were set by the Statistical Offices of the Federation and the Federal States. They are binding. Please have a look at the “Overview of the applicable rules of statistical confidentiality per statistic” (see insert page) to find out which confidentiality rules are applicable for a given statistic. Your supporting RDC can give you detailed information if the statistic you are interested in is being kept confidential with statistic-specific special rules.

4. General terms of use

Below, we want to bring our general terms of use to your attention. These contain additional, partly contractually defined regulations that are important for the use of data in the RDC. Please read your user contract additionally or have a look at the English working translation that can be provided by your supporting RDC location.

The use of German official statistics is

1. protected by law
2. bound to a specific user or group of users
3. bound to a specific purpose
4. temporary
5. contractually agreed
6. liable for costs

4.1 Use is protected by law

Prohibition of re-identification

The RDC are legally bound to check all statistical results that were created within the context of scientific projects based on provided microdata for statistical confidentiality. This serves the protection of data according to section 16 (6) of the Federal Statistics Law (BStatG). Should individual cases be part of the output then they have to be blocked. These blocks have to be consistent for all analyses in a project. Data users who deliberately plan to re-identify individual cases are liable to prosecution and are expelled from further data uses. If a data user unintentionally re-identifies individual cases they have to immediately inform the RDC about it. The protection of data also includes the fact that external variables may only be matched to the requested data if this was brought into agreement with the RDC in advance – if possible in the context of the request for data access.

4.2 Use is bound to a specific user or group of users

Use by scientific institutions

Only scientific facilities assigned with independent scientific research are eligible for use. These are universities (also colleges and equivalent institutions) and scientific institutions. If a scientific facility submits an application for access to microdata for the first time, the eligibility for use will be legally assessed. The assessment may take several weeks. The data may only be used by persons who belong to the eligible facility, meaning they are enrolled in the institution, their thesis or dissertation is supervised by the institution, they are employees of the institution or have a guest researcher's status.

In addition, it is necessary to commit users to statistical confidentiality in accordance with section 16 (7) of the Federal Statistics Act (BStatG). This commitment can be carried out in any statistical office.

4.3 Use is bound to a specific purpose

Use for scientific projects

The use is restricted to scientific projects only. These may be, for example, qualification works like master or doctoral theses, but also projects funded with own resources or third party funded research projects (e.g. research projects on behalf of the German Research Foundation, of foundations, associations or ministries). A separate application form must be submitted for every research project. Several publications may originate from one project.

Correct citation of the data

In publications, the used official microdata is to be cited as follows:

> *Source: RDC of the Federal Statistical Office and Statistical Offices of the Länder, [name of statistic used], [survey year(s) yyyy-yyyy], own calculations.*
>
> *Resp.*
>
> *Quelle: FDZ der Statistischen Ämter des Bundes und der Länder, DOI: [DOI der verwendeten Statistik(en)], eigene Berechnungen*

Publication copy

In In addition, it is necessary to make at least one (printed or electronic) copy of the publication available to the Research Data Centres of the Federal Statistical Office and the Statistical Offices of the Federal States.

4.4 Use is temporary

No data for permanent tasks

The regular duration of a data use is three years. There is a possibility of lengthening this period for another three years (liable for costs). Projects that use the students discount may use the data for only one year; lengthening this period is not possible. The use of the data for research projects is temporary because it is bound to a specific purpose – in other words, a project of limited duration. The data may not be provided for permanent tasks. During the regular duration of the project further statistics, new survey years or external variables can be added to the project (liable for cost). If publications that are based on RDC microdata are engaged in a peer review process or if a review process is planned, then a ‘review stage’ of not more than three years can be requested (liable for cost).

4.5 Use is contractually agreed

Rights and obligations of the contractual partners

For the use of data, a user contract with a given duration is concluded between the scientific institute requesting the data and the Research Data Centres of the Federal Statistical Office and the Statistical Offices of the Federal States. This contract regulates the rights and obligations of both contractual partners. Those include, for example, the obligation of the RDC to supply the data and the obligation of the institution or data user to follow the rules of statistical confidentiality. It is also contractually agreed that when visiting a safe centre, no mobile devices (e. g. mobile phones, laptops) may be brought that would allow the user to acquire or record additional knowledge or that could take photographs.

4.6 Use is liable for costs

See homepage for costs

A fee is charged for the use of German official microdata. The amount of the fee depends on the number of statistics, survey years and ways of data access that are used. Furthermore, it is relevant whether the requested data are part of the standard supply of the RDC or are prepared specifically for a project. Adding further statistics, new survey years or external variables is also liable for cost. For more information on the charge please see **www.forschungsdatenzentrum.de.**