

Guidelines for the Design of Evaluation Programmes and Analysis Results

Table of contents

Introduction	3
1 Why users should read the guidelines?	4
2 Data use in the FDZ infrastructure	6
2.1 Workflow from the analysis to the release of the results	6
2.2 Contractual obligations and sanctions	7
3 Guidelines for the design of evaluation programmes	8
3.1 The master file	8
3.2 Programme Files	9
3.3 Efficient Data Handling	10
4 Guidelines for privacy-compliant results	11
4.1 Number of observations in general	11
4.2 Frequency counts	12
4.3 Descriptive statistics	14
4.4 Percentiles / quantiles	15
4.5 Weighting	16
4.6 Graphs	16
4.7 LaTeX output	16
4.8 Aggregated data	17
4.8.1 Importing aggregated data into the project	17
4.8.2 Documentation of aggregation steps within the project	17
4.8.3 Export/transfer of user-generated aggregate data sets	18
4.9 Regression output	19
4.10 Event data analysis	20
4.11 Sequence pattern analysis	20

Introduction

The Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB) provides external researchers with access to microdata for non-commercial research in the area of social insurance as well as labour market and employment research. The FDZ provides social data that are subject to the data protection regulations of the German Social Security Code X (SGB X ¹) or the requirements of statistical confidentiality.

Apart from the creation and documentation of data products for labour market and employment research, the main task of the FDZ is to support data users in ensuring the confidentiality of information on statistical units such as persons, establishments or enterprise.

¹Zehntes Buch Sozialgesetzbuch - Sozialverwaltungsverfahren und Sozialdatenschutz

1 Why users should read the guidelines?

In practice, to ensure data protection, the deanonymization potential of result files must be checked after evaluations using controlled remote execution before their release. Since this work ties up a lot of resources due to the large number of users and the volume of results generated, the processes must be standardized. This is the only way that the FDZ can process all requests quickly and still have sufficient free capacity for updating and further developing the data products. To do so, the FDZ relies on the support of the users.

What can you contribute to a smooth and fast process?

- Read the following documents **before** you start with the analyses in the FDZ infrastructure:
 - the ‘[Manuals for Remote Data Access and On-Site Use at the FDZ of the BA at the IAB](#)’ (from here on: ‘manuals’)
 - the ‘[Guidelines for the Design of Evaluation Programmes and Analysis Results](#)’ (from here on: ‘guidelines’)
 - the ‘[Stata Templates of the FDZ](#)’ (from here on: ‘FDZ templates’)
- Design your programme files according to these documents.
- Check the runnability of your programmes using test data or subsamples.
- Check the runnability of the programmes and the data protection compliance of your results before submitting the results to the FDZ for review and approval. This means that the analysis results should normally be designed in such a way that subsequent manual deletions by the FDZ team are no longer required.
- Reduce the results that need to be subjected to a data protection review to the absolute minimum. Only ask for results that are intended to be published or presented to be checked. Results of preliminary analyses, quality checks during data preparation, etc. should not be submitted to the FDZ for review and approval.
- If the programme stops running, first try to find and correct the cause of the error yourself. First, check that your syntax is correct and that the technical specifications for the FDZ infrastructure (e.g. the use of “globals” and path specifications) have been implemented correctly. Second, consult the [FAQ list for JoSuA](#). Only if all measures do not lead to success, contact the FDZ.

The FDZ supports you with [templates](#), which allow an easy start for the implementation of the syntax design as well as the output of data protection compliant results. A detailed description of the FDZ templates can be found in the [zip archive](#) of the templates. In addition, you

will find references to corresponding programming examples in the templates in the following sections.

This data protection manual serves to explain and deepen the requirements for syntax creation and the output of data protection-compliant analysis results, which are illustrated by examples in the FDZ templates. Here you will learn how the analysis process ideally runs in the IT infrastructure of the FDZ while complying with the requirements. In addition, results classified as critical from a data protection perspective are explained, and strategies for avoiding such results are outlined.

What are the advantages of complying with the guidelines?

- You receive your results faster.
- Your results contain few or no deletions and additional blocking to prevent re-calculation of deleted values.
- Your results are based on sufficiently high numbers of observations and therefore have a higher significance.
- By following the guidelines, your programmes meet the requirements of “good scientific practice” and are thus easily reproducible. Moreover, your results can be reproduced at any time.
- By efficiently reviewing the results, the FDZ can devote more resources to other tasks, such as updating existing and developing new data products or improving the quality of services.

! Important

If you have any questions, please contact the Research Data Centre (iab.fdz@iab.de). Please always include your FDZ project number (**fdzXXXX**) and, if applicable, the **job ID** from JoSuA.

2 Data use in the FDZ infrastructure

2.1 Workflow from the analysis to the release of the results

As a rule, the data should be used in Stata. If individual analyses are not feasible with Stata, it is possible to work with R or Octave. To do so, please contact the FDZ in advance.

The FDZ's data are analysed through controlled remote execution, during an on-site use or remote desktop access. For both access methods, the FDZ uses the browser-based software JoSuA (Job Submission Application). For more information on JoSuA and the different modes, please refer to the [manuals](#). Please use the Internal Use Mode in JoSuA for data preparation¹, preparation of on-site visits and for testing the runnability of the evaluation programmes. The results visible to you in the Internal Use Mode are censored through an automatic data protection check. You can look at the results on the JoSuA web interface, but you are prohibited from downloading them². The data sets stored in your project directory in the Internal Use Mode are also available in the Presentation/Publication Mode. Please note that the datasets saved in remote desktop access must be generated again via JoSuA.

Please upload the programmes of the selected results that you want to publish in a presentation or publication in the Presentation/Publication Mode. Before doing so, please make sure that your do-files run correctly in the Internal Use Mode and that the number of observations underlying the results is sufficient³ (further information on the required number of observations can be found later in this document). The programmes and result files uploaded in the Presentation/Publication Mode are subject to a data protection check by the FDZ team.

As a rule, **no** further manual deletions or censoring are performed during this process. In limited exceptional cases, the FDZ carries out manual deletions in the log files. For this purpose, sufficient reasons why the results cannot be displayed in a different form must be provided in the programme files. After successful review by the FDZ, the programmes and result files will be made available for download. If critical results are still contained, the affected log file or, if applicable, the entire affected job will not be released. You then have the

¹At the end of the project, you can upload the data preparation programmes via the Presentation/Publication Mode for verification requested by reviewers of journals. Please note this in the JoSuA comment field.

²The results generated in the Internal Use Mode may not be copied, photographed, printed or duplicated. It is also prohibited to allow third parties to see them. Any use of these results except for the further development of your programmes constitutes a breach of contract.

³We reserve the right not to release any result files within JoSuA jobs that have obviously not first been tested in the Internal Use Mode for runnability and compliance with data protection.

option of revising the corresponding programme file accordingly and uploading it to JoSuA again.

2.2 Contractual obligations and sanctions

Despite the checking routines in the FDZ, you are obliged to check the transmitted results again yourself. In case of doubts, the results must not be published, but reported to the FDZ. In addition, you agree to maintain confidentiality in the data use agreement, which includes, among other things, that no re-calculations will be made to determine deleted values. In case of violations, you or, in severe cases, your entire institute or university will be excluded from any further data use for a period of up to two years. Furthermore, the penalties specified in the use agreement may apply. Information about violations of the above obligations will also be passed on to other research data centres.

3 Guidelines for the design of evaluation programmes

The FDZ staff, who check the output, rotate daily. In order to be able to check the output, it must be possible for the FDZ employees to understand the programmes without having knowledge of the project content. Therefore, sufficient commentary on the do-files is necessary. Examples for the structure of the do-files can be found in the [FDZ templates](#) on our website at.

The following general principles must be considered when generating analysis programmes, so that an analysis programme complies with the guidelines of the FDZ¹.

3.1 The master file

The master file must be designed as follows:

- All programmes used in the course of the project must be run from a master file. This file must always be named `master.do`. The `master.do` file in the FDZ templates illustrates its structure.
- Use the Stata globals `$orig`, `$prog`, `$data` and `$log` to access your files. These globals are automatically made available for you as well during remote execution via JoSuA as on-site use and remote desktop and refer to the correct storage location. The globals must not be overwritten.
- Do not use `cd` commands or relative path definitions.
- Never use the `set max_memory` command.
- Commands not permitted are censored by JoSuA. In these cases, you'll see the line commented out with the tag `*** CENSORED ***`.
- The master file must contain all analyses programmes used during the project. It is important to add a short description of the contents of the respective routines in the master file after each programme start command. Do not use the `nostop` option.
- The evaluation programmes must be numbered in the order in which they run. If several users are working together on a project, their programmes may be distinguished by adding abbreviations to the file names, e.g. `jd02_describe.do`. This does not apply to the `master.do`.

¹Further information can be found in the [manuals](#)

- All files you generate (result files, analysis data sets, etc.) must be replicable based on your programmes and the original data.
- Programme steps that have already been executed and have not changed since the last submission in JoSuA must be commented out in the `master.do`. Data sets stored in the folders **orig** and **data** can be accessed using JoSuA. Do- and log-files are stored on separate servers though.
- For JoSuA, we recommend to set the linesize to 120 in Internal Use Mode, because that corresponds with the width of the output windows in JoSuA. Use the command set *linesize*.

3.2 Programme Files

- Result files (including graphs) must have the same file name as the programme with which they were generated. For example, `jd02_describe.do` will generate the result file `jd02_describe.log`. The graphs generated in `jd05_graphs.do` can be named e.g. `jd05_graphs_02_boxplot1.gph` and `jd05_graphs_02_boxplot2.gph`.
- The programmes must be documented in German or English language for the privacy review (see [FDZ templates](#)). Please take the following into account:
 - Open a log-file **at the beginning** of every do-file (before every other command and before the content-related description of the do-file)
 - write central information (e.g. author, data set used) at the beginning of each do-file (see also example in `jd02_describe.do`)
 - Use subheadings (such as “Preparation of the variable X” or “Establishment-level aggregation”)
 - Describe the content and function of loops and programme descriptions of tables into your loops (e.g. using the *display* command)
 - Create variable labels and (if appropriate) value labels for all important variables that were newly generated
 - Use meaningful variable names whenever possible
 - Briefly describe self-generated variables directly before frequency tables
 - Describe the content of each graph and complex table (see the examples in the [FDZ templates](#))
- Result tables and graphs produced in Presentation / Publication Mode must be carefully documented (see also examples in `jd02_describe.do` and `jd05_graphs.do`):
 - Each table/graph must have a heading with a consecutive number within a do-file.
 - Each table/graph needs a description with the name and content of the designated variables, the relevant selection criteria and the purpose of the table/graph.
 - All other commands for results that are only relevant in Internal Use Mode, during on-site use or remote desktop access must be deleted or commented out.

- For the output of regression results, we recommend the following procedure (see examples in `jd03_analyses.do`):
 - Combine - if possible - several regressions in one table (e.g. with *estimates table*, *estout*, etc.). If possible, do not output these results twice! You can use “*quietly*” before the regression command for this purpose.
 - If you only need a subset of the regression coefficients for your publication, please only have these displayed; you can have more detailed results displayed in the Internal Use Mode for your review.
 - Please describe any special features that might be relevant for data protection review (fully interacting model, the dependent variable is dummy, etc.). The header of the regression table must clearly indicate which results are involved (if there are several).
 - We accept regression tables as TeX code, but only if it is easy to read; make sure that the columns have fixed widths.
- In principle, evaluation programmes (as well as ado-files sent in) must not contain any data or numbers of observations. External aggregated data sets must be sent in separately as Stata files. This does not apply to classifications and conversion tables (e.g., contribution assessment ceilings, DM-Euro translation, GDP deflators, CPIs, transcoding of industry codes, districts or employment agency districts). Please comment on which classifications and conversion factors you use in your programming. Please note our [manuals](#) on importing aggregated data sets in section 2.2.2.

3.3 Efficient Data Handling

- Use the storage space in your **data** directory sparingly. Once the maximum of 30 GB is exceeded, no more jobs can be processed (except deletion programmes, see `jd07_cleanup.do`) until the storage space is reduced.
- Temporary data sets should only be saved if necessary for subsequent analyses. Data sets need to be reproducible with your do-files at any time. Delete data sets that you generated in the meantime as soon as you no longer need them (see `jd07_cleanup.do`).
- Limit your data set to those variables and observations that you actually need for your analyses (see `jd01_dataprep.do` and `jd04_mergebhp.do`).
- When developing programmes, it might be helpful to test with a subsample (see `jd00_subsample.do`).
- Use the *compress* command at least before saving data sets.
- When generating new variables, please use the most economical format (e.g. *generate byte* or *generate int*).
- The available RAM memory is limited to **30 GB per person and project**.

4 Guidelines for privacy-compliant results

In order to release the results from jobs in Presentation/Publication Mode, a data protection review is carried out at the FDZ in a two-stage procedure: First, a specially developed test script is used to automatically delete critical values. However, this script can only check simple output formats in Stata (e.g. for the commands `tabulate oneway`, `summarize`) and delete critical values. Afterwards, a manual check of all results is always carried out by the FDZ employees.

As a rule, the FDZ **does not make any manual deletions** in the result files, but only checks whether the requirements for the minimum numbers of observations are met or whether too small numbers of observations have been censored by the FDZ's automatic check script. In exceptional cases, the FDZ undertakes manual deletions if it can be justified that a grouping of categories of a variable is not possible in order to avoid the output of too small numbers. Subsequent manual deletions by users are not possible.

You must ensure that your programmes do not generate output with too small numbers of observations. Examples for the design of the do-files are included in the [FDZ templates](#). In the following, the most important guidelines of the FDZ for the data protection review of analysis results are presented. The following examples are intended to illustrate, on the one hand, which results are classified as questionable under data protection aspects and, on the other hand, to help you avoid the output of such results.

4.1 Number of observations in general

- For all descriptive or multivariate analyses, always provide the number of observations they are based on. If the number of cases is missing, the corresponding results will not be released.
- For data protection reasons, all results must be based on **at least 20 observations**. This applies to all statistical figures, cross-tabulations and also multivariate analyses. The minimum requirement of 20 observations of the unit under consideration applies to enterprise-level, establishment-level and individual-level data offered by the FDZ.
- As part of the data protection check at the FDZ, log files or, if necessary, the entire job will not be released if values are contained that are based on fewer than 20 observations (individuals and/or establishments and/or enterprises).

- For analyses of linked data containing enterprise, employer and/or employee data, the numbers of observations for individuals, establishments and enterprises have to be reported in descriptive tables.

4.2 Frequency counts

In frequency counts, values < 20 are usually not released. In cross tables, the number of cases of each individual cell must be ≥ 20 , the total number of the table is not sufficient here. Therefore, if evaluations are based on a too small number of observations, the programmes for their generation must be adapted by you in such a way that insufficient case numbers are no longer displayed. For example, in descriptive statistics, characteristics can be summarised or excluded so that manual deletions are no longer necessary. **Example 1** shows how the categories “100-499 SVB” (SVB = „Sozialversicherungspflichtig Beschäftigte” / Employees subject to social security) and “500-999 SVB” can be combined to prevent the output of values that are too small. Further instructions can be found in the FDZ template `jd02_describe.do`.

Example 1: East Germany

Before				After			
No. employees (SVB)	Works Council		Total	No. employees (SVB)	Works Council		Total
	Yes	No			Yes	No	
1 1-4 SVB	43	1380	1423	1 1-4 SVB	43	1380	1423
2 5-9 SVB	39	547	586	2 5-9 SVB	39	547	586
3 10-99 SVB	594	1322	1916	3 10-99 SVB	594	1322	1916
4 100-499 SVB	573	175	748	4 100-999 SVB	715	175	906
5 500-999 SVB	142	16	158	Total	1391	16	4831
Total	1391	3440	4831				

If it is not possible to group categories for important reasons, manual censoring of the results can be carried out by the FDZ team in exceptional cases. First, all values (number of observations, distribution statistics and regression coefficients) based on a number of observations below 20 have to be deleted. Second, to prevent the re-calculation of deleted values by using subtotals or marginal sums, additional values have to be deleted or rounded.

Example 2 explains the procedure of the FDZ for data protection checks of descriptive cross tables. First, values < 20 are replaced by “/” due to the insufficient number of observations (primary suppression). In order to prevent a re-calculation of the deleted value by using the marginal sum, further values are subjected to secondary suppression (“*”).

Additionally, the re-calculation across several tables must be prevented. **Example 3** illustrates this case. In West Germany, the number of observations in the individual cells is ≥ 20 in each

case, but since the table for all of Germany is also shown in addition to East and West Germany, a re-calculation of the deleted values from East Germany (from example 2) is possible. For this reason, the corresponding cells in the table for West Germany are also censored (secondary suppression: “*”).

Example 2: East Germany

Before				After			
No. employees (SVB)	Works Council			No. employees (SVB)	Works Council		
	Yes	No	Total		Yes	No	Total
1 1-4 SVB	43	1380	1423	1 1-4 SVB	43	1380	1423
2 5-9 SVB	39	547	586	2 5-9 SVB	3*	54*	586
3 10-99 SVB	594	1322	1916	3 10-99 SVB	594	1322	1916
4 100-499 SVB	573	175	748	4 100-499 SVB	573	175	748
5 500-999 SVB	142	16	158	5 500-999 SVB	14*	/	158
Total	1391	3440	4831	Total	1391	3440	4831

Example 3: West Germany

Before				After			
No. employees (SVB)	Works Council			No. employees (SVB)	Works Council		
	Yes	No	Total		Yes	No	Total
1 1-4 SVB	64	2461	2525	1 1-4 SVB	64	2461	2525
2 5-9 SVB	54	847	901	2 5-9 SVB	5*	84*	901
3 10-99 SVB	859	1985	2844	3 10-99 SVB	859	1985	2844
4 100-499 SVB	793	255	1048	4 100-499 SVB	793	255	1048
5 500-999 SVB	198	22	220	5 500-999 SVB	19*	2*	220
Total	1968	5570	7538	Total	1968	5570	7538

Germany Total

Due to the risk of deanonymization, you need to report the number of observations for all subgroups, if the re-calculation of deleted values is possible. In the example above, you need to show the table for both West and East Germany, since the table for Germany allows the re-calculation of values for East Germany. If, for example, instead of an east-west comparison, you want to analyse only the construction industry, the output of all industries of the economy is unnecessary, since no re-calculation is possible.

Before				After			
No. employees (SVB)	Works Council		Total	No. employees (SVB)	Works Council		Total
	Yes	No			Yes	No	
1 1-4 SVB	107	3841	3948	1 1-4 SVB	107	3841	3948
2 5-9 SVB	93	1394	1487	2 5-9 SVB	93	1394	1487
3 10-99 SVB	1453	3307	4760	3 10-99 SVB	1453	3307	4760
4 100-499 SVB	1366	430	1796	4 100-499 SVB	1366	430	1796
5 500-999 SVB	340	38	378	5 500-999 SVB	340	38	378
Total	3359	9010	12369	Total	3359	9010	12369

4.3 Descriptive statistics

At first glance, descriptive statistics do not allow any conclusion to be drawn about the underlying case numbers. However, this does not mean that the output of descriptive statistics, such as means, cannot be problematic. Here, too, the principle applies that the statistics addressed are only classified as safe if the calculation basis comprises at least 20 observations. A special case is the output of descriptive statistics for dummies. In the case of binary-coded variables, the values of these are distributed over only two categories. Even if the total number of observations of a dummy variable is greater than 20, it may be that due to a skewed distribution only three persons or establishments or enterprises fall into one of the categories. In this case, the output is considered uncertain, even if the total number of observations does not directly indicate a low occupancy of a category. This is because this can be easily determined using the means. In order to be able to identify and check these cases, it is necessary to document whether a dummy variable is involved when means are displayed.

Example 4 illustrates the problem with the output of mean values. In the case of the dummy variable r61, all values had to be censored because only 12 establishments ($140 \cdot 0.0857143$) fall into one of the two categories due to the skewed distribution.

Example 4

Table 4.9: Before

Variable	Obs	Mean	Std. Dev	Min	Max
r60	201	2.373134	0.9192794	1	3
r61	140	0.0857143	0.2809469	0	1
r62a	73	2.219178	2.340742	1	15

Table 4.10: After

Variable	Obs	Mean	Std. Dev	Min	Max
r60	201	2.373134	0.9192794	1	3
r61	140	/	/	/	/
r62a	73	2.219178	2.340742	1	15

When using the `summarize` command in Stata, corresponding cases are automatically checked by the FDZ check script and deleted if necessary. If equivalent results are generated using other commands (e.g. `table`, `tabstat`), no automatic deletion can be carried out. In these cases, you must refrain from outputting results with insufficient numbers of cases by deleting the corresponding code from your evaluation programmes.

4.4 Percentiles / quantiles

When displaying percentiles, make sure that **at least 20 observations** are included **in the respective percentile**. A detailed output (1% percentiles) requires at least 2000 observations to guarantee the data protection requirements of the FDZ. In principle, the more detailed the analysis, the more observations must underlie the entire distribution:

- At least 20 observations for mean values (except dummies)
- At least 40 observations for 50% percentiles
- At least 80 observations for 25% or 75% percentiles
- At least 200 observations for 10% or 90% percentiles
- At least 400 observations for 5% or 95% percentiles
- At least 2000 observations for 1% or 99% percentiles

If you want to describe quantiles, the smallest distance between the selected percentages (also to zero and to 100) determines the number of observations needed- e.g. for the percentiles 10 – 15 – 30, the sample must be $(15-10)/100 * x \geq 20 \rightarrow x \geq 400$ observations to ensure the necessary 20 observations in each quantile.

When using the `summarize` command in Stata, corresponding cases are automatically checked by the FDZ check script and deleted if necessary. If equivalent results are generated using other commands (e.g. `table`, `tabstat`), no automatic deletion can be carried out. In these cases, you must refrain from outputting results with insufficient numbers of cases by deleting the corresponding code from your evaluation programmes or adapting it so that percentiles with larger intervals are displayed for which sufficient numbers of cases are available.

4.5 Weighting

When using any extrapolation or weighting factors in descriptive analyses, you have to report the corresponding unweighted results. The weighted results **must directly follow** the corresponding unweighted results because this facilitates the data protection check and thus speeds it up - also for you. This must also be taken into account and commented on in loops. If the unweighted output is missing, the corresponding log file or, if necessary, the entire job cannot be released. Further notes can be found in FDZ template `jd02_describe.do`.

For weighted regressions, you have to provide the unweighted number of observations if `iweights` or `fweights` are used.

4.6 Graphs

You have to provide the underlying numbers of observations for each individual value in a graph. The total number of observations is not sufficient, rather you need to show the **number of observations of each individual bar or point**. You can show the numbers of observations either directly in the graphs or in the log-file by providing tables directly before or after the command for the graph. When displaying very large numbers of data points in graphs, it is not necessary to output the number of cases for each individual data point. Here it is sufficient to prove that the smallest underlying case number is at least 20. You can find an example of this in the FDZ template `jd05_graphs.do`.

The rule of at least 20 observations also applies to graphs. Thus, scatter plots on the individual level are prohibited, because each individual data point is based on less than 20 observations. For histograms, you need to show the number of observations for each individual bar. For kernel density plots, the total number of observations is sufficient. If the number of observations in the corresponding regression table is sufficient, coefficient or margin plots are allowed. In addition, the use of the `asis` option is mandatory when saving the graphs in `gph` format to ensure that the graphs are saved in their current form and that no further editing is possible. Further instructions can be found in the FDZ template `jd05_graphs.do`.

4.7 LaTeX output

Stata result files must have the ending “.log” or “.txt”. Results written in separate files outside the log files created by Stata cannot be released. You can include the results of these separate files in the log file with the command “`type path`”. Make sure that the output is **clearly arranged and easy to read** (e.g. through fixed column widths). If this is not possible in exceptional cases, the results must be displayed again in an easily checkable form directly before the included text. An example can be found in the FDZ template `jd03_analyses.do`.

4.8 Aggregated data

4.8.1 Importing aggregated data into the project

External data on an aggregated level (e.g. unemployment rates by district) may be merged to the FDZ data if they comply with the data protection guidelines. For more information on the requirements and the procedure, please refer to our [manuals](#) (section 2.2.2).

4.8.2 Documentation of aggregation steps within the project

If you aggregate the original microdata for further analysis (e.g. to the level of industries, regions, etc.), please make sure this is properly documented in the log-file (captions; explanatory notes). All log-files created after such an aggregation step (and using the aggregated data set as the main data source) must contain the following information in the top header:

- the level of aggregation,
- the programme step (i.e. do-file), in which the aggregation takes place (please also mark this step in the master file).

If this information is missing or hard to find, the results cannot be checked and released.

If you want to display those aggregated variables later, note that the rules for the minimum number of observations still hold (see chapter 4.1 to 4.3). Therefore, the following requirements apply:

- An additional variable must be created for each aggregated variable to be displayed, which contains the number of the corresponding non-missing observations in each cell.
- If the aggregate variable is a quota, share or rate, tabular representations must also show the number of non-missing values in the individual subgroups used to construct the outcome (e.g. the number of men and women in federal state X alongside the women's share in federal state X).

Examples on how to generate these variables correctly can be found in FDZ template `jd02_describe.do`. The number of aggregated tables should be kept to a minimum to comply with the principle of data parsimony¹.

¹It should be noted here that the automated check script also deletes or censors aggregated outputs using certain Stata commands such as `summarize` or `tabulate`, for example at the district level, despite a sufficient number of persons or establishments or enterprises within the districts. In such cases, we would ask you to contact the FDZ in order to avoid such a deletion of results through adapted programming.

4.8.3 Export/transfer of user-generated aggregate data sets

If you wish to receive a more extensive aggregated data set as a .dta data file because it cannot be properly displayed and reviewed as a table in a log-file (or if you would like to transfer that data set to another project directory), please discuss the procedure with us **in advance**, ideally already when submitting the application. It must be clarified what the purpose of the aggregation is, at which level it is aggregated and how the included variables were generated (sum, mean, etc.). Explain why the export/transfer is necessary.

The generation of the aggregated data can be prepared with the test data and should be done during on-site use. The programmes must then be started again with JoSuA. Please use the Presentation/Publication Mode for this and note in the comment field that you want to receive or transfer a generated data set and the corresponding file name. Always save this data set in the subdirectory **data**.

Since checking aggregated data sets is very time-consuming, the following rules must be implemented when creating them:

- The entire data preparation for this aggregate data set, from the original data to the export data set, should be performed in **one** do-file/log-file. Please omit any unnecessary steps in this do-file. An example can be found in the FDZ template `jd06_export_aggregate.do`.
- The rules outlined in the previous chapters concerning documentation and reporting of number of observations apply equally.
- Use informative variable names, variable labels and value labels for the export variables.
- Avoid stacking aggregations (e.g. state + labour market region + districts), at least if the finest level of aggregation has cells that need censoring.
- Do not add variables for additional subgroups using the 'wide' format. Keep the format 'long'.
- The variables defining the aggregation level have to be named "lvl_1_*, lvl_2_*, ...", e.g. an aggregation to year and federal state should lead to grouping variables `lvl_1_year` and `lvl_2_fstate`.
- Each aggregated outcome variable has to be followed by a variable containing the respective number of observations used for constructing it (if it is not a count variable itself). This variable has to be named the same way but starting with an `N_`.
- Try to keep the aggregated data set as small as possible. In particular, values that can be calculated from other information in the aggregated data set (e.g. totals or ratios) should be generated after export or transfer to simplify disclosure control.

In addition, for the export/transfer of an aggregated data set, you must ensure that each cell is based on at least 20 observations. If this is not the case, you have the following options:

- **Group categories** (preferred by the FDZ). For example, if you want to obtain aggregates at district level, group together districts that have too few cases.

- Censor cells that have too few cases yourself (not preferred by the FDZ). Proceed as follows:
 - If the number of observations in some cells of the data is < 20 , set the count variable and all other outcomes for that cell to missing category “p”.
 - Keep the number of cells with insufficient observations at minimum. Keep in mind that our staff might have to perform secondary suppression (see chapter 4.2).
 - Add an explanation why further aggregation/grouping is not justifiable from a project standpoint.
 - Please note that we reserve the right to decline releasing data for export or transfer if it contains low cell counts that are not appropriately justified and documented, or we decide that the time needed for appropriate disclosure review would be excessive.

For each project, such a data set can be forwarded **only once**. This also applies if you want to have an aggregate data set transferred to a different FDZ project folder in order to merge it with other data sets. The export of the data set is done by email. Transfers to other project directories are carried out by FDZ staff members.

4.9 Regression output

The total number of observations used in regressions must be at least 20. Results of multi-variate analyses based on fewer than 20 observations can generally not be released and should therefore be removed from the programmes.

However, with the help of regressions, small numbers of cases can also be determined indirectly under certain circumstances. Therefore, in these cases it is not sufficient to consider only the total number of observations. The following regressions are problematic:

- Regressions whose coefficients result in the unconditional mean, such as in a regression with only one dummy or a group of dummies, e.g. for the federal states.
- Regressions with two or more variables in fully interacted models (all possible combinations of variable values).

Here it must be ensured that the individual values of the dichotomous or categorical variables are based on at least 20 observations. This is to be shown by frequency tables of all variables after the regression output using `if e(sample)`. Each additional variable in the model no longer allows for conclusions. Further instructions can be found in the FDZ template `jd03_analyses.do`.

4.10 Event data analysis

When using the `sts list` command in event data analysis to identify the Kaplan-Meier estimator results, the value at the beginning of the table (population at risk) is relevant for the data protection check. In addition, when groups are shown in graphs, the `sts list` command must also be used for each group to provide the underlying number of observations.

When results from survival analyses are displayed, only the **initial value must be ≥ 20** , not every single step.

4.11 Sequence pattern analysis

Graphical representations of sequence pattern analyses must be based on a sufficient number of observations. There are two ways to do this. First, you can show sequence patterns per unit of time as proportions of states (e.g. in month 1, 30% of people are in unemployment and 70% in employment). Second, you can depict aggregated sequence patterns for groups if each group contains at least 3 observations. Graphical representation of sequence patterns for less than 3 persons or establishments or enterprises is prohibited.