INEXDA
The Granular Data Network

09.09.2025

Standardized guidelines and best practices for output checking

INEXDA working group on Statistical Disclosure Control (SDC)

Contents

1 Introduction

The INEXDA Working Group on Statistical Disclosure Control (SDC) successfully achieved its first mandate goal by identifying the current needs, procedures, and tools used among INEXDA members. The survey conducted by the WG revealed notable differences in how RDCs implement output control—particularly regarding quantitative thresholds and the types of statistics permitted for release. The findings were published in a document and an interactive dashboard, both available on the INEXDA website [1][2]. WG members decided to advance towards the second goal outlined in the mandate: fostering harmonization in SDC with a proposal for common principles and rules for output checking, especially for cases where identical datasets are accessed across multiple institutions.

This document outlines the standardized guidelines and best practices that researchers must adhere to comply with these requirements within the INEXDA scope. The document is structured as follows: the next section details the anonymization rules that researchers must follow before releasing their output; the third section provides recommendations for best practices to ensure compliance with anonymity standards; and the final section outlines the rules for the publication phase.

2 Anonymization rules [1]

All calculation results that leave the secure environments must be absolutely anonymized. This means that any data that is shared externally must be presented in such a way that it is not reasonable to identify any individual from the data. External researchers are always responsible for ensuring that their results meet the criteria ensuring anonymization and their results don't contain microdata. To ensure this the following rules are required:

2.1 Minimum number of observations units

All the results to be extracted should be based on at least three different (non-imputed values for) observational units. This applies both to aggregate results [2] (averages, medians, etc.) and to charts and tables (at least three observational units per cell/information node). The simplest way to demonstrate compliance with this criterion is to always generate the frequency table associated with each result.

2.2 Degrees of freedom

Regression models must be calculated with at least ten observations, and they must also have at least ten degrees of freedom. Regression results that can be produced with descriptive statistics must also fulfil 2.3 and 2.8, if applicable. Regression models must be calculated with at least ten observations and three parameters, and they must also have at least ten degrees of freedom.

2.3 Dominance Rule

The largest observation unit must not exceed 85% of the analyzed value.

[1] The theoretical framework of this section has been informed by several documents [3,4,5,6,7], which provide the conceptual foundation and scholarly context

[2] From now on, all references to observations concern only to those not imputed

This rule applies to non-perturbed variables presented as magnitudes[3] and frequencies. For ratios, this rule should be checked separately for nominator and denominator. For magnitudes that can take negative values, dominance must be verified using the absolute value of the variable.

Weighted statistics (weighted means, index numbers, etc.) should be analyzed separately for the dominance of the weighted variable and the one used as a reference.

2.4 Confidentiality in multiple tables, control of differences

If results are calculated based on a population (G) but are subsequently recalculated for a subset (X) of G, the rules explained above must be met for observations of the difference. Otherwise, individual observations could be identified based on the differentiation.

Example: *If there is a table with all firms in a given sector and another with firms exceeding a certain sales volume, a third table with firms not reaching that volume must be created to check confidentiality criteria*.

2.5 Dichotomous (0-1) Categorical Variables (Dummies)

When calculating averages of these variables, there should be a minimum of three observation units for each category (three observations with 0 and three with 1)

2.6 Treatment of missing values

Missing values must be excluded when calculating the number of observational units (application of rule 2.1).

If missing values are imputed, the number of non-imputed observations must be at least 3. Imputed observations should be reported.

2.7 Percentiles and median

These measures can be released once the minimum number of observations and dominance rule is met in the interval between the previous percentile and the actual.

2.8 Maximums and minimums

Publishing individual maximum or minimum values is not allowed. However, it is permitted to publish the average of the three highest or three lowest values — including the maximum and minimum — as long as the highest value in that group complies with the dominance rule.

[3] A magnitude should be understood as any measurable characteristic of an observation unit that can be expressed by a numerical value.

3 Best practices for ensuring anonymity compliance and reproducibility

To verify compliance with anonymity requirements, the results provided for review and extraction must be easily verifiable and reproducible. To ensure this, researchers are required to follow these best practices:

3.1 Creation of a master file

Researchers are encouraged to create a master file containing all relevant information about the research project, including calls to all subprograms used, description of the software and data. Additionally, it is advisable to sufficiently comment on the program code so that even individuals unfamiliar with the project can understand it within a reasonable amount of time.

3.2 Specification of Anonymization Compliance

Researchers should include code that justifies compliance with the anonymization requirements described in section 2. This can be achieved by providing frequency tables, demonstrating fulfillment of dominance rules, or including any other elements that show adherence to the requested output standards.

3.3 Limitation on the Output requested for review

Researchers should be prudent in the quantity of output requested for review. Only results with the potential for publication should be submitted. Exploratory data analysis that is not intended for sharing with third parties must be excluded from the output to be extracted.

4 Publication control rules

The following guidelines aim to assist researchers in more easily complying with the publication control standards:

4.1 Referencing sources

Researchers must mention the ultimate data source in any publication resulting from the study, as indicated in the respective guide of each database.

4.2 Copy of Publications

Researchers should provide a copy of any published works that that contain research results from analyses conducted on the accessed datasets.

References

[1] Final_Report_INEXDA_WG_SDC
[2] Dashboard_INEXDA_WG_SDC
[3] Green, E.; Ritchie, F.; White, P. The statbarn: A new model for output statistical disclosure control. Conference contribution (2024)
[4] Green, E; Kendal, C; Ritchie, F; Alves, K. Guide to output checking processes (2024)
[5] Ritchie, F; 10 is the safest number that there's ever been. Journal Article (2022)
[6] Desai, T.; Ritchie, F.; Welpton, R. Five Safes: Designing data access for research (2017)
[7] Ritchie, F. Statistical disclosure control in a research environment (2011)