مـــركـــز الإحـــصـــاء
STATISTICS CENTRE

Statistical Data Editing Guide

Methodology and Quality Guides - No. (5)

5

www.scad.ae

Table of Contents

Introduction

The collection of data from its primary sources is a key process in the production of official statistics, whether the data is collected from administrative records (data sets generated from the management processes of various entities especially Government agencies), such as records of births, deaths, etc.; or data collected directly from the field from the statistical unit under study (individual, entity, ...), or obtained via other means, such as personal, or telephone interview , etc.

The data collected is of paramount importance as the primary source for the provision of indicators and statistics to policy - makers in economic, demographic, social and environmental areas, as well as in measuring the performance of the related policies. The data collected is also made available to researchers and the business community to support their investment decisions, as well as for entrepreneurs, businessmen and investors to make appropriate decisions and to evaluate their investment decisions.

The importance of data editing lies in the fact that it helps to maximize the usefulness of data, making it imperative to ensure that the data used is free of the errors arising during their collection or entry, and is coherent and consistent, since such characteristics have a favourable impact on the quality of the output of such data, which in turn reflects positively on the quality of the decisions based on the data in question.

Before embarking on data collection, a plan for the methodological aspects of the collection process must be set. It must indicate the intended purpose from collecting the data, identify the targeted statistical population , the variables for which the data will be collected and the outputs desired from this data. The plan must also include organizational matters such as determining who will collect and who will edit the data, besides setting a timeframe for the collection process, etc.

In view of the critical importance of statistical data editing, this guide will serve as a general manual for those working in the collection and tabulation of data in various entities, in order to introduce the essentials of data editing for statistical purposes.

1 | Definition of data editing

Data editing is the process of reviewing the data for consistency, detection of errors and outliers (values that are extremely larger or smaller than the rest of the data) and correction of errors, in order to improve the quality, accuracy and adequacy of the data and make it suitable for the purpose for which it was collected. Data editing also includes measures and indicators to assess the level of accuracy of the data, such as the number and percentage of fields with errors detected to the total number of fields in the database. It is notable that databases contain two types of variables: qualitative and quantitative variables. Quantitative variable are either discrete or continuous. Discrete quantitative variable are integers, i.e. whole numbers with no fractions, such as: the number of factories in each city of a given country, the number of car accidents during quarter one, etc. Continuous quantitative variables may contain any values between whole numbers and can therefore include whole numbers and fractions, such as income, student heights in a particular school, etc. Qualitative variables on the other hand have no numerical values, such as gender (male, female), educational attainment, etc. The editing of each type of variables will be discussed later in a dedicated section.

2 | Objectives of data editing

- Detect of errors that would affect the validity of outputs
- Detecting inconsistent values and outliers and adjust them.
- Provide information enabling assessment of the overall level of accuracy of the data
- Validate the data for the purposes it was collected for

3 | Types of data editing

There are several types of editing; they include the following:

Validity and completeness of data

The validity of the data refers to the correctness of the responses obtained, based on the possible range of answers for each variable. This type of checking verifies validity by ensuring the absence of non–numerical answers in fields devoted to numerical answers and vice versa (see Table 1: column 1, row 2: the answer is textual although the entire table is numerical). Data completeness is means ensuring that all of the fields have been filled and that there are no inappropriate missing values for any fields (see Table 1: column 3, and row 3 / no response).

Table (1): A hypothetical example of numerical data

	Column 1	Column 2	Column 3
Row 1	26	22	12
Row 2	Green	8	7
Row 3	84	60	-

Range

The range sets the minimum and maximum expected values of the variable. In this type of editing, the items on the questionnaire are individually checked (the questionnaire is the data collection tool, and includes all the forms used to record or collect the data) to verify that data in a given field are within the boundaries specified for that field. For example, the respondent's age can be checked to ensure that it falls within a set range of 0 to 125 years (see table 2: column 2, row 3 / age is outside the specified range). This means ensuring that all of the fields have been filled and that there are no missing values for any fields.

Table (2): A hypothetical example of age data			
	Column 1	Column 2	Column 3
Row 1	8	22	12
Row 2	55	36	16
Row 3	84	127	61

Duplicate data entry

Verifying that the data of each unit of the register or the database was entered only once, with no duplication, especially when there are variations in some index fields of the unit within the record (see table 3: Sr. No. 1 and Sr. No. 4 / note repeated data for school 1 with a different index number). Another case of duplication is the repeated entry of the index number of a given variable (see table 3: Sr. No. 2, Sr. No. 5 / ID, with different school data).

Table (3) Schools data					
Sr. No.	Index No.	School name	Number of teachers	No. of students	No. of classrooms
1	313	School 1	26	720	40
2	545	School 2	15	363	18
3	633	School 3	31	912	52
4	444	School 4	26	720	40
5	545	School 5	35	1363	78

Logical consistency

Consistency is the presence of logical relationships and interdependence between the variables. This type of editing takes into account the connections between data fields or variables. This check is based on logic. The following is an explanation of the key aspects of the logical consistency check:

1. Logical level checking
   - Examine the consistency of the data internally within the database, i.e. linking interdependent variables together to determine the extent of consistency and agreement. This means checking correspondence between the nature of the field and the types of answers/responses or data, as well, in addition to the agreement between different data or responses, (see Table 4: the answers on social status vs. age for Ahmed, 55 cm for Omar Salim's age and the educational attainment vis-a-vis the age of Muhammad, are all inconsistent answers).

Table (4) Household members' data			
	Age	Educational level	Social status
Ahmad	8	Primary	Married
Salim	55 cm	Diploma	Married
Muhammad	4	Doctorate	Unmarried

   - Examining the data by cross-checking it with other data from different records available within the entity.
   - Checking of data by testing its consistency with an existing time series of the same record (as Table 5 below shows, the educational attainment of Salim in 2014 is consistent with 2013).

(5) Household members' data			
	Age	Educational attainment 2013	Educational attainment 2014
Ahmad	8	Primary	Primary
Salim	55	BA	Diploma
Muhammad	4	Doctorate	Doctorate

- Cross checking the data with databases from other sources (third party data).

2. Revision of totals for the main and sub-figures is to ensure consistency with the other variables, (See Table 6: the total earnings of Company 3 is not consistent with the value of total sales).

Table (6): Monthly operating income (AED)		
	Total sales (in millions)	Total earnings (in millions)
Company 1	8	2
Company 2	15	3
Company 3	11	13

Outliers:

This type of editing follows other checks and is used for the detection of extreme values, based on the distribution of the current data and previous data series, which makes it easier to detect the values that can be considered unusual or extreme, so that they can be checked and verified. (See Table 7: income of employee 2 and employee 6).

Table (7) Monthly income from work (AED)	
Employee 1	18200
Employee 2	440
Employee 3	26100
Employee 4	35000
Employee 5	16300
Employee 6	235500

Other types of data editing

There are types of data editing where the focus is on other checks not discussed above, such as ensuring correct data classification, change in physical addresses, contact details, clarity (i.e. to make sure of numbers or labels are commonly known and easy to read, etc.)

4 | Stages of data editing

The data editing process follows pre-determined rules of scrutiny, including several fundamental stages:

Setting the editing rules

Setting the editing rules is a twofold process, the first includes the instructions given to desk editors to enable them to check responses in order to ensure coherence and consistency in the results.

The second set of rules are the automated validation rules built by establishing logical relations between different variables according to various criteria (social, economic, demographic, etc.). Such relations are then formulated into accurate rules to achieve a high standard of output consistency and quality. This type of editing seeks to detect any errors on the form made during data entry, and is likely to screen such errors. For example, the minimum age for a person to be the head of a household is 15 years or above. The program therefore denies the household head registration if the stated age is below 15 years. When dealing with expenditure data, the average price of a given item can be determined by dividing the total value paid by the quantity purchased. Highest and lowest prices assigned to that item with only values within that range to be accepted.

The data editing stage

The data editing stage can be divided into two steps:

- Manual desk editing

  The traditional method of manual editing is implemented by a specialized editing team that covers all the data under review. If the data is on paper the forms are checked after the data is collected and before it is entered into the database. If the data were collected by electronic means, the forms entered into the database are revised individually. Sample forms are re-entered to verify initial entry and keep errors at a minimum.

- Automated data editing

  In this method, the data is checked all at once after being entered electronically, through computer systems and programs, which incorporate the audit rules that have been identified and applied to these systems and programs beforehand. The editing consists of validating the data entered against such rules to detect errors or determine unacceptable responses. The editing teams submit reports on the errors requiring action for correction. In addition, reports on repetitive errors are submitted to the appropriate team members, so that preventive measures can be taken to avoid them in future.

Data editing is performed at several levels:

- Checking the form items independently without linking them to each other, in which case the items are evaluated based on the scope and coverage.
- This level involves a review of the entire questionnaire, with certain items checked and cross-referenced with other items according to the predetermined validation rules.
- Editing at the database level by checking the unit's data in the record with data pertaining to other units in the in the same record.
- Editing at the level of overall totals, sub-totals and entire questionnaires in order to determine forms with inconsistent or unusual content.

5 | Determinants of data editing

Certain limitations influence the data editing process. These may be summed up as follows:

- Available resources: limited time, budget and human cadres.
- Available software: there are several options for specialized data editing and imputation software. However the software becomes a challenge if it has to be designed by the entity because of the extra effort, time and cost incurred.
- Burden on respondents: one of the critical factors in data editing is the possibility to follow - up with respondents to treat missing or incorrect data, since in most cases the respondent is the more accurate source of information. However, follow-up may prove stressful to respondents.
- Intended objective of the data: the extent of editing should depend to a large extent, on the intended uses of the data. For example, some data does not require extensive editing unlike data with strategic importance in the decision - making process. Moreover, within a given set of data, some items may be much more important than others. Therefore it may be desirable to devote more time and resources to ensure that important data is accurate and free of errors.
- Coordination of the error correction process: the procedures and methods to be followed in handling data errors, such imputing missing data should be included in the survey plan from the start of the project.. The editing process would be of little value, if no action is taken to adjust items where the rules fail. For instance in situations where no follow-up is done with the respondent to correct data, the editor will generally turn to imputation and estimation.

In manual editing, it is necessary to:

- Develop and document the edit rules and procedures to be followed
- Train the editing team
- Establish a mechanism for control and verification of the work progress of the editing team members.
- Establish a method for assessing the impact of the edits on the original data.

In automated editing, it is necessary to:

- Develop and document the editing rules.
- Develop a dedicated software or customize a specialized program, besides regular testing of existing data editing programs.

6 | Data editing guidelines

6-1 General guidelines

- The editing rules must be set by staff members with expertise in data editing, questionnaire design and data analysis.
- Ensure that the editing rules are consistent and free of contradictions
- Taking into account the types of variables (in terms of quantity and quality), when setting the editing rules.
- To ensure enough time is given for completing the various stages of the process (data collection, entry and analysis). A quick check is needed at the end of each of these processes to ensure edits have been made and there are no empty fields within the form's questions.
- In the early stages of editing questionnaires are edited in full. If it turns out there are still errors, a sample of the forms is subjected to editing, the size of which is determined by the expected remaining errors.
- Re-run the desk editing, to make sure that the data are almost error-free.
- The questionnaire must also be subjected to detailed desk editing and, at a later stage go through the automated editing rules built in the data entry programs.
- In the final stage, quick checks are performed periodically to determine whether there are missing or unavailable values that may have been omitted during any of the stages of data editing and processing.

6-2 Knowledge of editing methodology

Before starting any editing assignment, it is important for the editor to be familiar with the editing methodology and receive training in related technical and methodological topics in order to perform competently the editing task and to be able to identify interdependent variables. Below some aspects of editing that an editor needs to have knowledge about are given:

- Purpose of the data collected: whether the purpose is organizational, statistical, .etc.
- Identify the target population: the population from which the organization is collecting the data, e.g. collection of data on establishments in the Emirate of Abu Dhabi, or one of the emirate's regions, etc.
- Identify the data unit, i.e. the statistical unit of observation from which data is collected, e.g. individuals, households, establishments.
- Know the variables for which the data is being collected about: e.g. collection of data on individuals such as age, education, sex; and establishments, such as the establishment's name, capital, etc.
- Know the timing scheduled for data collection; it must be clear whether the data is collected continuously, or for a specified period of time, and whether the collection takes place on a seasonal basis, etc.
- Go through the data coding guide and find out if there are codes for the variables or data and learn the coding pattern and system.

6-3 Editing guidelines

Data editing is a process based on logic, common sense and adherence to a written procedure. Below are some guidelines that must be borne in mind in the formulation of data editingrules:

- Verification of coverage: i.e. ensuring that all fields with mandatory questions have been answered.
- Examining the data for internal and external the consistency of all the core topics of the questionnaire, and related variables to ascertain accuracy and logical consistency, which is achieved by:
  - Ensuring that responses match the questions asked.
  - Checking internal consistency; i.e. consistency between different answers (data) if they are from the same source (e.g. it is not acceptable, for example, to find that the answer is "50 cm" when the question is about age).
  - Checking external consistency: verifying that various answers (data) are in agreement with other sources.
- Ensure that transitions between questions follow the right paths (in the event there are skips and routing to other questions non-sequentially)
- Ensure that answers are free of spelling mistakes and typos.
- Check the coding of answers and coding any text that is not coded according to the applicable coding system
- Correction of data errors: suggested adjustments must be checked after the data is entered, and, if necessary data errors corrected through:
  - Contacting the respondent once again
  - Cross checking the respondent's data with his data from the previous year, if any.
  - Cross checking the respondent's data with the data of a similar household
  - Using the editor's knowledge about the topic.
- Determine the final approval status: data ready to be handed over to the data entry team of data entry in the case of paper forms, ready for analysis in the case of electronic forms.
- Document error types and size of errors and, unsuitable validation rules and any issues observed in the forms to control future amendments and improve the questionnaire design.

- Prepare a list of data errors indicators in order to assess the level of accuracy of the data and continue improvements in data accuracy. Examples of such indicators include:

  1. Number and percentage of erroneous (invalid) data to the total data
  2. Number and percentage of outliers to the total data
  3. Number and percentage of data outside the range to the total data
  4. Number and percentage of missing data to the total data
  5. Number and proportion of consistent data to the total data