Statistical data editing

Rudi Seljak
Statistical Office of the Republic of Slovenia

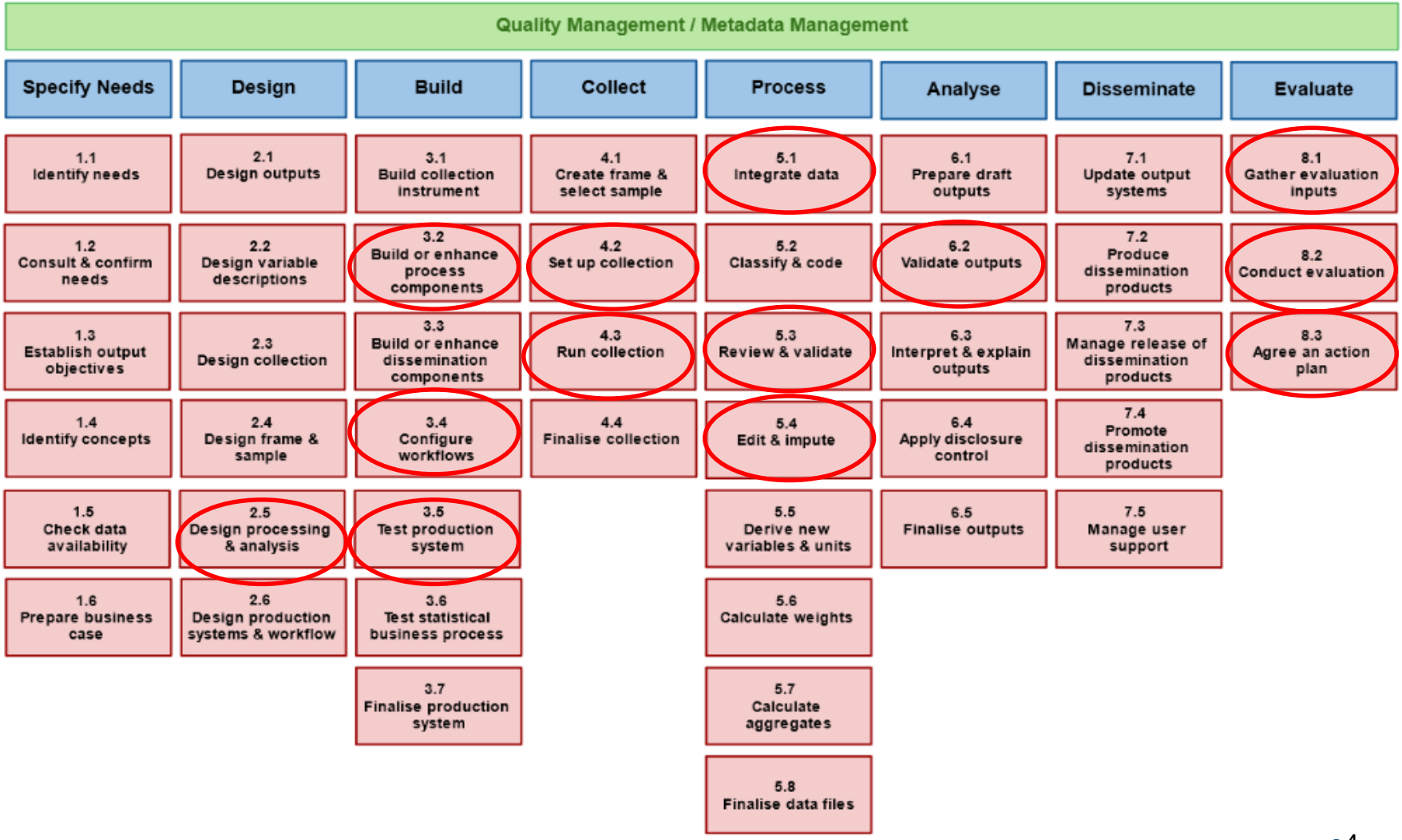
Summary of the content

- Data editing in the statistical process
- Basic concepts and definitions
- Modernisation of the data editing process
- Imputation methods

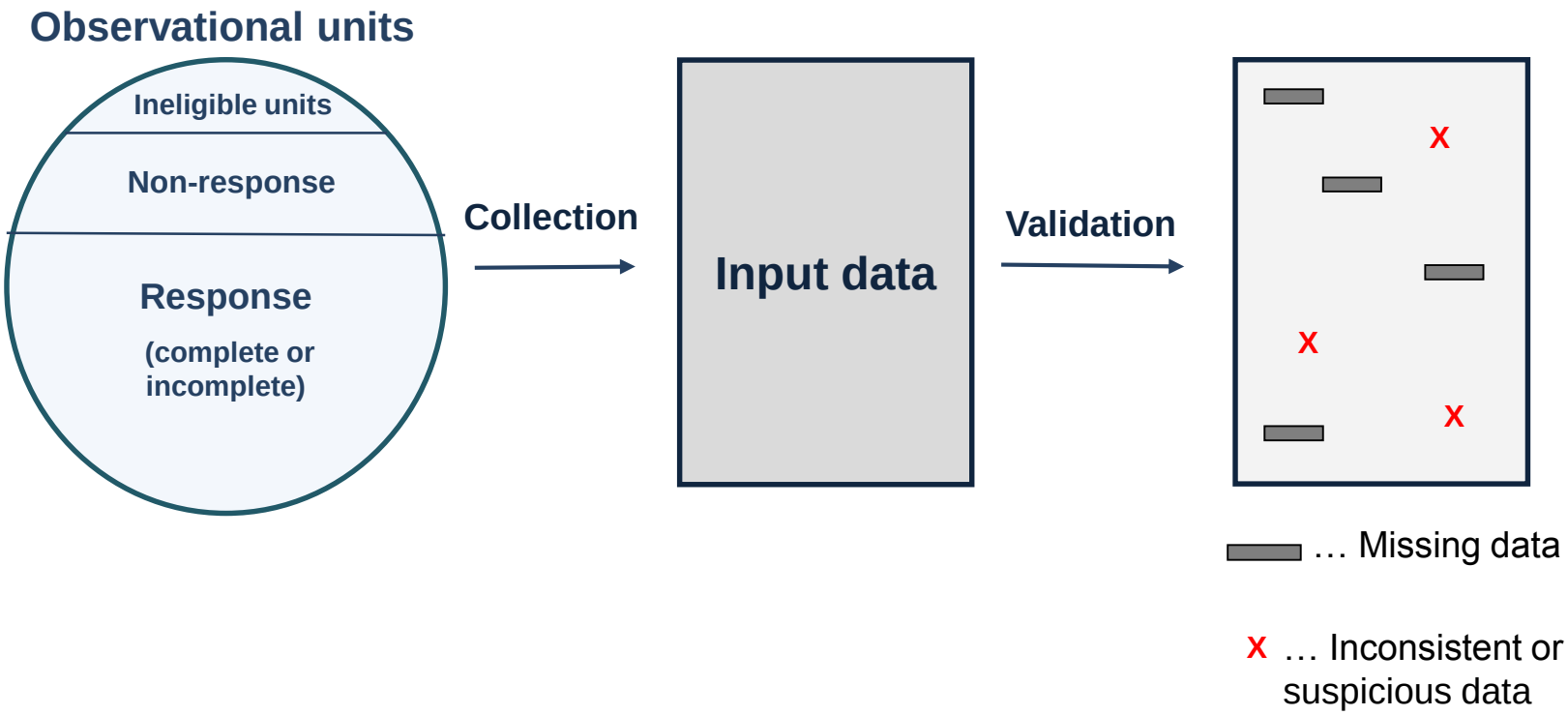
What is data editing

- Data – basic material for statistical production
- Almost each data set „infected“ with errors
- Errors occur in different stages of survey implementation
- Term „data editing“ denotes all activities aiming at detecting and correcting errors in the data

GSBPM model

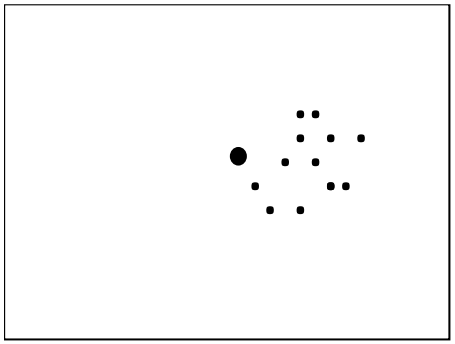

Quality Management / Metadata Management
Specify Needs
Design
Build
Collect
Process
Analyse
Disseminate
Evaluate
1.1 Identify needs
1.2 Consult & confirm needs
1.3 Establish output objectives
1.4 Identify concepts
1.5 Check data availability
1.6 Prepare business case
2.1 Design outputs
2.2 Design variable descriptions
2.3 Design collection
2.4 Design frame & sample
2.5 Design processing & analysis
2.6 Design production systems & workflow
3.1 Build collection instrument
3.2 Build or enhance process components
3.3 Build or enhance dissemination components
3.4 Configure workflows
3.5 Test production system
3.6 Test statistical business process
3.7 Finalise production system
4.1 Create frame & select sample
4.2 Set up collection
4.3 Run collection
4.4 Finalise collection
5.1 Integrate data
5.2 Classify & code
5.3 Review & validate
5.4 Edit & impute
5.5 Derive new variables & units
5.6 Calculate weights
5.7 Calculate aggregates
5.8 Finalise data files
6.1 Prepare draft outputs
6.2 Validate outputs
6.3 Interpret & explain outputs
6.4 Apply disclosure control
6.5 Finalise outputs
7.1 Update output systems
7.2 Produce dissemination products
7.3 Manage release of dissemination products
7.4 Promote dissemination products
7.5 Manage user support
8.1 Gather evaluation inputs
8.2 Conduct evaluation
8.3 Agree an action plan

Data editing – graphical presentation



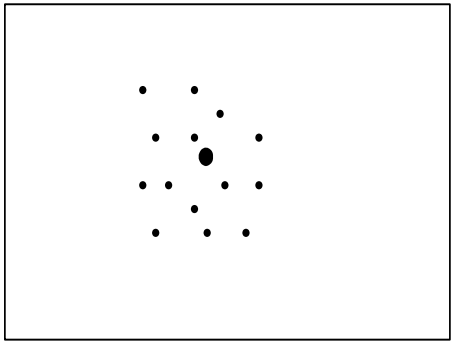
Types of errors

Systematic error



- Measurement unit error
- Misunderstanding error

Random error



- Typing errors
- Rounding errors

Methods for data editing

- **Micro editing**
  - Each individual record is verified and corrected (if needed)
- **Macro editing**
  - Validation of aggregates
    - Aggregated data are analysed and verified
  - Analyses of distributions
    - Usually used to detect outliers
  - Usually placed at the end of statistical process
  - Only used to detect suspicious data → should be followed again by the procedures on micro level

Manual vs. automated editing

- Detection of errors: to a great extent automated
- Correction of errors: still to a great extent manual
- Manual editing (also called interactive editing)
  - Each suspicious unit treated individually
  - Suspicious items verified and eventually corrected:
    - By re-contacting reporting unit
    - Using other (e.g. administrative sources)
    - Using „expert estimates“
  - Very time consuming and costly
    - A lot of space for improvements

Data validation

- Procedure for detection of erroneous or suspicious values
- Based on the set of validation rules
- Set of validation rules is cornerstone of the data editing process
  - Hard rules: Certainly or with high probability detect erroneous data

    Turnover >= Profit

    Age=16 AND Status=‚Student‘
  - Soft rules: Only detect suspicious values

    Number of employees (Y) / Number of employees (Y-1)>1.5

Validation rules

- **Integrity checks –** basic database constraint rules
  - Non-existence of unique identifier
  - Duplicates
- **Plausibility checks –** check if the value is in the prescribed range
  - Gender = {“Male”; ”Female”}
  - Turnover ≥0

Validation rules cont‘d

- **Consistency checks –** we check relations among several variables
  - Status=“Employed” → Age ≥15
  - $Export_{EUR\_area} + Export_{Non\text{-}EUR\_area} = Export_{Total}$ (balance checks)
- **Checks in distribution –** we examine distribution of the variable(s)
  - Mostly used for detection of outliers

Outliers (extreme values)

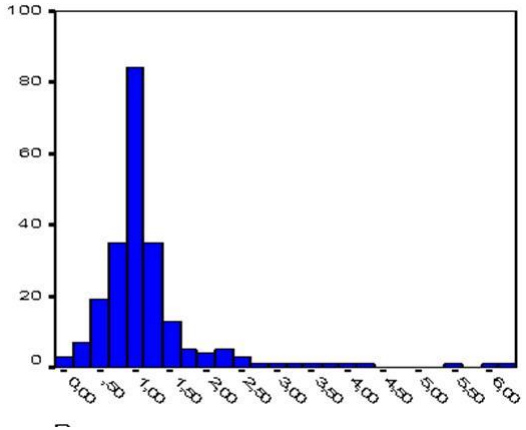
- Value that significantly deviate in the data distribution
- Can be:
  - Correct value
    - Representative outlier
    - Non-representative outlier
  - Incorrect value (e.g. due to the measurement unit error)
- Some statistics are more sensitive on outlying values (e.g. mean) as others (e.g. median)

How to detect outliers

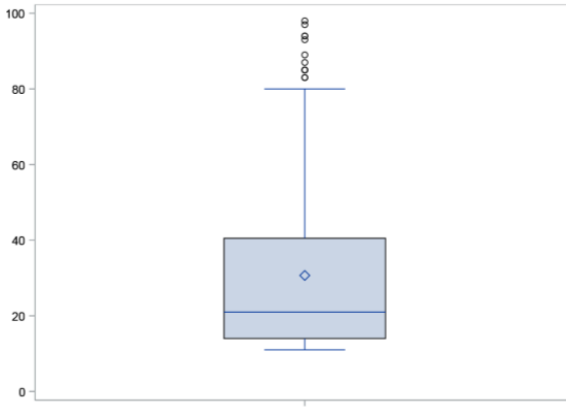
- Graphical methods (manual detection)
- Automatic detection
  - Univariate detection
    - Distance from the median
    - Distribution of ratios with historical values/auxiliary variable → asymmetric distribution → appropriate transformation needed
  - Multivariate distribution
    - Several methods, most of them based on multivariate distance (e.g. Mahalanobis distance)

Types of charts for detection of outliers

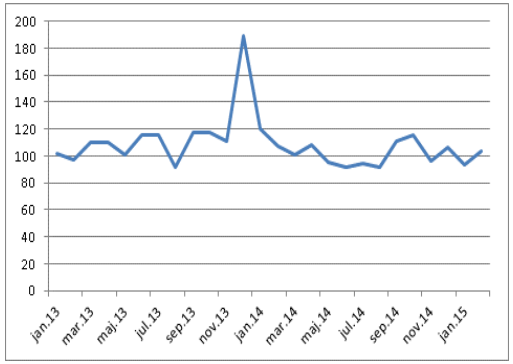
Histogram



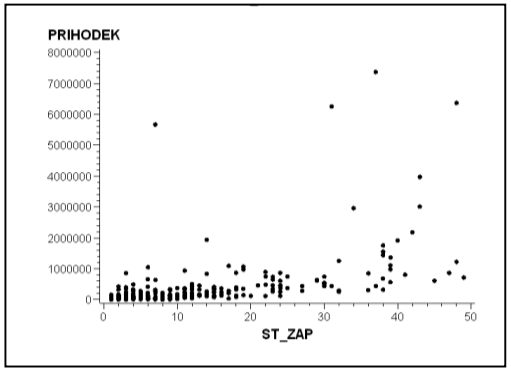
Box-plot



Line chart



Scatter plot



Acceptance region

- The set of acceptable values determined by the set of validation rules
- Each validation rule determines one sub-region
- Rules with one variable: absolute (independent) sub-region
- Rules with several variables: relative (dependent) sub-region

Acceptance region - example

ID	X	Y	Z
1	5	10	4
2	6	4	7
3	11	11	10

Validation rules :

Z>5 and Z<10 ;

X>0 ; Y>0 ;

X>Y

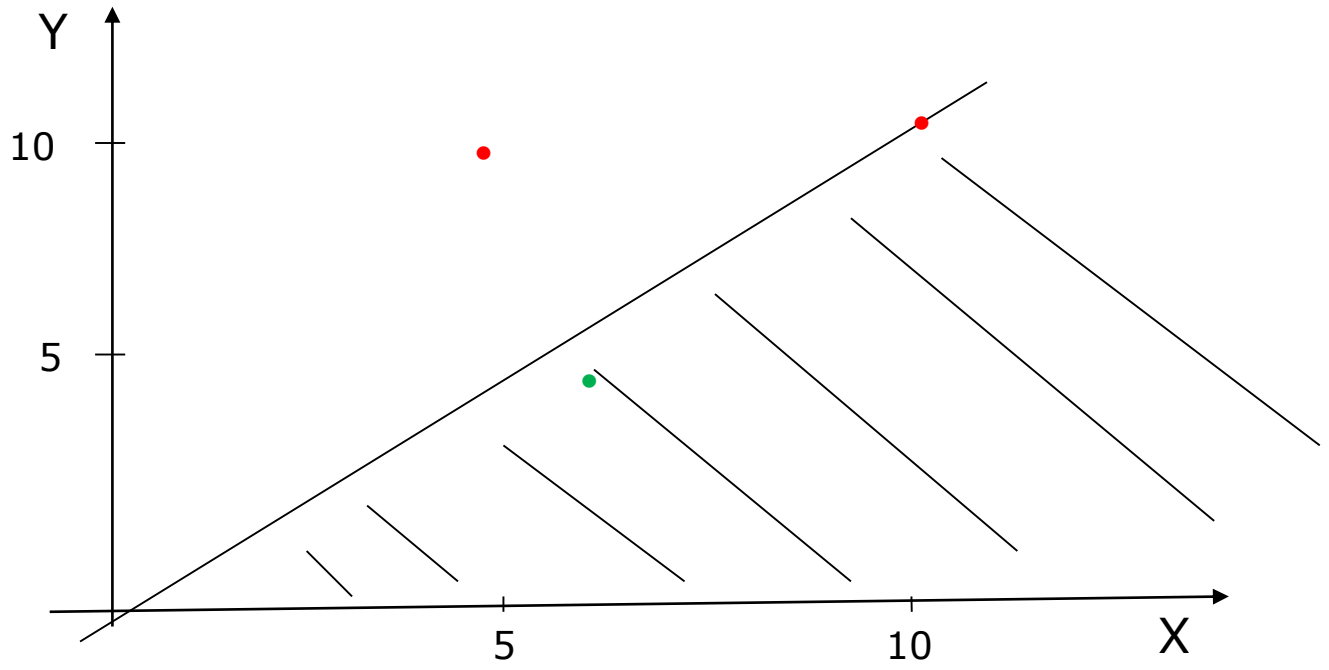
Acceptance region – example cont'd

Acceptance region:

Z : (5,10)

X: Unit 1: (10, ∞) ; Unit 2: (4, ∞); Unit 3 : (11, ∞)

Y: Unit 1: (0,5); Unit 2 : (0,6); Unit 3 : (0,11)



Categorisation of data validity

- Correct/Incorrect
  - Correct: data would be confirmed as plausible, if could be checked with any disposable means
- Acceptable/Suspicious
  - Acceptable: data have passed all the validation rules

n – number of edited units

	Acceptable	**Suspicious**
Correct	n_1	n_2
Incorrect	n_3	n_4

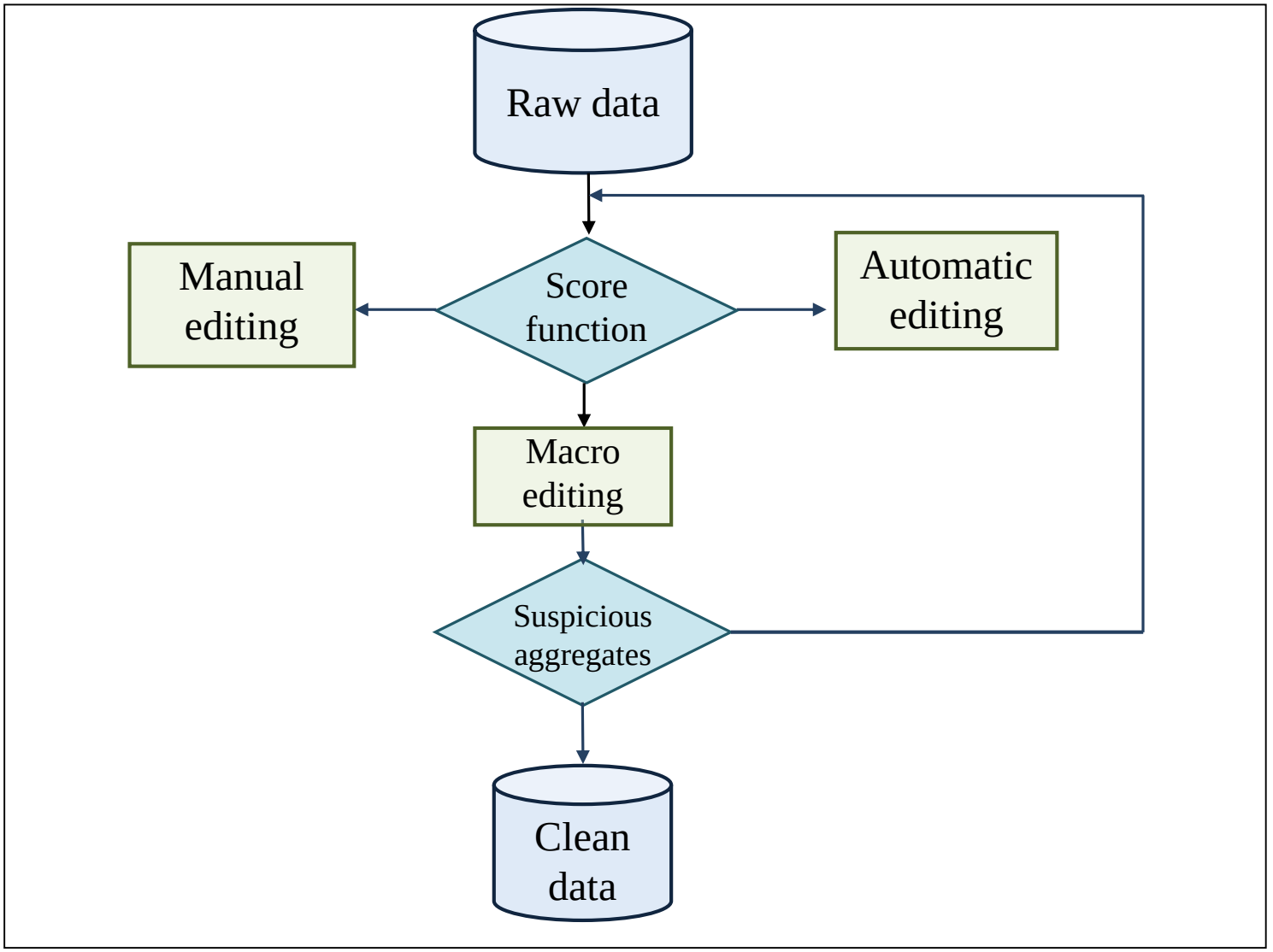
Overediting

- Too many validation rules with too strict conditions
- Too much interactive verification (manual editing)
- Negative consequences
  - Input resources are not in proportion with improvements of quality
  - If too many edit rules have to be verified → the significant errors can be left in the data
- Remedies:
  - Careful planning and testing of controls
  - Restriction of individual verification to the influential errors

Selective editing

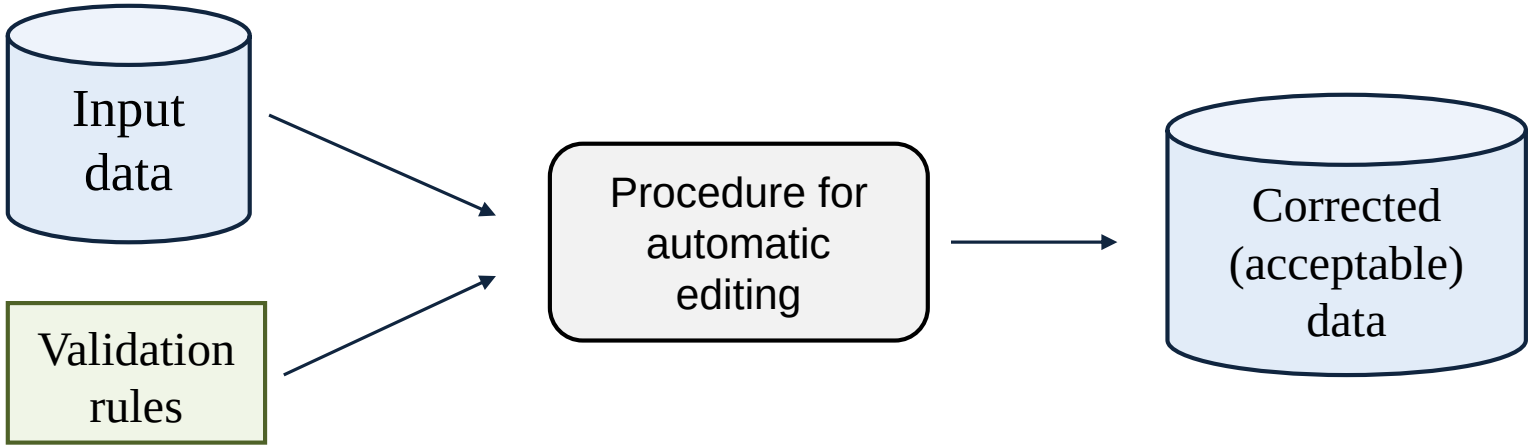
- Procedure for selection of significant errors
- Based on the difference between reported and expected value
- Score function is the mathematical tool to detect influential error
  - Local score: score for particular variable
  - Global score: combines the local scores into the score of the unit
- Selective editing can significantly improve efficiency of data editing

Implementation of selective editing –basic schema



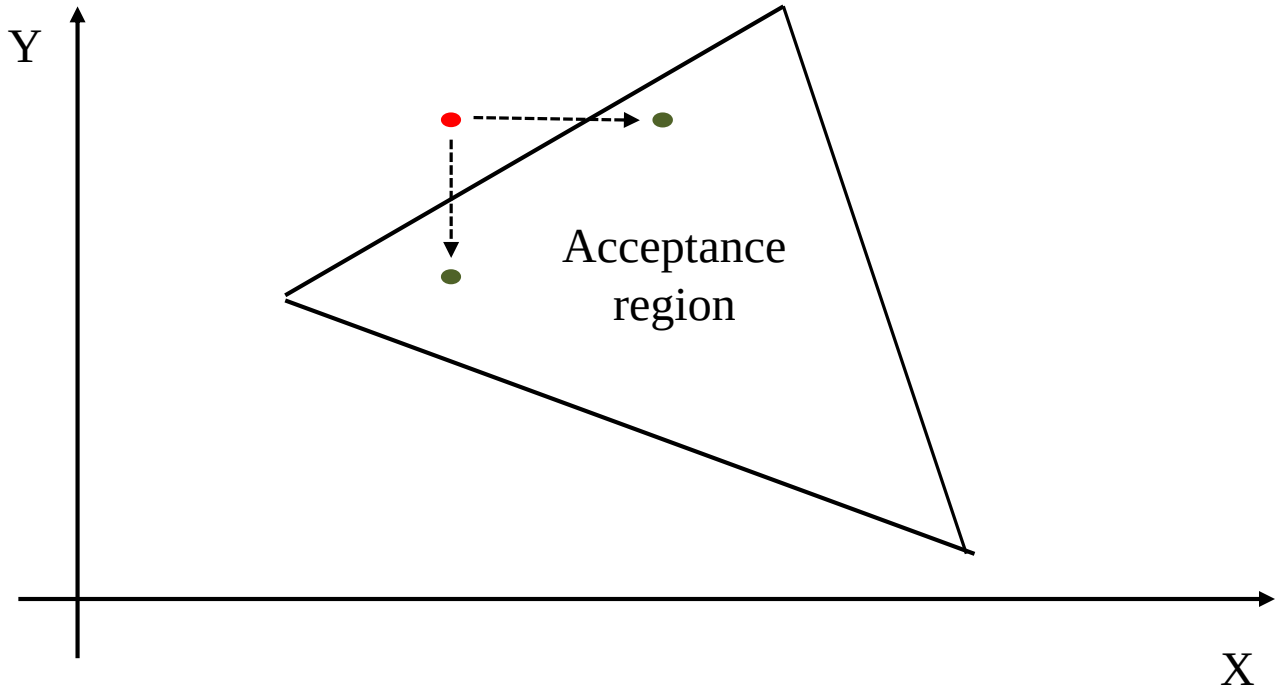
Automatic editing - fully automated procedure

- Correction are implemented only on a basis of a set of validation rules



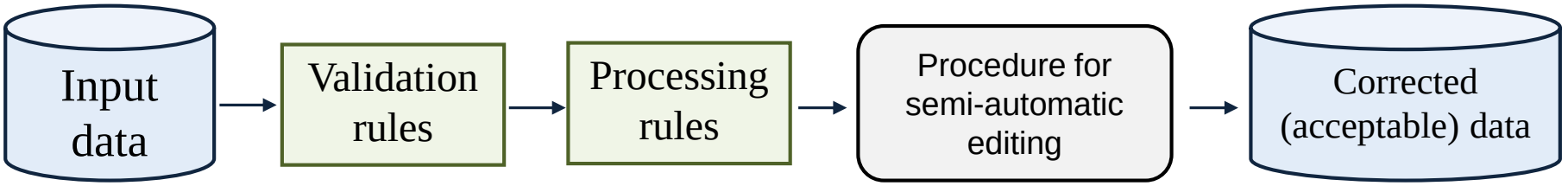
- The most known approach for automatic editing:
  - Fellegi-Holt approach (minimum change approach)

Fellegi-Holt approach – graphical presentation



Automatic editing - semi-automated procedure

- Corrections are implemented on the basis of the processing rules defined and provided by the statisticians



- Processing rules:
  - Deductive correction rules
  - Imputation rules

Imputation methods

- Mean imputation
  - Mean value of responding units imputed
  - Decreases variability
  - Sensitive to outliers
    - Trimmed mean
    - Median
- Ratio imputation
  - If highly correlated auxiliary variable on disposal
  - Average ratio of responded units calculated first
  - Average ratio multiplied with the value of auxiliary variable of „imputed unit“

Imputation methods cont'd

- Hot-deck imputation
  - Imputed value taken from (appropriately selected) donor
  - Several implementations:
    - Random hot-deck
    - Sequential hot-deck
    - Distance based hot-deck
- Historical imputation
  - Used in the case of periodical surveys
  - Reported value from previous period(s) used
  - Several implementations:
    - Average trend imputation
    - Donor trend imputation

Imputations – final remarks

- Imputations can be used to replace missing or erroneous data.
- The aim is not to predict the correct value
  - The main goal is to reduce bias
  - The focus should be on macro level
- Imputation should always be performed by automated procedure
- Each imputed valued should be appropriately flagged.

Data editing and quality dimensions

- Editing have significant impact on quality dimensions
  - Accuracy
    - Improvements of accuracy is a basic goal of data editing
    - More editing → higher accuracy ???
  - Timeliness
    - Important factor in trade-off between timeliness and accuracy
    - Especially significant in short-term business surveys
  - Costs and burden
    - Costs reduction is the main driver of improvements of editing procedure
    - Reducing number of re-contacts → reducing respondent‘s burden
  - Implicit impact to other dimensions:
    - Coherence, Comparability

Toward effective statistical editing

A few guidelines

- Include data editing in different stages of the statistical process
- Invest in building an optimal set of validation rules
- Combine manual and automatic editing
- Each editing process should be accompanied with the set of performance indicators:
  - Edit failure rate
  - Hit rate
  - Data editing impact
- Each survey cycle should be followed by the:
  - Analysis of the data editing process
  - Improvement actions