Data corruption, correction and imputation methods.

Yerevan 8.2 – 12.2 2016

Enrico Tucci – Istat

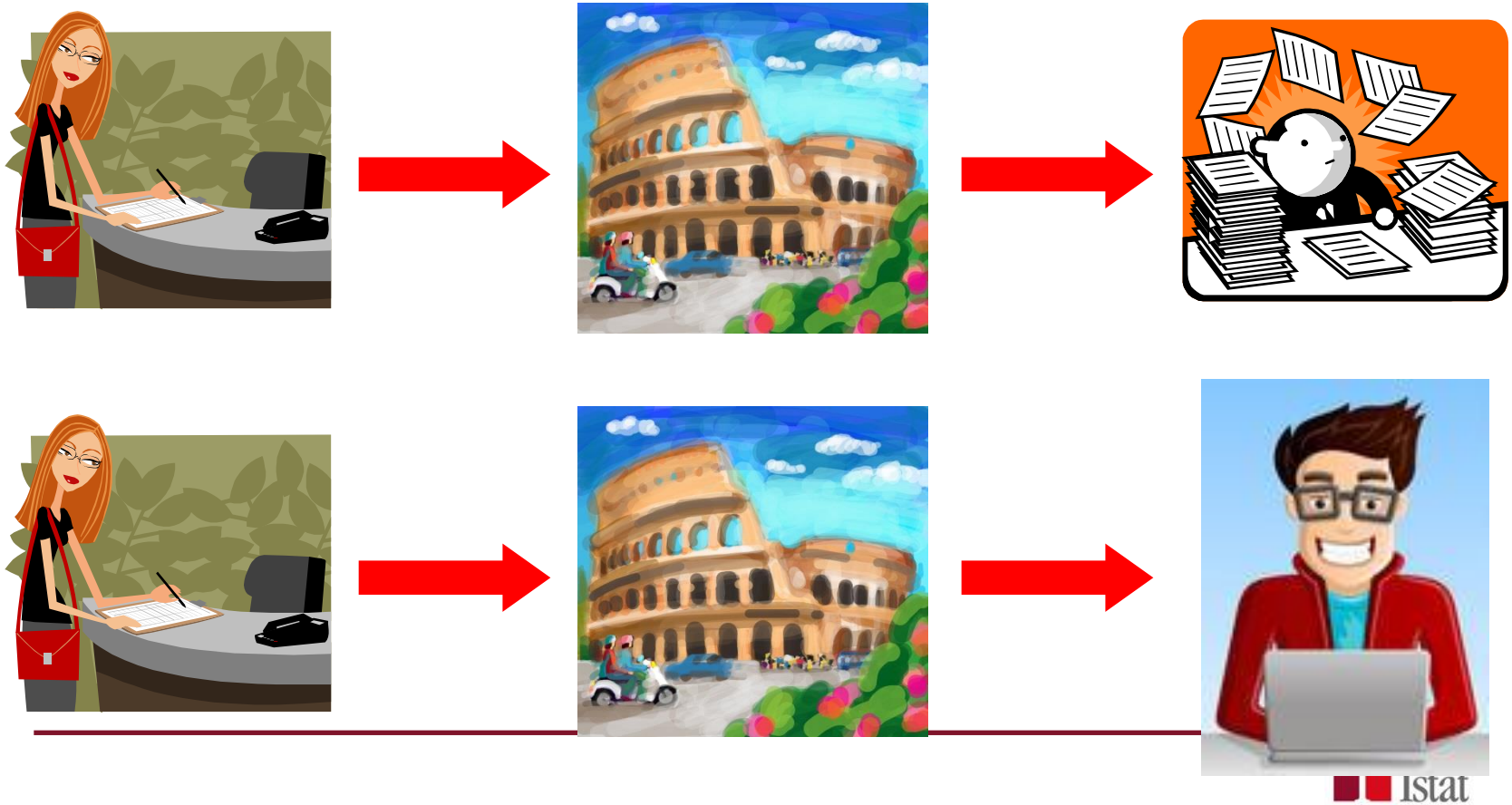
Istat

Outline

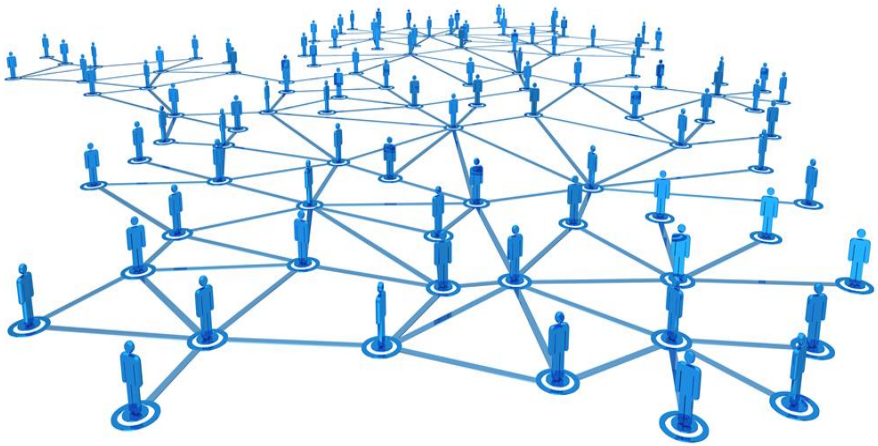
- Data collection methods
- Duplicated records
- Data corruption
- Data correction and imputation
- Data validation

Data collection methods

- From paper to electronic data



From local to central database



Data corruption 1

Data corruption refers to errors that occur during writing, reading, storage, transmission, or processing, which introduce unintended changes to the original data.

Computer, transmission use a number of measures to provide data integrity

Data corruption/2

Validation rules:

- **Size check**. Check the number of characters in a data item value
- **Format checks**. Data must conform to a specified format.
- **Consistency check**. Codes in the data items which are related in some way can thus be checked for the consistency of their relationship.
- **Range check**. data must lie within maximum and minimum preset values.

What to do with found errors?

- The correct value can be deduced from other answers on the same record
- It is necessary to look back at the original data source: if the error is found, the reason for the error needs to be localized

    In case the correct value is entered the correction should be traced

What to do with found errors?

- If the error cannot be corrected, it should be marked as a missing value. Later, the missing values may be replaced by imputed values.

- N.B. Imputations do not make the data set correct but more usable

Missing data and imputation

Missing data is a source of error in any data set requiring correction as they can lead to serious problems in statistical analysis. Imputation methods can be used in order to fill those gaps and provide a complete data set.

It is usual to distinguish missing data caused by unit non-response (total non-response) and missing data caused by item non-response (partial nonresponse).

The former is usually corrected by imputation whereas the latter is usually dealt with by reweighting or estimation methods.

Simple imputation methods/1

Deterministic imputation:

it refers to the situation, given specific values of other fields, when only one value of a field will cause the record to satisfy all of the edits.

It imputes a missing value by using logical relations between variables and derive a value for the missing item

- For instance, it might occur when the items that are supposed to add to a total do not add to the total. If only one item in the sum is imputed, then its value is uniquely determined by the values of the other items.

Simple imputation methods/2

Imputation is a method to fill in missing data with plausible values to produce a complete data set.

Mean imputation

1. Unconditional mean imputation – The missing values are replaced by the mean of the observed (i.e., respondent) values.
2. Conditional mean imputation – Respondents and non-respondents are previously classified in classes (strata) based on the observed variables and the missing values are replaced by the mean of the respondents of the same class.

In order to avoid the effect of outliers, the median may be used instead of the mean. For categorical data, the mode is used for the imputation.

Simple imputation methods/3

Regression imputation

It involves the use of one or more auxiliary variables, of which the values are known for complete units and units with missing values in the variable of interest.

Deterministic regression imputation: this method replaces the missing values by predicted values from a regression of the missing item on items observed for the unit.

Simple imputation methods/4

Hot deck imputation: missing data are replaced by values drawn from similar respondents called “donors”.

Hot Deck Imputation is a simple way is to impute for each missing item the response of a randomly selected case for the variable of interest. Alternatively, imputation classes can be constructed, selecting donor values at random within classes. It was useful to create homogeneous strata (imputation cells), where both the recipient and donor must belong.

Nearest-neighbour imputation: or distance function matching, is a donor method where the donor is selected by minimising a some ‘distance’

the stratification and matching variables must strongly characterize the observations.

Other techniques for data imputation…

Data validation

Outlier detection

An outlier is an extreme observation in the sense that it is surprisingly different from the other observations, leading one to think that it may have been generated by errors due to measurement, collection, coding, recording, transcription, processing … or model.

Exploratory data analysis

Examination of the main characteristics of the data based on graphical displays and on numerical measures and coefficients will most likely uncover the majority of the problems the data may have.

Conclusion

- Electronic data improve the quality
- We will always find errors
- We can use imputation methods to fill gaps
- We need to choose the method that better fits our needs
- Always check the result of the imputation