Data Quality

on the quality of the dataset

Lucas Mello Schnorr, Jean-Marc Vincent

INF/UFRGS
Porto Alegre, Brazil – October 30th, 2017

DATA PRODUCTION

Global Process

$$\text{Question} \Longrightarrow \text{Experiment, Survey} \Longrightarrow \text{Decision}$$

$$\text{Decision} = \text{Risk}$$

Quality of Data

Specification of the Data

- Error model for the values
- Experimental/Survey bias
- Analysis limitations

Evaluate the Quality of the Decision

Criteria for the Quality of Data (from EuroStat)

Relevance

- degree to which statistics meet current and potential needs
- could extend to varying needs

Accuracy

- Closeness of computations or estimates to the (unknown) exact or true values
- Variability (random error) and bias (systematic error)
- Sources of errors (experimental, coverage sampling...)

Timeliness

- delay between the reference point and the availability date
- trade-off against accuracy,

Criteria for the Quality of Data (from EuroStat)

Comparability

- measuring the impact of differences in applied statistical concepts and measurement tools/procedures when statistics are compared between geographical areas, non-geographical domains, or over time

Coherence

- adequacy to be reliably combined in different ways
- compatibility of measures

Accessibility

- Accessibility refers to the physical conditions under which users can obtain data
- Clarity refers to the data's information environment

Extracted from *Handbook on Data Quality Assessment Methods and Tools* EuroStat Report (2013)

Other Criteria for the Quality of Data (from Berti-Equille (2007))

Interpretability

- availability of the supplementary information and metadata
- covers the underlying concepts

Unicity

- one physical observation is represented by a unique object in the Dataset
- no duplicates

Conformity to Norm

- use the standardized encoding (reals, strings, statistical variables)

Consistency

- duplicated informations have the same value

Pre-processing of Data

Before any analysis : check the Data

Question on the Quality

- Are there missing values ? almost yes
- How many sampling are missing ?
- Is there a bias for missing data or randomly spread ?
- Is the bias in the dataset sufficiently important to modify the analysis (estimators, tests,...) ?

Give potential explanations

Identification of Data Problems

Model of the Dataset (types, semantic,...)

- Missing Data (none or partial value)
- Non relevant
- Duplicated

Give potential explanations

Université Grenoble Alpes

PRE-PROCESSING OF DATA (2)

Distributions of Data Problems

Analyse the position of missing values in the Dataset

- MCAR, Missing Completely at random (unpredictable missing)
- MAR, Missing at random (predictable values : model)
- MNAR, Non missing at random

Processing Missing Data

- Do nothing
- Remove samples with missing values
- Weighted analysis
- Value imputation (central tendency, EM, regression, random hot deck, neighbouring,...)

Report the method that has been used

UNIVERSITÉ Grenoble Alpes