

Data Characterization

on the nature of observations

Lucas Mello Schnorr, Jean-Marc Vincent

INF/UFRGS
Porto Alegre, Brazil – October 21th, 2017



DATA PRODUCTION

First question : Why this dataset has been produced ? (purpose)

- ▶ Who organized the study ?
- ▶ What was the question to be answered by the statistical analysis ?
- ▶ Who will be the target of the analysis ?

Second question : Which approach has been used ? (method)

- ▶ Exhaustive collected information
- ▶ Designed survey on a population
- ▶ Designed Experiments

Third question : How this dataset has been practically produced ? (observations)

- ▶ Nature of the items in the Data set
- ▶ Characterization of data
- ▶ Semantic of Data

Take time to analyse the production process

ANALYSIS OF THE SET OF VARIABLES

Identification of the variables types

- ▶ Type of the variables (numbers, identifiers, ...)
- ▶ Set of values taken by the variables (bounds, sets,...)
- ▶ Properties of the variables (positive,...)

Identification of the variables role

- ▶ When these variables has been collected ?
- ▶ Why these variables have been chosen ?

Identification of the variables semantic

- ▶ What is the interpretation of the variables values ? (size, weight, ...)
- ▶ What are the relations between variables (structure) ?

Take time to build a serious metadata document

ANALYSIS OF THE TYPE OF VARIABLES

Nominal Variables : classification, membership (qualitative)

- ▶ Values in an unstructured set
- ▶ Examples : color, gender, ...
- ▶ Methods : grouping
- ▶ Operators : $=$, \neq

Ordinal Variables : Comparison, Level (qualitative)

- ▶ Values in an ordered set
- ▶ Examples : ranking, opinion, ...
- ▶ Methods : sorting
- ▶ Operators : \leq , \geq

Quantitative Variables : Quantities

- ▶ Real values (ratio is significant)
- ▶ Examples : amount, duration, cost ...
- ▶ Methods : sum, difference
- ▶ Operators : $+$, $-$, $(\times, /)$

Take time to define precisely the variables properties

USAGE OF VARIABLES

Response Variables

- ▶ Quantity asked by the question
- ▶ Examples : response time, iteration duration, ...

Explanatory Variables

- ▶ Variables that could affect the response variable
- ▶ Examples : size, load, ...

Univariate or Multivariate

- ▶ Univariate : one variable is involved
- ▶ Multivariate : several variables are involved

Take time to identify the response/explanatory variables