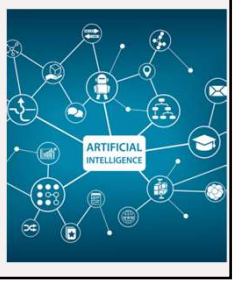
Análise de Dados em Informática

Engenharia Informática

Ano Letivo: 2023/2024

Ana Madureira



1

Data Science vs. Big Data vs. Data Analytics

1

https://youtu.be/X3paOmcrTjQ

2

“

Data science is a field that deals with unstructured, structured data, and semi-structured data. It involves practices like data cleansing, data preparation, data analysis.

3

3

“

Big data is high-volume, and high-velocity or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. data analysis, and much more.

4

4

“

Data Analytics is the science of examining raw data to reach certain conclusions.

5

5

Exploratory Data Analysis

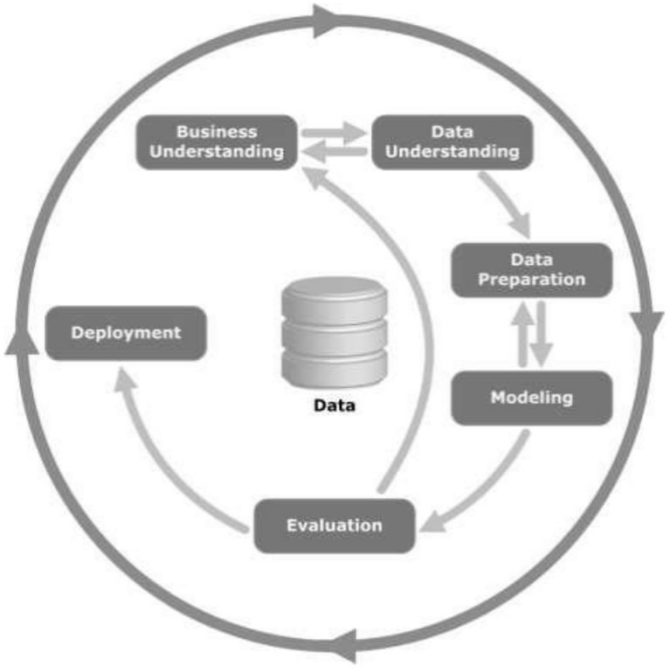
Data Analysis in Informatics

- Usability Analysis
- Reliability Analysis
- Software Quality
- Performance Analysis
- …

6

6

CRISP-DM methodology that stands for Cross Industry Standard Process for Data Mining



7

7

Statistics

The statistic aims to provide information (knowledge) using numerical quantities, divides the study and analysis of the data (numerical facts) into three phases:

1. Data collection
2. Description, classification and presentation of the data ⇒ **Descriptive Statistics**
3. Conclusions from the data ⇒ **Statistical Inference**

8

8

Statistics

1-Descriptive statistics

- Refers to the description and organization of data
- Set of techniques that aim to synthesize and represent in an understandable way the information contained in the data
- We often use summary measures (averages, medians, variances), tables and graphs to synthesize and represent information.

2-Statistical inference

- Corresponds to a set of techniques that aim to make estimates and draw conclusions about a population from the information contained in a sample.
- It serves to draw conclusions from the reality of a whole, starting from the knowledge of a part. Ex: Estimation of parameters and hypothesis tests.

9

9

Descriptive statistics

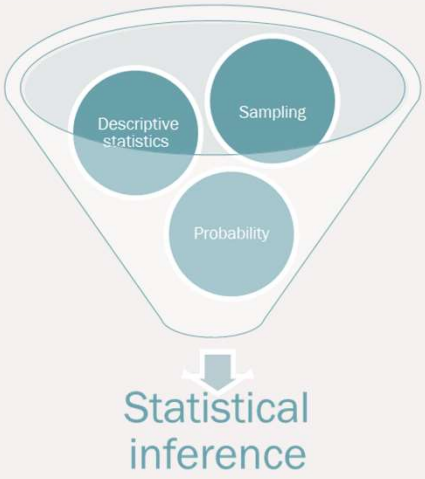
Objective: To represent the information contained in the data in a synthetic and organized way.

- Percentages
- Graphical analysis
- Frequency Tables
- Measures of Central Tendency (mean, median, mode)
- Dispersion measures (standard deviation, variance, amplitude)
  - How to summarize the characteristics of large datasets

10

10

Statistical inference



Objectives:

- control and quantify inference errors
- make the best of it (minimizing errors) of available information
- dimension the information necessary to guarantee pre-specified error levels
- regulate the processes of collecting information

Types of inferences:

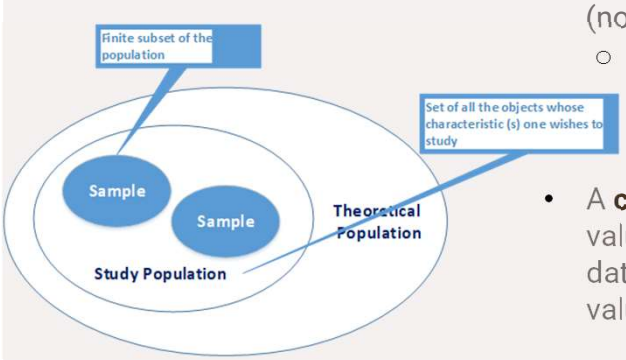
- Estimation of parameters
- Hypothesis Tests

Draw conclusions on populations based on the results observed in the sample(s)

11

11

Basic Concepts



- **Statistical variable**: represents a characteristic of a population that can take several possible values. It can be **qualitative** (non-numerical) or **quantitative** (numerical).
  - Example: the height of a person, their weight, the number of accidents in a place, etc.
- A **class** is any of the variable's possible values (discrete qualitative or quantitative data) or any range of possible variable values (continuous data).
- The number of elements the sample is referred to as of **sample size**.

12

12

Basic Concepts

- **Qualitative variable** - It is a variable whose measurement scale only indicates its presence in exhaustive and mutually exclusive discrete classification categories.
  - Nominal (eye color, sex, race, ...)
  - Ordinal (education, satisfaction, ...)
- **Quantitative variable** - It is a variable whose measurement scale allows the ordering and quantification of differences. Examples: daily number of births in Portugal, the temperature at noon in a certain location, height, age or weight.
  - Discrete (number of children, number of failures, ...)
  - Continuous (waiting time, weight, ...)

Note: Qualitative variables are discrete. Quantitative variables may be discrete or continuous.

- **Discrete variable**
  - It is a variable that can take only a finite set or an infinite number of values / categories.
  - Examples: the number of accidents, the number of births, the number of people, etc., ...
  - Definition 0.10 –
- **Continuous variable**
  - It is a variable that can take values from a real range or breaks.
  - Examples: the temperature of a given material, height, weight, etc., ...

13

13

Basic Concepts

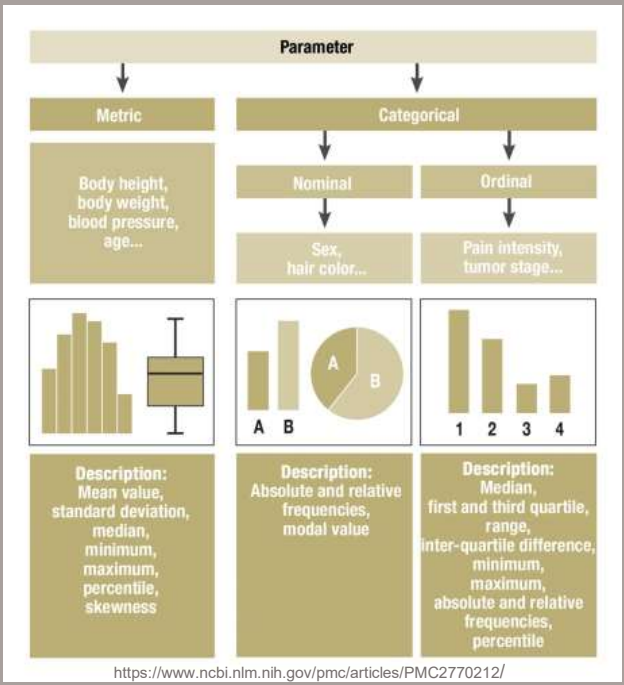
There are 4 scales or measurement levels of the variables:

- **Nominal scale** - Elements are attributes or qualities. It is not possible to establish any kind of quantification or ordering at the outset. For example: Gender (F or M).
- **Ordinal scale** - The variable assumes different categories with a relation of order between them, according to a descriptable but not quantifiable relation. For example: Socio-economic strata (low/medium/high).
- **Interval scale** - These scales do not have absolute zero, i.e. they do not have the absence measure. For example: Temperature in degrees
- **Ratio scale (absolute)** - Similar to the interval scale but zero has real existence, it corresponds to the absence of the measured characteristic. For example: Weight, Height, ...

14

14

Basic Concepts



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2770212/

The property of a parameter is specified by its so-called **scale of measure.**
Generally two types of parameters are considered:

- **Metric or quantitative** - A variable has a metric level (= quantitative data) if it can be counted, measured or weighed in a physical unit (as in cm or kg) or at least can be recorded in whole numbers.
  - **continuous and discrete variables.**
- **Categorical or Qualitative** - The gender cannot be measured but is classified into two categories. Parameters which can be classified into two or more categories are described as categorical parameters (= qualitative data). A further classification of a categorical parameter is into nominal characteristics (unordered) and ordinal
  - **Nominal or Ordinal**

15

Exploratory data analysis

Sampling

↓

Huge amount of information (raw data)

- They may contain high potential, but usually have little interest, given the difficulty of handling associated with the volume of data collected.
- Synthesizing (or reducing) and representing comprehensively the information contained in the **raw data** uses the **methodology of descriptive statistics**
- The **raw data** result from the observation of quantitative variables (discrete or continuous) or qualitative (discrete), then classified as **qualitative or quantitative data** (**continuous or discrete**).

16

Organization of qualitative data

Tables

- Frequency tables

Graphical Representation

- Barplots (% or absolute frequencies)
- Pie charts

17

17

Organization of qualitative data – Frequency Tables

Região de residência

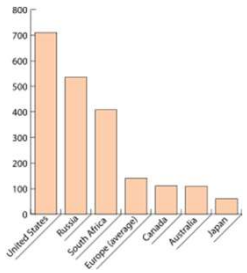
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Norte	63	20,9	20,9	20,9
	Centro	68	22,6	22,6	43,5
	LVT	65	21,6	21,6	65,1
	Alentejo	56	18,6	18,6	83,7
	Algarve	49	16,3	16,3	100,0
	Total	301	100,0	100,0	

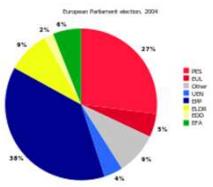
Frequency(ni)

Relative Frequency=fi=ni/n

Relative frequency accumulated (Fi)

Does not make sense to consider cumulative frequencies





18

Organization of discrete quantitative data

Tables	• Frequencies tables (simple and accumulated)
Graphic Representation	• Barplots (% or absolute and cumulative frequencies) • Pie charts • Boxplot

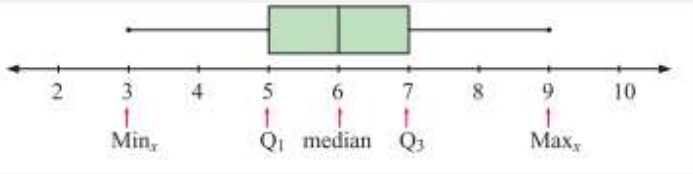
Frequency tables and graphs (barplots and Pie charts). It makes sense here to consider cumulative frequencies.

19

Organization of discrete quantitative data - Boxplot

A graphical mode that allows to easily interpret the location, the dispersion and the asymmetry of a data set, simultaneously making its synthesis → the diagram of extremes and quartiles

- Sample minimum Min_x but not less than Q1-1.5 (Q3-Q1)
- Sample maximum Max_x but not more than Q3 + 1.5 (Q3-Q1)



- The rectangular box represents the 'middle' half of the data set.
- The lower whisker represents the 25% of the data with smallest values.
- The upper whisker represents the 25% of the data with greatest values.

20

Organization of continuous quantitative data

Tables

- Frequencies tables (simple and accumulated)

Graphic Representation

- Barplots (% whose height of the bar represents, e.g. the average)
- Histogram
- Boxplot

Frequency distribution, absolute frequency and relative frequency are defined in the same way as for discrete quantitative data. Data is presented in classes.

21

Exploratory Data Analysis

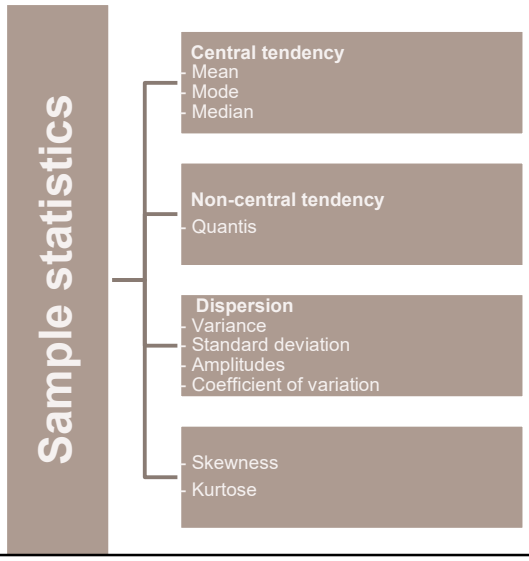
Summary of graphical representations of most frequent use for each type of variable (measurement scale).

Measurement range	Graphic representation
Nominal (without any relation of order) - Species; genre	Circular diagrams Bar chart (single frequencies)
Ordinal (ordainable but not quantifiable) - Hierarchical level; degree of satisfaction (Likert scale)	Circular diagrams Bar chart (simple and accumulated) Boxplot (in some situations)
Quantitative (ordinal, being possible to quantify the differences) - Height; temperature; number of cells	Histograms Boxplot Averages with S.E. (standard error of mean)

22

Exploratory Data Analysis

The sample statistics allow to summarize important characteristics of the samples.



23

Measures of position or central location

If you want to characterize a numeric set of data by a number, a typical value is chosen, which is usually a value around which distribute the data.

- The **arithmetic mean**, for a set of *n* observed values is

  $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$ for unclassified data

- The **arithmetic mean**, for a set of *n* observations grouped/classified in *c* classes is:

  $\bar{x} = \sum_{i=1}^{c} x_i n_i = \sum_{i=1}^{c} x_i f_i$ for the classified data, x_i represents the class mark

- Properties of arithmetic mean
  - Uses all observations.
  - It is easy and quick to calculate.
  - It is a value between the maximum and minimum of observations, but pose is not one of the possible values.
  - It is sensitive to extreme values (outliers). For this reason, the truncated mean is used (mean after eliminate extreme observations).

24

Measures of position or central location

- The **median** divides in half the set of observed values (after being ordered in magnitude - ascending or descending order).
- **Median** properties
  - It has little meaning for a small number of observations.
  - It is not very sensitive to outliers.

- The **mode** is the most common value of a set of observations

Or

- The mode, Mo, of discrete or continuous non-clustered data is the observed value of higher frequency.
  - The modal class is the highest-frequency class.
  - It may not exist (amodal joint) and if it exists it may not be unique.
  - If the frequency polygon only has "1 peak" the distribution is unimodal.
  - If you have multiple peaks it is called multimodal (bimodal in the case of 2 peaks).

25

Measures of variability

Central location measurements do not provide sufficient information about the data set. **Measurements of variability** help to visualize frequency distributions. They provide information on the dispersion of sample values.

- **Range r**: is the difference between the largest of the values of the data set of observations: $r = \max \{x_1, x_2, ..., x_n\} - \min \{x_1, x_2, ..., x_n\} = x_{(n)} - x_{(1)}$
- **Interquartile range** r_q: $r_q = q_3 - q_1$
  (reflecting the variability of only half of the observations).

26

26

Measures of variability

- The standard deviation *s* indicates the proximity to which the values are grouped around the average.
- A small value of the standard deviation means that the observations are scarcely scattered around the average.
- The variance s^2 is the square of the standard deviation.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$$

$$s = \sqrt{s^2},$$

27

Measures of variability

- Coefficient of variation
- Coefficient of variation
- Measurement of relative variability

  - Allows to easily compare frequency distributions since it is dimensionless (without units).

$$c_v = \frac{s}{|\bar{x}|} \times 100\%$$

28

Descriptive Statistics

Descriptive statistics (location and dispersion) measures used most frequently in each type of variable (measurement scale).

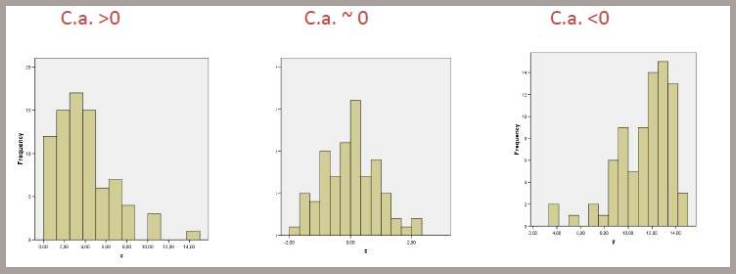
Measurement Scale	Descriptive Statistics	
	Measures of central and non-central tendency	Measures of dispersion
Nominal (without any relation of order) - Species; genre.	Mode	None
Ordinal (ordainable but not quantifiable) - Hierarchical level; degree of satisfaction	Mode Quartis	Range (in some cases)
Quantitative (orderly, being possible to quantify the differences) - Height; temperature; number of cells.	Mean Mode Quartis	Range Standard deviation Coefficient of variation

29

Measures of asymmetry (Skewness)

Asymmetry coefficient - is a measure that assumes zero value when the frequency distribution of the sample is completely symmetric and assumes values other than zero (positive or negative) when the distribution is not symmetric.

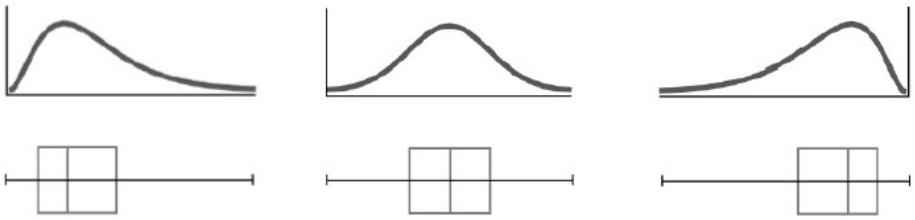
- Note that in a sample it is almost impossible to observe pure symmetry (coefficient of asymmetry = 0).
- It is dimensionless, which facilitates comparisons between different frequency distributions.



30

Measures of asymmetry

- Positive asymmetry: mode <medium <average
- Negative asymmetry: mean <median <mode
- Pure symmetry: mean = median = mode
- Approximate symmetry: mean ~ median ~ mode (in cases where mode calculation makes sense)



31

Measure of Kurtosis

- The kurtosis measures synthesize information about the weight of the distribution tails.
- The interpretation of this measure is not, in general, easy. This is how their value is compared to the normal curve.
- The kurtosis coefficient (C.k.) of the normal curve is 0. Thus
  - If C.k.> 0 the shape of the distribution is slimmer and with a heavier tail than normal.
  - If C.k. <0 the shape of the distribution is more flattened and with a heavy memos tail than normal.

32

Association measures (correlation coefficients)

- Quantify the intensity and direction of the association between pairs of variables.
- Linear and ordinal correlation:
  - Pearson correlation coefficient (r)
  - Spearman Correlation Coefficient (ρ)
  - Correlation coefficient V of Cramer (V)
  - Correlation coefficient Phi (ϕ)

33