BIG DATA STATISTICS FOR BUSINESS

AY 2023-24

Giovanni De Luca
Parthenope University of Naples

1

Data Handling

- Data analysts have to devote considerable time to "preparing" the data before carrying out the statistical analysis.
- Data Handling (or data pre-processing) aims to transform raw data into data that can be effectively analyzed.
- Data Handling is usually applied to a large amount of data and is a mix of ad-hoc interventions and automatic (therefore reproducible) procedures.

2

2

Data Handling

- We deal with three cases:

1. Extreme values (outliers)
2. Missing data
3. Inaccuracies

3

3

1. Extreme values (outliers)

- There is a large literature on the detection of outliers and many definitions of outliers exist.
- Barnett and Lewis (1994) define outlier "an observation (or a set of observations) not consistent with the set of data".
- Hawkins (1980) defines an outlier as "the observation (or set of observations) that is so different from the others as to allow us to hypothesize a different generating mechanism".

4

4

Extreme values (outliers)

- Outliers can have several causes. The most common causes are:

1. Human error in data collection (can sometimes be corrected).
2. Voluntary alteration by survey participants (often linked to sensitive issues).
3. Sampling error (some sample units come from a different population than the target population).

- An outlier can result from one of these causes or indicate large variability in the data.

Identification of outliers

- Failure to detect an outlier can lead to an incorrect specification of the model, distorted parameter estimates and therefore incorrect results.
- There are graphical and analytical techniques for detecting outliers in numerical variables.

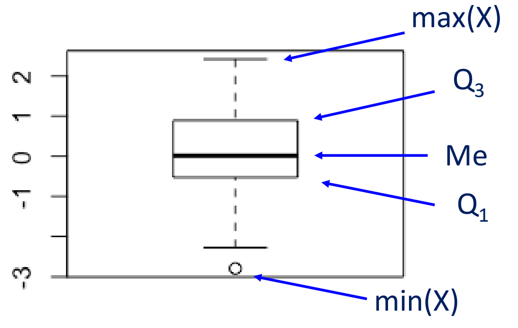
Box plot

- The **box and whiskers plot** (or **box plot**) is a graphical representation to describe the distribution of a set of data through simple indexes.
- Given a variable X, we compute:
  - min(X)
  - 1° quartile of X (Q_1)
  - Median of X, Me(X)
  - 3° quartile of X (Q_3)
  - max(X)

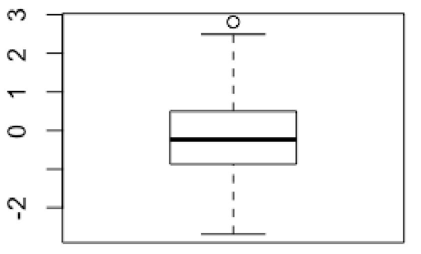
Outliers: box plot

- In a box plot used to detect the presence of univariate outliers, whiskers are **not** plotted at the observed minimum and maximum values.
- The lower whisker corresponds to $Q_1 - k(Q_3 - Q_1)$
- The upper whisker corresponds to $Q_3 + k(Q_3 - Q_1)$
- In practice, k is generally set equal to 1.5, but higher values can be used.
- Values outside the whiskers indicate the presence of possible outliers.
- The quantity k defines the sensitiviness of the plot.

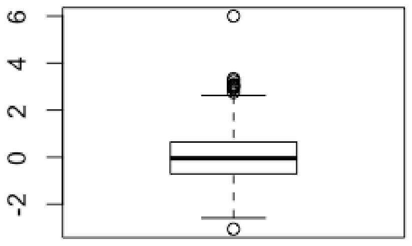
Box plot (example 1)



Box plot (example 2)



Box plot (example 3)



Outliers

- After detecting an autlier, should we remove it or keep it?
- The decision is not easy. The knowledge of the context can help to take a decision.

2. Missing values

- Possible reasons for the lack of data:
  - misfunction of the equipment
  - error in the data-entry step
  - …
- There are two possible approaches:

1. "Leave out records with missing data"
2. Imputation: process of estimation of missing values.

13

13

Imputation process

- The imputation process essentially depends on the availability of auxiliary information.
- There are 2 important approches of imputation.

14

14

Imputation process

1. **Cold-deck** imputation that consists in estimating:

   - the numerical missing data as the arithmetic mean or the median of the variable (the mode if the numerical variable is discrete)

   - the categorical missing data as the mode of the variable

Imputation process

1. **Hot-deck** imputation that consists in estimating:

   - the numerical missing data as the arithmetic mean or the median of the variable using the similar cases (the mode if the variable is discrete)

   - the categorical missing data as the mode of the variable using similar cases

   Alternatively, we could also randomly sample the numerical or categorical missing data using the similar cases.

Imputation process

***k*-Nearest Neighbour (k-NN)**

This method can be used for the imputation process of a categorical variable using the information from numerical variables.

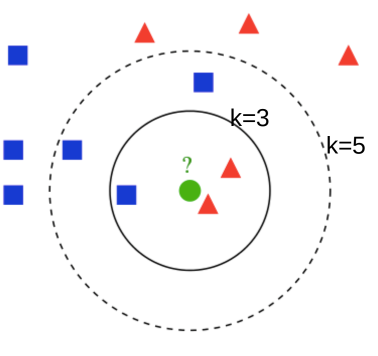
- It consists of an algorithm that identifies *k* most similar observations to the missing one.
- The input is the *k* nearest cases.

17

17

Imputation process

- The k-NN technique follows the "Learning by analogy" approach ("*Tell me who your friends are, and I'll tell you who you are*")



18

18

Imputation process

- The procedure is sensitive to the value of *k*.

- k too small ➔we make a decisions based on very few cases
- k too high ➔ many points of other classes are included.

- A possible solution is

$$k = \sqrt{n}$$

19

19

Imputation process

- We compute the distances of the numerical variables for the unknown category from each observation
- The distances are sorted (in increasing order).
- We use ordered distances to select the *k* nearest neighbors.
- Then, we apply the majority rule.

20

20

3. Inconsistencies

- The inconsistency does not have a precise definition. Anything that is not consistent or logical falls into this case.
- For example, the categorical variable *Gender at birth* has two categories: male and female. If categories are more than two, it is necessary to do a merge (for example, the categories are male, female, M, F).
- It is possible to find *Pepsi Cola* and *Pepsi*. We need to merge.
- A negative height is an inconsistency and has to be removed.