ANALYST

Co-funded by
the European Union

MODULE 2/Pre Processing and Data Exploration

Carlo Drago PhD

University Niccolo Cusano, Rome

MASSIVE OPEN ONLINE COURSE (MOOC)

Project N. 2023-1-IT02-KA220-HED-000161770
ANALYST - A New Advanced Level for Your Specialised Training

HOZINT HORIZON INTELLIGENCE
fyg consultores
UNIVERSITY OF WESTERN MACEDONIA
Centro Studi AMIStaDeS APS
UNIVERSITÀ CUSANO
European Institute for Local Development
UNIVERSIDAD DE ALMERÍA

Co-funded by the European Union

PRE-PROCESSING AND DATA EXPLORATION IN DATA ANALYTICS

This module introduces three relevant topics: data preprocessing, data description and exploration

- **Data preprocessing** it is fundamental to handle eventual problems data can have
- **Data description** is important to obtain from the data a relevant “picture” which can be used for the analysis
- **Data exploration** is very useful to identify “the data structure” but also on evaluate if some assumptions for data analysis can be fullfilled or not (for instance normality)

See also Giudici (2005)

DATA QUALITY

In real cases data can be characterized by problems like:

- Missing Data
- Data with errors
- Outliers

These problems need to be handled and considered before the analysis. We refer to these problems as problems on “data quality” (on the their importance see Jain et al. (2020).

The quality of the data, in turn, determines the reliability and validity of any analysis that can be done. Enhancing Data quality is a process of constant refinement that can lead to better decision-making.

See Little & Rubin (2019)

MISSING DATA

Missing data are a typical characteristics of the typical real data

- Missing data are single observation not having any value
- These observations can be typically dependent on some phenomenon inducing the missingness or they can be MAR "missing at random"
- Statistical methods are not designed typically to work with datasets with missing data
- If data are not MAR but MNAR "missing not a random" a possible strategy it could be impute the data.

See Little & Rubin (2019)

DATA WITH ERRORS

Data with errors are another typical characteristics of modern data. There are two notable elements to take into the account:

- The errors on data can lead to inconsistent analyses
- It is necessary to handle these type of errors and correct the data if is possible.

These errors can significantly affect the accuracy and consistency of analyses. The well-known phrase "garbage in, garbage out" emphasizes the relevance of the data quality in obtaining reliable results. Therefore, it is important to identify, manage, and correct these errors whenever possible to ensure the validity of the analysis results.

See for instance Teh et al. (2020).

OUTLIERS

Outliers are another relevant issue on the data

- Data with outliers can be problematic for statistical methods
- They are not necessarily "wrong" data but they need to be considered with care
- It is possible to repeat the analysis with and without outliers (sensitivity analysis)
- It is possible as well to use robust methodologies to avoid statistical problems linked to the presence of outliers on the data

See Aguinis et al. (2013).

DATA RECODING

Data can be usefully recoded

- The recoding is useful when modalities are not sufficient to express some relevant real characteristics
- Recoding can be done in different ways (use of qualitative variables with different classes and also creation of dummy variables)
- It is also necessary to transform sometimes textual variables in qualitative variables
- There can be ordinal and not ordinal variables

DATA TRANSFORMATIONS

Typical transformations are:

- **Normalization** allows the data to fit into a specific range, typically between 0 and 1. This measure ensures that all features can contribute equally, especially when dealing with data from different scales.
- **Standardization**, on the other hand, involves rescaling the data with a mean of zero and a standard deviation of one. It benefits algorithms that assume normally distributed data, such as specific regression models.
- The **logarithmic transformation** stabilizes the variance and compensates for skewed data by applying a logarithmic function.

FEATURE ENGINEERING

A relevant point in the realm of the analysis with large datasets is to manage many variables. In many cases there is the need to synthesize different variables, priorly to the analysis. This can be done with different analytical techniques

- **Principal Component Analysis**
- **Composite Indicators**
- This also includes the creation of **new characteristics**, such as the calculation of ratios, differences, or interactions between variables.

Analysts can improve model performance and gain deeper insights into the data by carefully selecting and transforming features.

EXPLORING DATA (EXPLORATORY DATA ANALYSIS)

Data exploration is a very relevant issue in data analysis. Sometime we do not need to confirm any hypothesis and we just to analyze data with the aim to discover the data structure of the dataset considered. In this sense this type of analysis can be useful to create hypotheses for future studies. Relevant aims are:

- Discover the data structure
- Identifying relevant outliers
- Assess the reasonability of the hypothesis of the analytical methods used

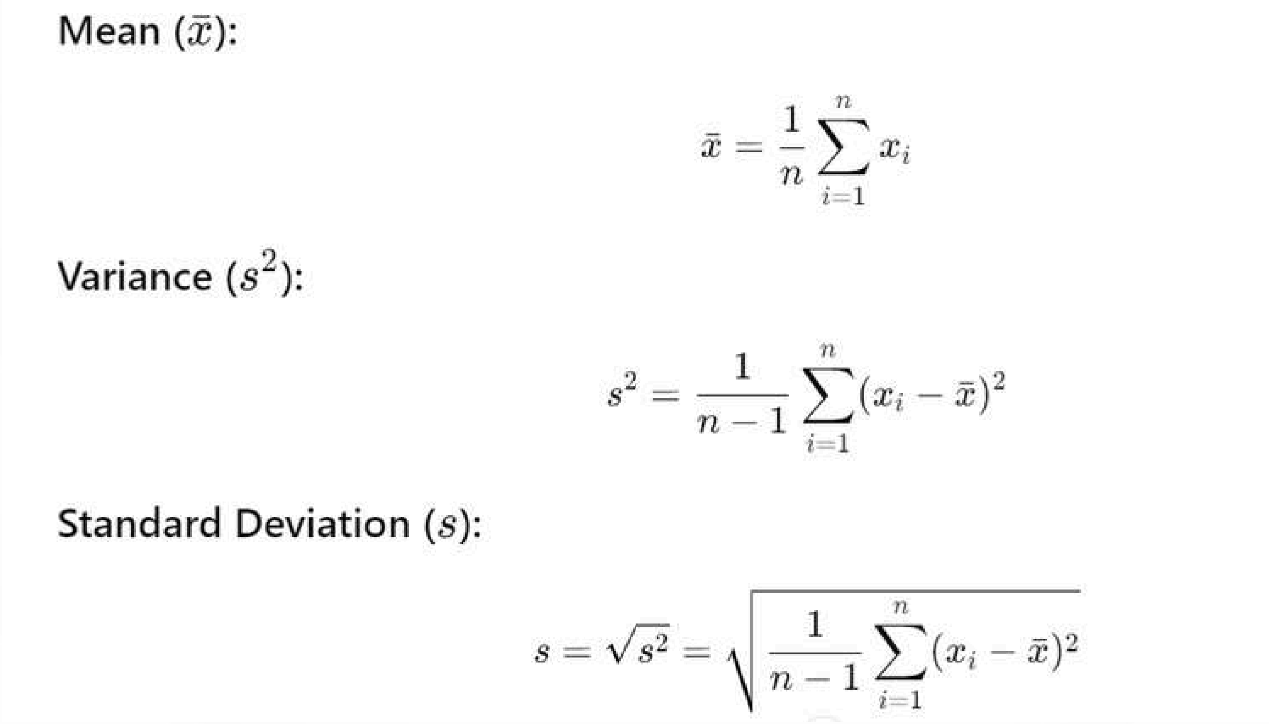
Tukey (1977)

DESCRIPTIVE DATA ANALYSIS

A very relevant approach in data analysis is also the descriptive analysis of the datasets. In this case we are just interested to show the relevant characteristics of the data. Typical methods here are

- Mean
- Variance
- Standard Deviation
- Minimum
- Maximum

See Cicchitelli et al (2021)



Mean ($\bar{x}$):

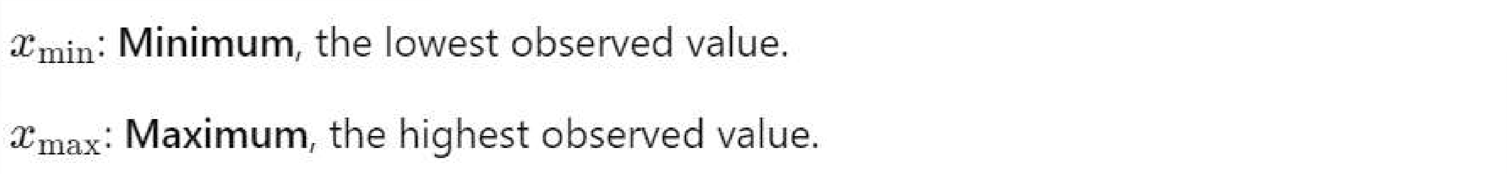
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Variance (s^2):

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

Standard Deviation (s):

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



$x_{\min}$: **Minimum**, the lowest observed value.

$x_{\max}$: **Maximum**, the highest observed value.

UNIVARIATE, BIVARIATE AND MULTIVARIATE DATA ANALYSIS

Exists different analyses can be considered for a dataset, univariate statistics, leading to the analysis of a single variable, then it is very important to study the existing relationships between different variables both considering the bivariate and the multivariate

- Univariate data analysis
- Bivariate data analysis
- Multivariate data analysis

See Cicchitelli et al (2021)

DATA VISUALIZATION

Data visualization is very relevant to improve the understanding of the data structure of the data. It is very relevant to optimize the visuallzations avoiding the biases.

- **Good visualization principles** (Tufte 1983) are key to designing clear and insightful graphics that effectively communicate information. Adhering to best practices in data visualization, such as selecting appropriate scales, choosing the right chart type, and ensuring data accuracy, makes complex data sets easier to interpret and analyze.
- **Adding comparative elements to visualizations** can further engage audiences by enabling them to explore data from various perspectives. A well-crafted visualization transforms raw numbers into a narrative that captures attention, clarifies understanding, and inspires action.

VISUALIZING DISTRIBUTIONS

Visualization can be relevant also to show the distribution of the single variables. In this sense we can use:

- **Boxplots**
- **Histograms**
- **Beanplots**
- **Violinplots**
- **Rug-plots**

All these statistical tools provide unique insights that facilitate a deeper understanding of patterns and variability underlying the data.

VISUALIZING VARIABLE RELATIONSHIPS

Relationships between different variables can be analyzed using:

- **2-D Scatterplots**
- **3-D Scatterplots**
- **Matrices of 2-D Scatterplots**

These visual tools are very useful for analyzing data. They allow analysts to examine patterns and perform further statistical analysis.

See Cicchitelli et al. (2021)

OPTIMIZING DATA VISUALIZATION

Following Tufte (1983) optimizing visualizations can be obtained:

Maximizing the ratio of data to ink: Tufte emphasizes that the proportion of ink used to represent data (data ink) should be as high as possible in relation to the total ink used in a graphic. All unimportant decorative elements should be reduced to a minimum or removed.

Deleting non-data ink: Following on from the previous point, visualizations should avoid "chart junk" In other words, unnecessary or distracting decorations that do not contribute to understanding, such as 3D effects, shadows and superfluous labels, should be avoided.

OPTIMIZING DATA VISUALIZATION

Data density and clarity: Effective graphics can display large amounts of data in a small space without compromising clarity. High data density should not lead to confusion, but rather highlight patterns, trends or relationships that would otherwise be overlooked.

Encourage visual comparisons: Good data visualizations make it easy for viewers to compare quantities. This can be achieved, for example, by using consistent scales.

Avoid distortions: The visualization should not be misleading. Axes should be clearly labeled, proportions respected and transformations indicated. Tufte (1983) insists that graphic integrity is crucial for truthful data communication.

VISUALIZATION FOR THE COMMUNICATION

Storytelling in visualization is about selecting the right visual representation to clearly communicate data and its message.

By aligning the visualization with the data's characteristics and the story it is possible to share information and help the audience connect the understand the information more effectively.

REFERENCES

- Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. Organizational research methods, 16(2), 270-301.
- Cicchitelli, G., D'Urso, P., & Minozzo, M. (2021). Statistics: principles and methods (pp. 1-533). Pearson.
- Giudici, P. (2005). Applied data mining: statistical methods for business and industry. John Wiley & Sons.

REFERENCES

- Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., ... & Munigala, V. (2020, August). Overview and importance of data quality for machine learning tasks. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 3561-3562).
- Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data. John Wiley & Sons.
- Teh, H. Y., Kempa-Liehr, A. W., & Wang, K. I. K. (2020). Sensor data quality: A systematic review. Journal of Big Data, 7(1), 11.
- Tufte, E. R (1983). The visual display of quantitative information (Vol. 2, No. 9). Cheshire, CT: Graphics press.

REFERENCES

- Tukey, J. W. (1977). Exploratory data analysis (Vol. 2, pp. 131-160). Reading, MA: Addison-wesley.