Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2023/2024

Theoretical-Practical Sheet 1

Descriptive Statistics

Objectives:

- Familiarization with the Phyton programming language in the support of Exploratory Data Analysis;
- Brief review of Descriptive Statistics;
- Analysis and discussion of results.

1. PORDATA, the Database of Contemporary Portugal (https://www.pordata.pt/en/home), is a certified statistical database about Portugal, its Municipalities and Europe. In **Pordata>Portugal>Environment, Energy and Territory>Environmental Pollution and Climate**, you will find the number of rain-free days in Portuguese weather stations, in the last decades. Select the Export to Excel button.

a) Using the `read_excel` function, import from the score sheet the numeric data with the corresponding rows and columns labels. Define the `range` parameter to include only this data;

b) Observing this data, check the columns labels. Use the function `colnames` to rename column 1, if necessary;

c) Create a boxplot graphic to compare the number of rain-free days in the weather stations of the database. Which station displayed more and which displayed less rain-free days in the last decades? Analyze the graphic, referring to the data concentration and dispersion;

d) Comment on the existence of zeros, in the data and in the graphics. How can we read these values? Replace all the zeros with **NA** and repeat the previous exercise. What are the more relevant differences?

e) Recreate the graphics only with the mainland weather stations and removing the outliers. Comment on the results;

f) Regarding the quartiles, describe the distribution of the number of rain-free days in Castelo Branco. Repeat with data related to Porto;

g) Create a frequency table for the number of rain-free days in Porto. Repeat the exercise with the data in empirical defined groups and compare when applying Sturges' rule.

h) Create a graphic to observe the distribution of rain-free days in Porto. What can you infer from the graphic? Analyze the table with the frequencies observed in the graphic.

i) Analyze graphically the evolution over time of the rain-free days in Porto. What do you conclude? Repeat the analysis for the number of rain-free days in Faro.

2. Our world in data (https://ourworldindata.org/), is simultaneously a database and the organization that supports it, held in the University of Oxford. According to its own description, their mission is to publish the "research and data to make progress against the world's largest problems". In *Technology Adoption* you can find data about the percentage of population with Internet access. Import to CSV this data set.

a) Create a data structure in R using the `read.csv` function;

b) Change properly the name of the column 4;

c) Compare the distribution of the percentile values of the population with Internet access in the year 2000, with the same distributions in the years 2010 and 2019;

d) Watching the data, check if all the rows represent data related to countries. How relevant is this aspect for the result of the previous point? Filter the data in order to consider only the data related to countries and recreate the graphics of the previous point;

e) Calculate:

i. The number of countries or organizations presented on the data;

ii. Among the previous, which are countries and which are regions or groups of countries?

iii. The minimum, maximum, median and mean values of the percentage of population with Internet access in 2019;

iv. The variance and amplitude interquartile of the same data.

f) Identify 10 countries with the highest percentage of population with Internet access;

g) Graphically represent the evolution over time of the percentage of population with Internet access in Portugal, Spain, Denmark, European Union and in the global population.

Consolidation exercises

1. In PORDATA, on **Pordata>Europe>Education>Educational Attainment**, we can find data about the European population with higher education regarding the percentage of population between 25 and 64 years old, by age group. Export to Excel the available data.

a) Using the function `read_excel`, import from the spreadsheet the numeric data with the correspondent rows and columns labels;

b) Properly change the columns labels;

c) Create a boxplot in order to compare the percentages of population with higher education in the year 2020 in the different age groups. Which is the age group with the higher percentage of population with higher education? Analyze the graphic, referring to the data concentration and dispersion;

d) Analyzing the data, comment on the existence of zeros. How can we read these values? Replace all the zeros with **NA** and repeat the previous exercise. What are the more relevant differences?

e) Regarding the mean, compare the global percentages of population with higher education in the year 1992 and in the year 2020. Considering the number of missing data, can these values be regarded unconditionally?

f) Regarding the quartiles, describe the distribution of the percentage of population with higher education in 2020. Compare with the values observed in 1992;

g) Using the Sturges' rule to clarify the data, create a frequency table with the percentage of European population in the age group 25-34 years old with higher education in 2020. repeat with the age group 55-64 years old and compare the results.

h) Graphically represent the distributions of the previous point. Customizing the `hist` function parameters, represent accordingly both frequency distributions in the same graphic and describe the result regarding the mode, symmetry and data dispersion.

2. In a practical laboratory class of Algorithm and Programming, the teacher decided to carry out a study of the students' performance, in order to evaluate which, type of error was most accomplished. For such purpose, he suggested the students the coding of a given algorithm in C++. Then, he asked these students to compile the program and analyze the number of lexical, syntactic and semantic errors committed.

Student	Lexical Errors	Syntactic Errors	Semantic Errors
1	2	5	1
2	3	2	0
3	0	1	0
4	0	0	0
5	3	2	1
6	2	4	1
7	1	5	0
8	2	6	0
9	1	3	1
10	2	6	0
11	2	4	1
12	3	7	1
13	4	12	1

a) Construct a graph with the ***box plot*** for the three classes of errors, that allows us to analyze student behavior by the number of errors of each type. What is the most common error? Analyze the graph, referring to the data concentration and dispersion.

b) Construct the ***frequency tables*** for each type of error. From the analysis of the table, indicate the median value of each type of error. What is the most common number of errors in each type of error?

c) Determine the mean, the standard deviation, the minimum and the maximum values for each type of error. Based on these measures, what can you say about the data?

d) Construct a suitable graph to visualize the shape of the sample frequency distribution. What can you see on the chart?