Data Analysis - Lab 2

ISEP – October 1st, 2019

Instructions: Prepare a report including the source code and the results. Deposit your report on Moodle and don't forget your binome's name or to make 2 deposits if you did not work alone.

Libraries

This lab requires the following libraries : Numpy, Matplotlib, Seaborn, et Scipy:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns; sns.set()
```

A Multivariate data set : Fisher Iris

In this exercice, we study the Iris data set.

1. Open the file "iris.csv" with a regular text editor to see what the data look like (how many rows, how many attributes, etc). Then, use the pandas command **read_csv(⋯)** with the right parameters so that you can open this data set as a data matrix.

2. Display the histograms of the different attributes. You may use the **displot** function from the seaborn library. What can you say about their distributions ?

3. Compute the coefficient of correlation between all attributes without using the dedicated Python functions. You must define your own functions.

4. Use the commands **pairplot()** and **heatmap()** from the seaborn library to confirm your previous results and visualize the correlation between the different variables. Comment your results.

5. Compute the confidence intervals for the correlation coefficients (we will suppose that the attributes are following a normal distribution). Comment your results.

B Multivariate data set : Anthropometric data

In this exercice, we study the "mansize" data set. These data described anthropometric features acquired in a famous medicine University based on a population of Bachelor students.

1. Open the file "mansize.csv" with a regular text editor to see what the data look like (how many rows, how many attributes, etc). Then, use the pandas command **read_csv(⋯)** with the right parameters so that you can open this data set as a data matrix.

2. Apply the function **describe()** to your data set. What does this function do ? Comment the results on your data.

3. Display the histograms of the different attributes. What can you say about their distributions ?

4. Use the commands **corr()**, **subplots()**, **heatmap()** and **pairplot()** to visualize the correlation between the different variables. Comment your results. In particular, what can you say about the use in archaeology of the femur length to predict the height of an individual ?

5. Compute the confidence intervals for the correlation coefficients (we will suppose that the attributes are following a normal distribution). Comment your results.

6. Based on the results of the previous questions as well as your analysis of the correlation and determination coefficients between the data, conclude on the links between the different variables in this dataset.

C Chi-squared test of independence and categorial variables

In this exercice, we want to assess whether there is a link between different meteorological variables measured in different cities.

1. Open the "weather.csv" data set and describe the different variables and their values using histograms.

2. Use the pandas command **crosstab()** to create the contingency table between the variables "outlook" and "temperature". Comment the repartition of the variables in the resulting table. How many degrees of freedom do we have in this problem ?

3. Use the command **chi2_contingency(·)** from the scipy.stats library on your table. From the result and if need be by computed other indexes, what can you conclude on the dependency between these two variables ?

4. Based on the methodology you used in the previous questions, assess whether there is a link between the other variables of your data set (outlook/humidity, temperature/humidity).