Data Mining
Practical laboratory 1

EFREI – November 25, 2020

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.

Remark: This lab can be done using either R or Python, but help on doing the exercices as well as the corrections will be provided for R language only.

A Discrete series

1. Generate a discrete series of 1000 random data (values included between 0 and 10):
   A = round(runif(1000,0,10))
2. Represent this series in the form of a histogram: to do so, you may use the R functions **barplot** or **hist**. See R help to find the right parameters.
3. Determine the mode, the median and the mean of this series without using the predefined R functions.
4. Verify the mean and the median value of your series using the functions **mean(·)** and **median(·)**. The results should be identical with these of question 3.
5. Explain why the mean and median values of this series may be very different.
6. Determine the range, the variance and the standard deviation:
   - Without using the predefined R functions.
   - Using the predefined R functions: **range(·)**, **var(·)** and **sd(·)**.
   - Comment the results.

B Grouped discrete series

Let B be the following data set:

Mark x_i	5	8	9	10	11	12	13	14	16
Number n_i	10	12	48	23	24	48	9	7	13

1. Input this series and respresent it as a histogram.

   - The R function **c**$(v_1, \cdots, v_N)$ creates a vector with N values. Use this function to generate the vectors for the marks and number of students having each mark.
   - The function **plot(data1,data2,type="h")** is the only one available to generate a histogram from 2 vectors. You can use the command "**? plot**" to learn more about this function.

2. Determine the position and dispersion measures.

3. Explain why this series has a bimodal distribution.

C Normal distributions

The R function **rnorm**(n,m,sd) generates a sample of n random variables that follow a normal distribution of mean m and standard deviation sd. In this exercise, we propose to generate a sample to simulate the human IQ. Human IQ has a mean value of 100 and a variance of 225.

1. Use the function "**curve**$(\cdots)$" to display the probability density function of this distribution ($dnorm(x, \mu, \sigma)$ for a Gaussian distribution).

2. Generate a sample of size 100000 and display its histogram.

3. Assess the mean value and the standard deviation of your sample. Comment.

4. Find the percentage of your sample that has an IQ bellow 60.

5. Find the percentage of your sample that has an IQ above 130.

6. Find the range of values that contains 95 percent of your sample around the mean.

D IQ analysis

In this exercise, we want to assess the affect of malnutrition on the human IQ. Knowing that the average IQ is of 100 with a standard deviation of 15, we will modelise the human population with random sample of different sizes and compare them with IQ sample data from people that suffered from malnutrition.

1 Generate 3 different samples of size 10, 1000 and 100000 with a mean value of 100 and a standard deviation of 15 (function **rnorm**$(\cdot)$).

   - For each sample, evaluate its mean value and its standard deviation.
   - Compare the values you found for the mean and standard deviation with the theoretical values.
   - Calculate the standard error and IC_{95} of the estimated mean values of each sample.
   - Comment on your previous results.

We now want to assess the effect of malnutrition on the IQ. To this end, we will analyze the data from a sample of people that suffered from malnutrition during their childhood.

2 Using the command **read.table(file)**, open the file *malnutrition.csv*.

3 Compute the mean and standard deviation of this new sample.

4 Using the statistical measures at your disposal, what can you conclude on the effect of malnutrition on the IQ when comparing this sample to your previous sample of 100000 elements ?

- Compare the mean and standard deviation of both samples.
- Compute the confidence intervals for both comparisons.
- Comment on your results.