# LFSAB1105 : Probability and statistics
# APP 2019-2020

## Objectives

In the first project, students analyze the performance of a data center, analytically and with simulations. This is an introduction to supply chain management which is a field actively studied by engineers (see e.g. LINMA2470 Stochastic modelling). In the second project, students use the statistical tools of hypothesis testing and linear regression to analyze the performance of supercomputers. They learn how to correctly model real data and assess the quality of a model fit by sound statistical methods.

## Instructions

- This APP consists of 2 parts and counts for **4 points in the final evaluation**.
- This APP has to be done in **groups of 3 to 5 students**. These groups do not have to be the same as the ones during the APEs.
- Don't exchange your work and/or code with other groups. Reports that are too similar will obtain 0 points.
- **Important :** Don't wait until the last week to start this project !

## Report contents

- Your report has to be in **English** and cannot exceed **10 pages** (Appendix excluded). It also has to contain a page with the first and last names and the NOMAs of all group members, and an Appendix containing your Matlab or R code. It needs to end with an Appendix containing 1) the plots asked in the project and 2) your Matlab or R code.
- Grades are granted to the members whose names are on the pdf. If your name doesn't appear on the pdf, you'll get a 0, even though you are in a group on Moodle.
- The clarity and conciseness of your analysis and of the tables and graphs that you present are very important.
- The absence of matlab or R code will lead to a lower grade.

## Report submission

- Every group needs to submit their work as a **PDF** on Moodle **for Sunday 5 January**. Submission after the deadline will not be accepted.
- To submit your report, go to the course on Moodle, then go to the section "APP" and the subsection "Soumission du rapport". You can upload your work in this subsection. Once you are certain that it is your final version, **click the button "Envoyer le devoir"**. It is important that you don't forget to click on this button !
- Reports that have not been uploaded through Moodle will not be corrected.

## Hints

You should take a look at the `MATLAB` document on Moodle.

## Part 1 - Probability

A European data center receives on average 312 computation requests per day from worldwide universities. These requests are treated by three supercomputers. As illustrated in Figure 1, computations are transmitted to one of the three computers with respective probabilities : 0.20, 0.50 and 0.30. The total number of requests is denoted by $N$. Whereas $N_1$, $N_2$ and $N_3$ are the number of computation requests respectively treated by servers one to three.

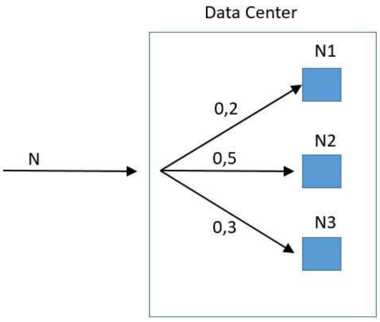

FIGURE 1 – Data center configuration

1. What statistical distribution do you select to model the number of computation requests per day, $N$, received by the data center ? What is its expectation and standard deviation ?
2. Determine the conditional distribution of $N_1|N$ : $P(N_1 = j|N = n)$. Find the conditional expectation $\mathbb{E}(N_1|N = n)$ and standard deviation, $\sqrt{\mathbb{V}(N_1|N = n)}$. Calculate the probability that the first supercomputer performs $N_1 = 64$ computations when the total number of requests is $N = 360$ ?
3. Find the expression of the exact joint distribution of $N_1$ and $N$ : $P(N_1 = j, N = n)$. Plot the pmf (probability mass function).
4. How would you calculate the pmf of $N_1$, $P(N_1 = j)$ with the result of the last sub-question ? As this calculation may be numerically unstable (why ?), you decide to approximate $P(N_1 = j|N = n)$ by a Poisson law. Using this approximation, compute numerically its expectation, $\mathbb{E}(N_1)$, and standard deviation, $\sqrt{\mathbb{V}(N_1)}$. Plot the pmf of $N_1$.
5. You have to report **weekly** statistics (one week = 5 days) to the managing authorities. They are interested in the weekly expectation, standard deviation and the 5% and 95% quantiles of the number of requests received by the data center. You consider two approaches :

   5.1 The first one consists in performing simulations of the number of daily requests. Choose an appropriate number of simulations. Plot the empirical distribution (pdf) and report the empirical mean, standard deviation and 5% and 95% quantiles.

   5.2 In the second, you use a Gaussian approximation for the number of requests. Can you justify this choice ? Compare the expectation, standard deviation and 5% and 95% quantiles with statistics obtained in question 5.1.

## Part 2 - Statistics

*This exercise is made of two parts.*

- *In question 1-4 you should use the dataset* ***CPU1.txt***
- *In question 5-7 you should use the dataset* ***CPU2.txt***

*More information on these two datasets is provided in section Datasets. Note that the two parts are independent. hence you should not use results you derived in the first part to answer questions in the second part.*

You are a researcher at UCLouvain and need a supercomputer to run a set of simulations. In order to perform your job (i.e. your set of simulations), you book a supercomputer and divide your job between a given number $m$ of CPU's for a certain time $t$. The time your job takes to run is represented by a random variable $T$ and is the maximal running time across the CPU's. Note that the $m$ CPU's are operating independently.

There are two types of CPU's (low-performing and high-performing). More specifically

- $X_1, \ldots, X_l \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$ represent the time for the $l$ low-performing CPU's ;
- $Y_1, \ldots, Y_h \overset{\text{i.i.d}}{\sim} N(\mu + \delta, \sigma^2)$ represent the time for the $h$ high-performing CPU's ;

where $\sigma^2 = 100$ is known and $\delta < 0$. The total number of CPU's you are using is thus $m = l + h$.

1. What is the MLE of $\mu$ and $\delta$ ?
2. What is the MLE of $P(T \leq t)$ ? Your answer should be expressed in terms of $\Phi(.)$, the CDF of the standard normal distribution.
3. Some of your colleagues argue that all CPU's actually have the same performance. Conduct a test to conclude whether this view or the assumption of low and high performance CPU's is the most relevant. You should
   - State what test your are using and write down the assumptions underlying its use.
   - Carefully write down the hypotheses your are testing.
   - Undertake the test (with a significance level of 0.05) and conclude.
4. Would your conclusion necessarily be the same with an ANOVA test at the same significance level ? Discuss.

Now we want to analyze the performance at the CPU-level, focusing on the low-performing ones. In each simulation, you perform operations on a matrix. As such, you would like to see how the size of the matrix influences the operating time of the CPU. Two models are postulated

$$\begin{aligned} X_k &= \alpha_0 + \alpha_1 S_k + \epsilon_k, \\ X_k &= \beta_0 + \beta_1 S_k^2 + \varepsilon_k. \end{aligned}$$

with $k = 1, \ldots, 75$, $\epsilon_k \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ and $\varepsilon_k \overset{\text{i.i.d}}{\sim} N(0, \tau^2)$. Moreover, $S_k$ is the matrix size (number of columns multiplied by number of rows) for simulation $k$.

5. Estimate these two models and, given the data you observe, conclude which one is the most relevant (provide at least two different arguments).
6. Consider a matrix with 100 rows and 10 columns. Using the model you have chosen, predict the operating time and provide a prediction interval of level 95%.
7. Undertake a simulation exercise in order to judge the coverage of your prediction interval. More specifically, you should

- Fix the matrix size $S_k$ to that given in 6. and simulate 1000-times the error term from a $N(0, 100)$.
- Compute the associated operating time using the model you have chosen.
- Examine the proportion of operating times falling in your prediction interval.
- Interpret the obtained results.

*You should carefully explain how you proceeded in practice, for example by commenting **small** excerpts of code.*

### Datasets

Two datasets are used throughout this exercise

1. `CPU1.txt` containing two columns
   — `OperatingTime` gives the operating time on each CPU.
   — `LowCPU` indicates whether the CPU comes from a low-performance CPU (value 1) or a high-performance CPU (value 0).
2. `CPU2.txt` containing two columns
   — `OperatingTime` gives the operating time on each low-performance CPU (i.e. $X_k$)
   — `MatrixSize` gives the size of the matrix that each CPU has to use (i.e. $S_k$)