Data Valorization
Solution Labs 1: Exercise 1.2

Cyprien GILET and Lionel Fillatre

Exercice 1.2

Recall 1. ***(Central Limit Theorem)***. *Let $Y_1, \dots, Y_n$ be n real independent and identically distributed (i.i.d.) random variables such that $\mathbb{E}[Y_i] = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$ for all $i \in \{1, \dots, n\}$. Then,*

$$\bar{Y}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \xrightarrow[n\to\infty]{\mathcal{L}} \quad Z \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Let $\mu = 5$ and $\sigma = 2$. The data matrix $X \in \mathbb{R}^{n\times k}$ is generated as follow:

$$X := \left.\overbrace{\begin{bmatrix} & \vdots & \\ \cdots & X_{ij} & \cdots \\ & \vdots & \end{bmatrix}}^{k}\right\} n$$

where, for all $i \in \{1, \dots, n\}$ and all $j \in \{1, \dots, k\}$, $X_{ij} \sim \mathcal{N}\left(\mu, \sigma^2\right)$. To generate X_{ij} according to a normal distribution, we can use the R command `rnorm`.

Then, the one-row matrix $\bar{X} \in \mathbb{R}^k$ is given by:

$$\bar{X} := \left.\overbrace{\begin{bmatrix} \cdots & \bar{X}_j & \cdots \end{bmatrix}}^{k}\right\} 1$$

where, for all $j \in \{1, \dots, k\}$,

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}.$$

From the central limit theorem, for n sufficiently large, we get

$$\forall j \in \{1, \dots, k\}, \quad \bar{X}_j \underset{n\to+\infty}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

We can note that, in practice, when n increases, the variance $\frac{\sigma^2}{n}$ of $\bar{X}_j$ decreases until 0.
In this exercise, since all the variables X_{ij} are normal, the variable $\bar{X}_j$ follows exactly a normal distribution. However, the proposed approach holds for any distribution satisfying the central limit theorem. For example, you can replace the R command `rnorm` by `rexp` which generates some random samples from the exponential distribution. In the case of an exponential distribution of parameter λ, the mean of $\bar{X}_j$ will be $\frac{1}{\lambda}$ and its variance will be $\frac{1}{n\lambda^2}$.