TD n°3

Exercise 1. *Normal law*

Let $m, \sigma^2 \in \mathbb{R}_+$, a random variable X follows a *normal law* $\mathcal{N}(m, \sigma^2)$ if X takes its values in $\mathbb{R}$ and has density $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$.

1. Compute the mean and variance of X.

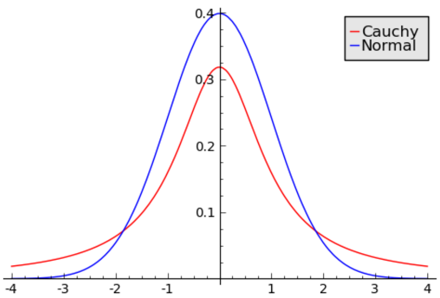
2. Let X, Y be independent normal random variables of laws $\mathcal{N}(m_1, \sigma_1^2)$ and $\mathcal{N}(m_2, \sigma_2^2)$, show that $X+Y$ follows the law $\mathcal{N}(m_1+m_2, \sigma_1^2+\sigma_2^2)$.

Exercise 2. *Normal law sampling*

1. Assuming that you can call a `Random` function which returns a random variable uniform over [0,1] independent of previous calls, can you apply the inverse transformation method to easily simulate the normal law $\mathcal{N}(0,1)$?

2. Using the inverse transformation method, simulate the law of density $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ over $\mathbb{R}$ (Cauchy law).

3. Deduce a way to simulate the law $\mathcal{N}(0,1)$ using von Neumann rejection method.



Exercise 3. *Gaussian errors*

You have done n independent experiments to measure an unknown parameter $\mu \in \mathbb{R}$. The measures may be subject to errors, each measure is then modeled by a random variable $X_i = \mu + \varepsilon_i$, $1 \le i \le n$, where ε_i are i.i.d. random variables with a law of mean 0 and variance 1. Given a sample $(x_1, \dots, x_n)$, $x_i \in \mathbb{R}$, you choose to use the empirical mean $\overline{x}_n = (x_1 + \cdots + x_n)/n$ to estimate μ, and you denote the associated random variable $\overline{X}_n = (X_1 + \cdots + X_n)/n$.

1. Give a quick justification to the fact that if you make a sufficient number of experiments, you can approach the exact value μ as close as you want.

You wish to have the guarantee with high probability ($\ge 99\%$) that the estimation error is small ($\le 0,1$), that is $\mathbb{P}(|\overline{X}_n - \mu| \ge 0,1) \le 0,01$.

2. Use Chebychev's inequality in order to set a number of experiments providing this guarantee.

3. Same question with the additional assumption that the random variables ε_i are normal: can you plan fewer experiments ? How many ?

Exercise 4. *Cantelli inequality*

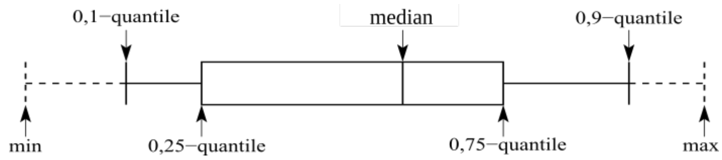
Let X be a real random variable of mean $\mathbb{E}(X) = \mu$ and variance $Var(X) = \sigma^2$, both finite.
Prove the Cantelli inequality (a unilateral Bienaymé-Tchebychev bound) : for any real $a > 0$,

$$\mathbb{P}(X - \mu \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

Exercise 5. *Mean vs Median*

Let X be a real random variable of mean $\mathbb{E}(X) = \mu$ and variance $Var(X) = \sigma^2$, both finite. Let $m \in \mathbb{R}$ be the median value of X, defined by $\mathbb{P}(X < m) \leq 1/2$ and $\mathbb{P}(X > m) \leq 1/2$.

1. Given the box plot of X drawn below, where can the mean locate ?



2. A statistician claims that the distance between mean and median is controlled by the standard deviation, that is $|m - \mu| \leq \sigma$ for any distribution where μ and σ are finite. Using Python, generate several samples from several laws (e.g. try 3 different laws and generate 10 i.i.d. samples of size 1000 per law). Then check the claim for the empirical law of each sample. Now what is your opinion about the claim ?

3. Depending on your intuition from the preceding question, either find a counter-example to $|m-\mu| \leq \sigma$ (as simple as possible), or prove it (mathematically).

Exercise 6. *Yule-Simpson school*

A selective school hires students through different exams: cooking, ping-pong, belote. Each candidate can take only one of those exams. The total number of male candidates over the three exams is equal to the total number of female candidates. For each exam, you observe that the success rate of boys is better than the success rate of the girls. Can you infer that the success rate of boys over all the three exams is better than the one of girls ? If YES, prove it. If NO, provide a counterexample.

Exercise 7. *Confidence in confidence intervals*

1. Recall how to construct a confidence interval of exact level 90% for the mean μ of a normal law $\mathcal{N}(\mu, 1)$ (standard deviation is known) using a sample of $(X_1, \ldots, X_n)$ with X_i i.i.d. $\sim \mathcal{N}(\mu, 1)$.

2. Choose some mean $\mu \in \mathbb{R}$ and some integer $n \in \mathbb{N}^*$, then generate one sample of $(X_1, \ldots, X_n)$ with X_i i.i.d. $\sim \mathcal{N}(\mu, 1)$. For how many people in the classroom does the confidence interval for the sample fails to cover the mean ? Is it a surprise ?