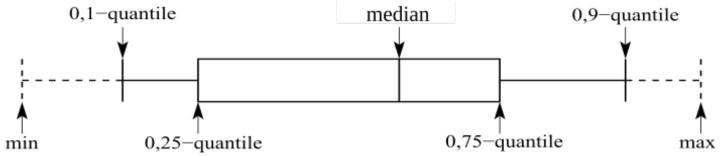
TD n°4 - More on statistics

Exercise 1. *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 2. *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 3. *Confidence in confidence intervals*

1. To estimate a parameter $\mu \in \mathbb{R}$ of your system, you perform n independent measures $x_1, \ldots, x_n$. Some limited noise disturb the measures and you choose to modelize this noise by a normal law of variance = 1. With this model, you suppose that the sample $x_1, \ldots, x_n$ have been generated by i.i.d. random variables $X_1, \ldots, X_n$ of normal law $\mathcal{N}(\mu, 1)$. Explain how to build for this model a *confidence interval for μ of level* $0 < \alpha < 1$ *from a sample of size n*, that is provide two functions $I^- : \mathbb{R}^n \to \mathbb{R}$ and $I^+ : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{P}(\mu \in [I^-(X_1, \ldots, X_n), I^+(X_1, \ldots, X_n)]) \geq \alpha$.

2. Let $\alpha = 90\%$ the confidence level. Test your confidence interval by choosing some mean μ and some integer $n \in \mathbb{N}^*$, then generate 100 samples of $(X_1, \ldots, X_n)$ with X_i i.i.d. $\sim \mathcal{N}(\mu, 1)$. For how many samples, does the confidence interval from the sample fail to cover the mean μ ? Is it a surprise ?

Exercise 4. *PID Enquiry*

For one day, a computer has dedicated his work to spam processes on your network. To analyze what happened, you have access to incomplete logs mentioning some of the process identifiers (PID) from this computer. From what you know, you assume that if n processes were launched, their PID started at 1 and then were incremented by 1 at each new process, yielding all PID from 1 to n. In your logs, you observe only k different PID which seem uncorrelated and you assume they have been picked uniformly at random in $\{1, \ldots, n\}$. Here is the list of PID you collected :

1195, 1250, 479, 1550, 1563, 1866, 1334, 154, 1693, 1720, 1855, 16, 737, 1090, 903, 349, 729, 1329, 1403, 1819

How many processes did the computer spammed during that day ? Explain your guess about n and give some arguments to convince us that we can trust you (experimental and/or theoretical arguments).