Project

You **must** return a single Jupyter notebook in Python on ecampus. No other format will be accepted. Mathematical formulas must be written in Markdown / LateX. Python code must run without errors.

The deadline is October 30, 2021.

We focus on the cost of nuclear accidents before the accident of Three Mile Island that occurred on March 28, 1979. The dataset is available on ecampus.

Exercise 1 (Statistical model)**:**

1. Load data as a vector $x = (x_1, \dots, x_n)$ of nuclear accident costs before the Three Mile Island accident and remove all missing values. You must get $n = 55$ observations.
2. Let F_θ be the cumulative distribution function of a Gaussian distribution with mean μ and variance σ^2, with $\theta = (\mu, \sigma^2)$.
   (a) Show that the quantile function F_θ^{-1} satisfies:
   $$\forall\, p \in (0,1), \quad F_\theta^{-1}(p) = \sigma F_{(0,1)}^{-1}(p) + \mu.$$
   This suggests that, if observations have a normal distribution, the corresponding Q–Q plot is well approximated by a line.
   (b) Show the Q–Q plot of data for the Gaussian model using the probplot function of SciPy.
3. Let F_θ be the cumulative distribution function of an exponential distribution with parameter $\theta > 0$.
   (a) Show that the quantile function F_θ^{-1} satisfies:
   $$\forall\, p \in (0,1), \quad F_\theta^{-1}(p) = \frac{1}{\theta} F_1^{-1}(p).$$
   Again, if observations have an exponential distribution, the Q–Q plot is well approximated by a line.
   (b) Show the Q–Q plot of data for the exponential model.
4. Discuss the results.

In the following, we use the **exponential model** for the cost of accidents before the Three Mile Island accident. Specifically, we assume that accident costs are i.i.d. samples of an exponential distribution with parameter $\theta > 0$. Let $\Theta = (0, +\infty)$ be the set of parameters. We denote by $X = (X_1, \dots, X_n)$ a random vector of n i.i.d. samples.

Exercise 2 (Point estimation)**:**
We first focus on the estimation of θ or some functions of θ.

1. Give the maximum likelihood estimator $\widehat{\theta}$ of θ.

2. Show on the same plot the histogram of data in density and the probability density function of the exponential distribution with parameter $\widehat{\theta}(x)$.

3. We seek to estimate the expected cost $g(\theta) = \frac{1}{\theta}$. Let $\widehat{g}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$. Show that $\widehat{g}$ is an efficient estimator of $g(\theta)$.

4. Compute $\widehat{g}(x)$ from the available observations.

5. For any $\eta > 0$, define the estimator
$$\widehat{g}_\eta = \eta\widehat{g}.$$
Show that for some values of η (to be specified), the quadratic risks associated with the estimation of $g(\theta)$ satisfy $R(\theta, \widehat{g}_\eta) < R(\theta, \widehat{g})$ for all $\theta > 0$. Discuss this result.

6. Give η so that $\widehat{g}_\eta$ is an unbiased estimator of the median cost. Compute $\widehat{g}_\eta(x)$ from the available observations. Compare with the empirical median.

7. For the previous value of η, compare the quadratic risks of $\widehat{g}$ and $\widehat{g}_\eta$, both viewed as estimators of $g(\theta)$, depending on the number of samples n.

Exercise 3 (Hypothesis testing)**:**
We wish to show that the expected cost of an accident is less that one billion dollars. This amounts to reject the null hypothesis H_0 that the expected cost of an accident is at least one billion dollars.

1. Give the null hypothesis H_0 and the alternative hypothesis H_1 as subsets of Θ.

2. Using Neyman-Pearson's approach, give a uniformly most powerful test at level α.

3. Apply the test at level $\alpha = 5\%$. Give the p-value and conclude.

4. For $n = 55$ samples and the parameter θ associated with an expected cost of one billion dollars, plot the probability density function of $\widehat{g}(X)$ and show the rejection region of H_0 at level $\alpha = 5\%$.

5. Plot the power of the test at level α with respect to θ for $n = 10, 50, 100, 1000$ samples. Explain the results.

6. Propose a new test using the approximation of $\widehat{g}(X)$ suggested by the Central Limit Theorem and give the result of this test.