QUALITY DATA ANALYSIS

22/01/2025

General recommendations:

1. Write the solutions in CLEAR and READABLE way on paper and show (qualitatively) all the relevant plots;
2. avoid (if not required) theoretical introductions or explanations covered during the course;
3. always state the assumptions and report all relevant steps/discussion/formulas/expression to present and motivate your solution;
4. when using hypothesis tests provide the numerical value of the test statistic and the test conclusion in terms of p-value.
5. Exam duration: 2h
6. **For multichance students only: Exam duration is 2h 30min**

Exercise 1 (15 points)

A manufacturing company operating in the food industry is interested in monitoring the energy consumption of one of its production lines. To this aim, they started recording the peak power during the most energy-demanding process, on a weekly basis. The data from the first 39 weeks is reported in \`energy.csv\`.

1) Inspect the dataset and verify the normality and independence assumptions for the available measurements.
2) Suggest a suitable model for the data.
3) Based on the model defined in point 2), design an appropriate control chart to monitor the peak power over time. Use an ARL_0 = 200. In case of violations of control limits, assume that there is no assignable cause.
4) Based on the model defined in point 2), estimate the 95% prediction interval for the peak power in the next week.

Exercise 2 (14 points)

The density of a new additively manufactured magnesium alloy is measured by an aerospace company. They collected density measurements of specimens manufactured in two different positions of the machine, namely in the front and in the back of the build volume.

The data are stored in the 'magnesium.csv' file. The mean and standard deviation of sample densities are known from historical measurements (front: μ = 1600 kg/m^3, σ = 860 kg/m^3; back: μ = 1400 kg/m^3, σ = 795).

1) Inspect the dataset and check the normality and independence assumptions. Perform also the LBQ test at lag 3 and report the p-values.
2) Using the information about known means and standard deviations, design two univariate control charts for the mean density with a family-wise ARL_0 = 250. Discuss the results.
3) The process engineers believe that the data exhibit a lower standard deviation than the one observed in historical measurements. Design appropriate statistical tests to confirm or reject their hypothesis.
4) Re-design the new univariate control charts for the mean density replacing known parameters with ones estimated on the available data (use a family-wise ARL_0 = 250). Discuss the results.

Exercise 3 (4 points)

In the following questions select one of the four possible choices as your answer and provide a short justification of your choice. Answers **without** justification will **not** receive any credit.

Question 1

We have a sample of n data points on the variables (Y, X_1, X_2) and we fit the linear regression model M$_1$: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where the coefficient of X_2 (i.e. β_2) is statistically significant. From model M$_1$ we remove the variable X_1 and we fit on the same n data points the partial model M$_2$: $Y = \beta_0^* + \beta_2^* X_2 + \varepsilon$. Which of the following statements will be valid?

- **a)** β_2^* will not be statistically significant in M$_2$
- **b)** β_2^* will still be statistically significant in M$_2$
- **c)** β_2^* will not only be statistically significant in M2, but it will have higher statistical significance (i.e. smaller p-value) from what β_2 had in M$_1$
- **d)** We cannot infer anything from the above, based on the given information.

Question 2

We observe a time series model for which is known to be a stationary AR(3) model. Which of the following descriptions of the pair of plots of ACF (Autocorrelation Function) versus lag and PACF (Partial Autocorrelation Function) versus lag is most likely?

- **a)** The ACF shows exponential decay and the PACF will have significant the first three lags only.
- **b)** The ACF show linear decay and the PACF will have significant the first three lags only.
- **c)** Both the ACF and PACF will have significant the first three lags, then they will decay exponentially fast.
- **d)** The ACF will have significant the first three lags then it will decay linearly and the PACF will show exponential decay, after the first three lags.