QUALITY DATA ANALYSIS

17/06/2024

General recommendations:

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

Exercise 1 (14 points)

The quality of a new pharmaceutical product is monitored by means of 3 quality indicators: "x0", "x1" and "x2". The data collected over 1 week of production is reported in "PCA_phase1.csv". Data are reported in the original data acquisition order.

1) *(All students)* The head of the quality assurance department would like to apply the Principal Component Analysis (PCA) to these data. Is it more appropriate to use the variance covariance matrix or the correlation matrix of the data? Discuss and motivate the choice.
2) *(All students)* Based on what you decided in point 1) apply PCA to the data and determine the number of principal components that should be retained to capture at least 80% of the total variance (report the eigenvectors and the eigenvalues of the retained components).
3) *(All students)* Check if the scores of the retained PCs are normal and independently distributed. In case of violation, suggest and apply a suitable approach to the original data so that the newly derived PC scores obey the assumptions (we need to keep capturing at least 80% of the total variance).
4) *(10 CFUs / 8 CFUs / 5 CFUs math.eng.)* Based on the results of previous points, design appropriate control charts on the PCA scores to monitor the process (with familywise Type I = 1%). Report the values of the control limits. *Note: in case of violations of control limits, assume no assignable cause was found.*
5) *(10 CFUs)* Assume that from historical data the following parameters are known:

$$\boldsymbol{\mu} = [10.2\ 30.1\ 1505]^T,$$
$$\sigma_1 = 3.5, \sigma_2 = 4.7, \text{ and } \sigma_3 = 5.5$$

It is also known that the 1st Principal Component is normally and independently distributed with:

$$\boldsymbol{u}_1 = [0.68, -0.17, 0.71]^T, \lambda_1 = 1.65$$

In case we are interested in monitoring the 1st Principal Component only, how does the design of the I control chart changes with respect to the I chart designed in point 4)? Is the process in control?

Exercise 2 (15 points)

You are the new process engineer of a company that produces heavy-duty vehicles. Your first task is to design a control chart to monitor the power consumption (in kW) of the machines installed at a metal forming station. The station works on a 24/7 schedule. The total power consumption of the station is measured once every hour. You are given a dataset (`power_phase1.csv`) containing the power consumption data collected from day 1 to day 4.

1) *(All students)* Inspect the dataset and verify the assumptions.
2) *(All students)* Knowing that from hour 12 (12:00 PM) to hour 18 (6:00 PM) the station operates at full capacity (with 2 machines working in parallel), suggest an adequate model to fit the data.
3) *(10 CFUs / 8 CFUs / 5 CFUs math.eng.)* Design an appropriate control chart to monitor the power consumption of the station such that the average number of days before a false alarm is 10. Report the values of the control limits. *Note: in case of violations of control limits, assume no assignable cause was found.*
4) *(10 CFUs / 8 CFUs / 5 CFUs math.eng.)* The data collected from day 5 to day 7 is stored in `power_phase2.csv`. Using the control chart designed at point 3) check if the new observations are in control and report the number of out-of-control points detected (if any).
5) *(10 CFUs)* Knowing that small increases in power consumption can anticipate a failure in the machines, suggest a different control charting approach that could be more effective than the one designed at point 3. Motivate your answer.

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 (2 points):
In a hypothesis testing problem, which of the following statements is valid?

- **a)** If the level of significance alpha (α) will decrease, the false alarm will increase.
- **b)** As the probability of type I error increases the power of the test will increase.
- **c)** The p-value of a test statistic is the probability of type II error.
- **d)** All the above are invalid

Question 2 (2 points):
We have available n data points from three random variables (Y, X_1, X_2) for which we have estimated the following correlations: $corr(Y, X_1) = -0.57$ and $corr(Y, X_2) = +0.39$. We run:

I. the simple linear regression of Y on X_1 and derive the SSE_1,
II. the simple linear regression of Y on X_2 and derive the SSE_2 and
III. the multiple linear regression of Y on $\{X_1, X_2\}$ and derive the SSE.

Which of the following is valid?

- **a)** $SSE_1 < SSE_2$
- **b)** $SSE_1 > SSE_2$
- **c)** We cannot tell if $SSE_1 < SSE_2$ or $SSE_1 > SSE_2$ from the given information
- **d)** $SSE = SSE_1 + SSE_2$