QUALITY DATA ANALYSIS

19/06/2023

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: skip exercise 2 point 4) and exercise 3 question 1).**

Exercise 1 (14 points)

A manufacturing company that produces electronic components is interested in monitoring the quality of the produced components using statistical techniques. The company has collected data on several variables related to the components, including dimensions, electrical characteristics, and performance indicators. The data are stored in "exe2_phase1.csv".

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). Discuss and motivate the choice of using either the variance covariance matrix or the correlation matrix of the data.
2) Based on the results of point 1), design a suitable control charting approach for these data, such that the familywise Type I error is at most 1%. Motivate the choice of the proposed control chart and discuss the result. *Note: in case of violations of the control limits, assume that no assignable cause has been found.*
3) You are now in phase 2 (usage stage of the CC). Test the control chart designed in point 2) on the new observations stored in the file "exe2_phase2.csv" and determine if the process is in-control or not. Report the index of the out-of-control observations (if any).

Exercise 2 (15 points)

In a plant for the production of hydraulic actuators, the head of the quality control department is aiming to implement a new statistical process monitoring tool. The quality characteristic of interest is the holding force measured in kN during acceptance testing. The quality department received holding force measurements that consist of five measurements for each tested actuator. The data for control chart design are stored in "exe1_phase1.csv".

1) Assume that actuators were manufactured and tested with the same sequential order displayed in the dataset, but no information about the time order of individual measurements for each actuator is available. Design an $\bar{X} - R$ control chart such that the average number of samples before a false alarm is 400 and discuss the result.
2) Based on the result of point 1), design a more appropriate control chart for the available data (with the same average number of samples before a false alarm used in point 1). Discuss the results.
3) After some investigations, the data analysts found that the five individual measurements reported for each actuator were carried out always with the same time order, and the order is the one displayed in the dataset. Based on this new information, identify and fit a model for these data and use it to design an appropriate

control chart (same average number of samples before a false alarm used in point 1) and 2)). *Note: in case of violations of control limits, assume no assignable cause was found.*

4) Using the control chart designed in point 3) determine if the new samples stored in "exe1_phase2.csv" are in-control or not.

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 data set, we fit the simple linear regression model of Y on X, and the Ordinary Least Squares (OLS) line is given by: $\hat{Y} = -0.138 - 1.33 * X$. In the ANOVA table the overall F statistic has the value: $F = 16.81$. Which of the following will be the T-test statistic value (T) that examines the hypothesis testing: $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$? (β_1 refers to the slope of the regression line).

a) $T = -4.1$
b) $T = +4.1$
c) $T = 16.81$
d) We cannot tell from the above output only.

Question 2 (2 points):
In a data set, we fit the simple linear regression model of Y on X, and the Ordinary Least Squares (OLS) line is given by: $\hat{Y} = 1.67 - 2.84 * X$. If the coefficient of determination was $R^2 = 0.81$, then which of the following is true for the (Pearson) correlation coefficient between X and Y, i.e. $r(X, Y)$?

a) $\mathrm{r(X, Y)} = +0.9$
b) $\mathrm{r(X, Y)} = -0.9$
c) $\mathrm{r(X, Y)} = 0$
d) None of the above