10 CFUs
QUALITY DATA ANALYSIS

04/09/2024

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 (14 points)

Automotive-Parts Co. has launched a new model of cast aluminum piston. The quality department is interested in monitoring the hardness (HB) of the pistons. The measurements of the first week of production are stored in `HB_phase1.csv`. Each row represents a sample with a sample size equal to 4.

1. Inspect the dataset and verify the assumptions.
2. Assuming randomness inside each sample, design an Xbar-R control chart for the data with $ARL_0 = 500$. Discuss the results.
3. Knowing that the data stored in each column correspond to hardness measured on pistons produced with one specific mold, check if the hardness in pistons produced with mold 2, 3 and 4 ('x2', 'x3', 'x4') is significantly different from the hardness measured in pistons produced with mold 1 ('x1').
4. Based on the results of the point 3), suggest a suitable model for the data.
5. Based on the model derived in point 4), design an appropriate control chart using the same ARL_0 used in point 2. Discuss the results. *Note: in case of violations of control limits, assume that no assignable cause was found.*

Exercise 2 (15 points)

The quality department of a company that produces dried fruit is interested in analyzing and monitoring three quality characteristics of their top-selling product, namely moisture content (x0), nutritional content (x1), and weight (x2).
A dataset consisting of 50 random samplings has been collected and stored in the file `PCA_phase1.csv`.

1. Using the data correlation matrix, how many Principal Components shall be retained to explain at least 80% of the overall data variability? Report the following information:
   a) eigenvalues and cumulative explained variance ratio
   b) loadings of the retained PCs
2. Using the `PCA_phase1.csv` dataset, design univariate control charts with a familywise Type I error $\alpha = 0.0027$ to monitor the scores of the PCs retained in point 1). *Note: in case of violations of control limits, assume that no assignable cause was found.*

3. A new dataset has been acquired. The data are stored in `PCA_phase2.csv`. Compute the scores by projecting the new observations onto the space spanned by the retained Principal Components **identified in point 1)**. Report the following information:
   a) sample means and variances of the computed scores
   b) sample correlation between the computed scores
   c) p-values of runs-tests performed on the computed scores
   d) p-values of the normality tests performed on the computed scores
   Discuss the results
4. Apply the control charts designed in point 2) to the scores computed in point 3). Is the process in control or not? Discuss the result.
5. The data scientists of the quality department are interested in using the PCA for data reconstruction. They are specifically interested in reconstructing the first variable, moisture content, using either only the first PC or the first 2 PCs. Using the data stored in `PCA_phase1.csv`, compute both these two reconstructions. Report the sample means and sample standard deviations of the variable reconstrued using either 1 PC or 2 PCs. Discuss the result.

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 regarding the p-value is valid?

a) If the p-value exceeds the level of significance, then we have evidence against the null hypothesis.
b) The p-value is the probability that the null hypothesis is true
c) The p-value is equal to the power of the test
d) The p-value depends on the observed test-statistic.

Question 2 (2 points)

In a control chart for the monitoring of the mean of a Normally distributed process, the lower and upper control limits (i.e., LCL and UCL) are designed so that under the in-control state we have ARL_0=100. For this chart, we have estimated the type II error β^* and the respective ARL_1^* in detecting a shift of size $\delta^* = -1.5$ standard deviations (i.e. a downward shift of size $\delta^* = -1.5\sigma$). If in practice, we will experience a shift of size $\delta = +2.5$ standard deviations (i.e. an upward shift of size $\delta = +2.5\sigma$), then which of the following statements regarding ARL_0, the type II error β and the respective ARL_1 of the $\delta = +2.5\sigma$ shift will be valid?

a) ARL_0 in the new shift will become smaller.
b) $ARL_1 < ARL_1^*$
c) $ARL_1 > ARL_1^*$
d) We cannot estimate the ARL_1 shift from the given information

→ 17.49/30 → 1 di...

- senza contare i CC
- dandomi 0 nel test d'Ip per X_1 e X_2 dell'esercizio 1

• senza contare punti per la lode