QUALITY DATA ANALYSIS

01/09/2023

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: you can skip Exercise 1, point 2), Exercise 3, question 2).**

Exercise 1 (14 points)

A process engineer is interested in monitoring the temperature and pressure of a chemical reactor. The values are measured at 15-minute intervals in 5 different locations. The data collected during the first 4 hours of operation are stored in `reactor_temp_phase1.csv` and `reactor_press_phase1.csv`:

Only MECH ENG STUDENTS:

1) Design two UNIVARIATE control charts to monitor both the temperature and pressure mean in the reactor with an overall ARL0 of 500. *Note: in case of violations of control limits, assume no assignable cause was found.*

Only OTHER STUDENTS:

1) Design one MULTIVARIATE control chart to monitor the temperature and pressure mean in the reactor with an overall ARL0 of 500. *Note: in case of violations of control limits, assume no assignable cause was found.*

ALL:

2) The process engineer suspects that the temperature measured at location 2 is greater than the temperature at location 4. Check if this is the case with an appropriate test (alpha = 0.05).
3) Using the control chart(s) designed in point 1 (phase 1), check if the data collected during the next 1 hour of operation (stored in `reactor_temp_phase2.csv` and `reactor_press_phase2.csv`) are in control.

Exercise 2 (15 points)

A service company is interested in implementing a novel statistical process monitoring approach to keep under control how some of their most important performance indicators evolve over time. The head of the quality department decided to start a pilot project focusing on one single indicator, that is measured on a daily basis. The values recorded in 120 consecutive days are reported in "exeKPI.csv".

1) Find a suitable model to fit the performance indicator data. Note: to verify the lack of autocorrelation, use an LBQ test with L (number of lags) = 10 and show both the value of the test statistics and the p-value.
2) Based on the result of point 1) design a suitable control chart methodology for the performance indicator, and using the designed approach determine if the underlying process is in-control or not (use K = 3). *Note: in case of violations of control limits, assume no assignable cause was found.*
3) The head of the department is interested in evaluating a batching approach on the original performance indicator time series, with batch size = 4. After applying the batching operation, design a suitable control chart (K = 3) *Note: to verify the lack of autocorrelation, use an LBQ test with L (number of lags) = 5 and show both the value of the test statistics and the p-value; in case of violations of control limits, assume no*

assignable cause was found. Discuss the result as well as the difference with respect to the result in point 2).

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

When we apply PCA, the goal is to approximate the available data space with one of smaller dimension. In deciding of how many components we need to keep in this approximation, which of the following is **not** useful:

a) Percentage of variation explained by each component.
b) Cumulative percentage of variation explained at each of the ordered components.
c) To know whether PCA was applied on the standardized variables or on the original.
d) Scree plot

Question 2

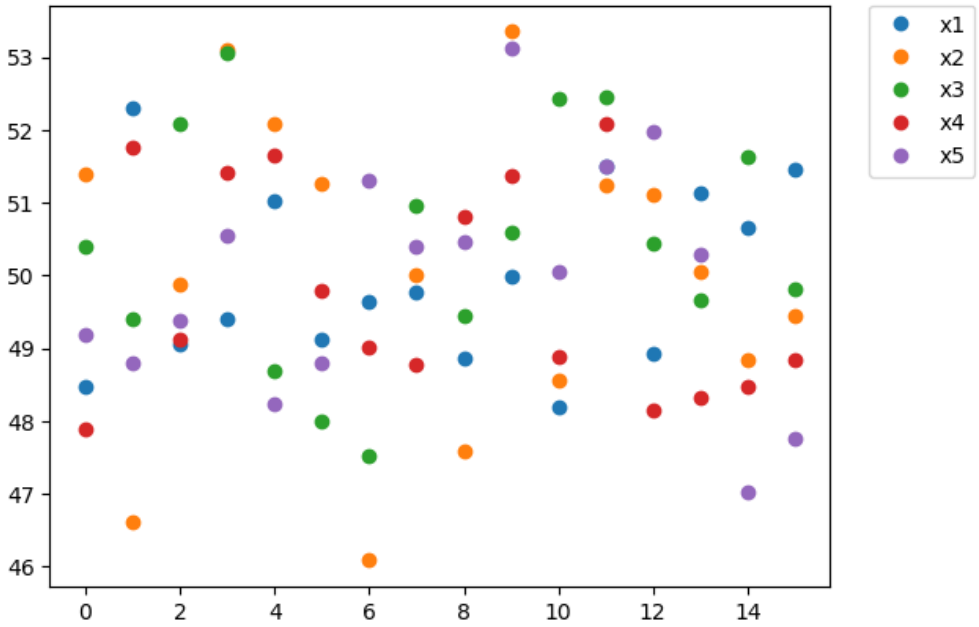
In a data set with five variables (X_1, X_2, …, X_5), where correlation $|r(X_i, X_j)| < 1$, for $i \neq j$, we applied PCA. Which of the following can represent the ordered proportion of variance explained by the first four components?

a)	0.48	0.31	0.12	0.09
b)	0.52	0.18	0.13	0.10
c)	0.42	0.29	0.22	0.08
d)	0.68	0.19	0.14	0.04

Exercise 1 solution

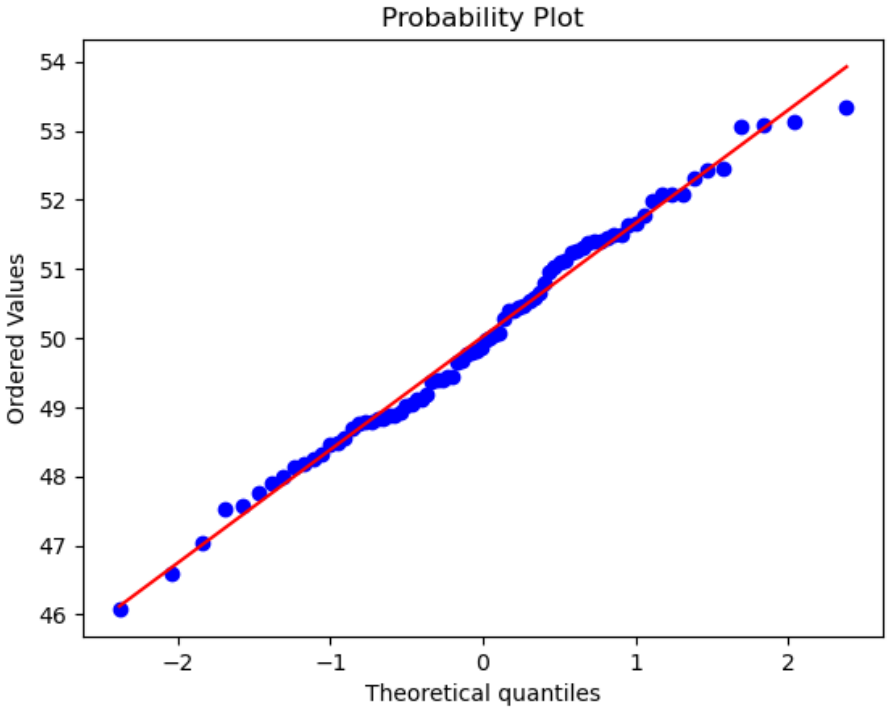
1)

Import the temperature data and plot them.



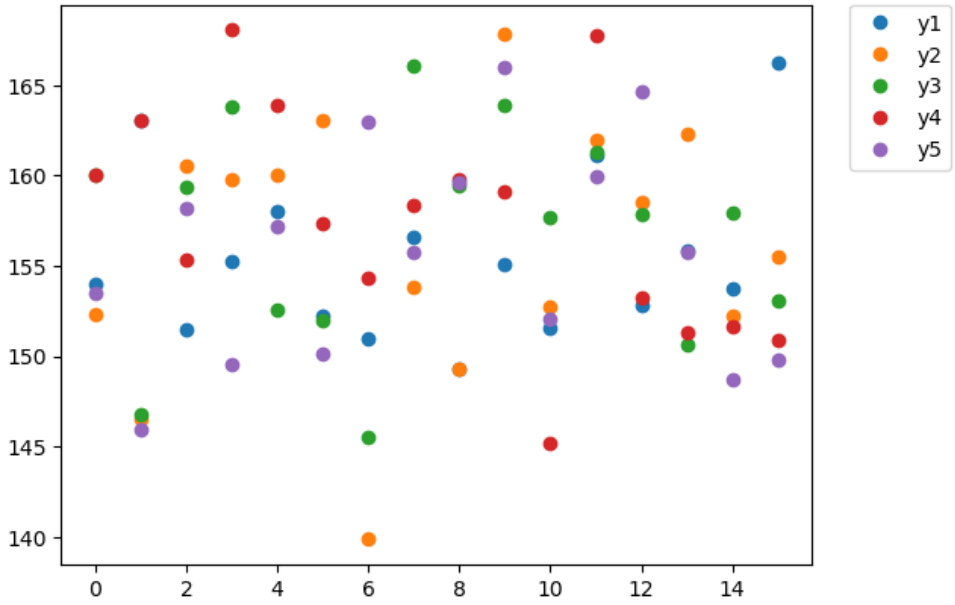
No information is given out about the time order within each sample and we need to rely on a qualitative assessment to evaluate randomness. No strange patterns appear in the data and the process seems stationary. I have no reason to reject the randomness hypothesis.

Let's check the normality.



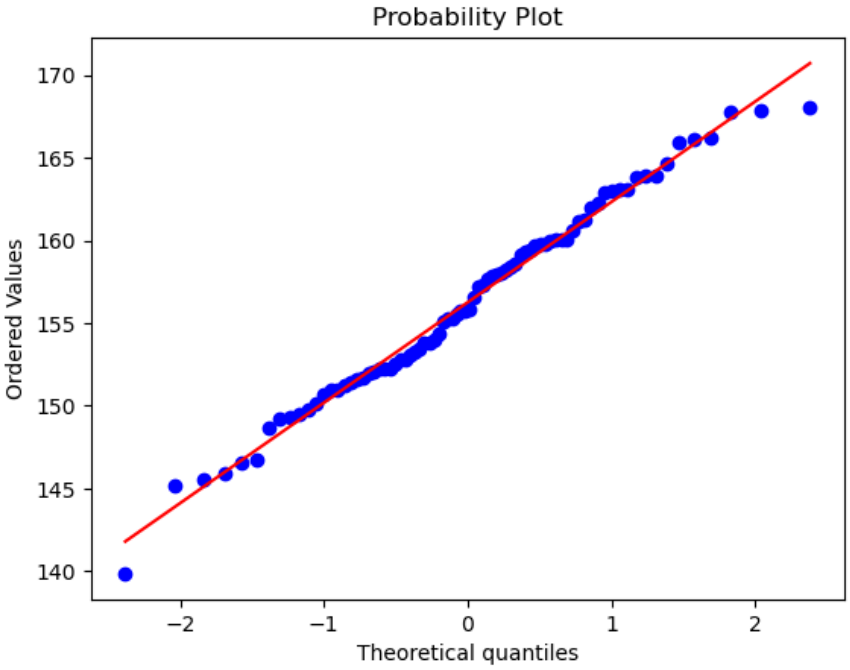
We cannot reject the normality hypothesis (SW p-value = 0.611).

Let's import the pressure data as well.



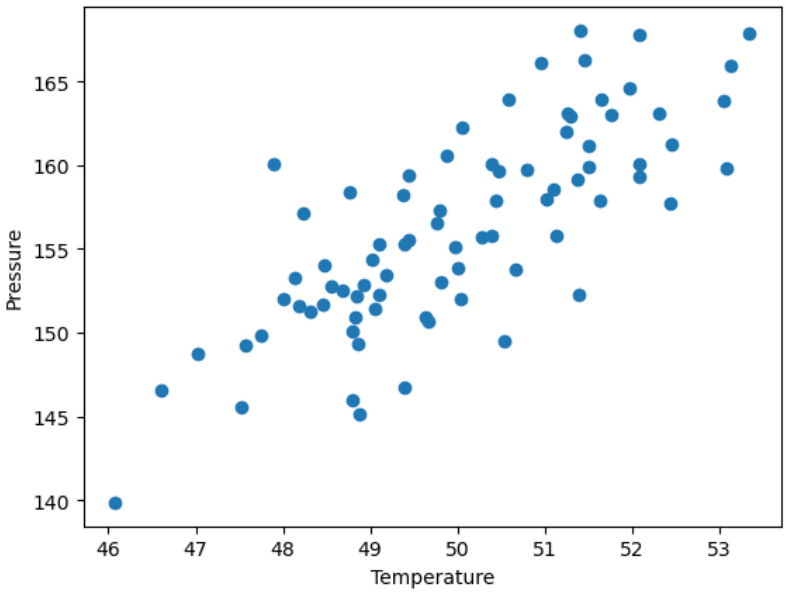
No trend in the data and observations appear to be random.

Let's check the normality.



Normality hypothesis cannot be rejected (SW p-value = 0.599).

Let's plot temperature and pressure data together. The scatterplot reveals that the two variables are positively correlated.



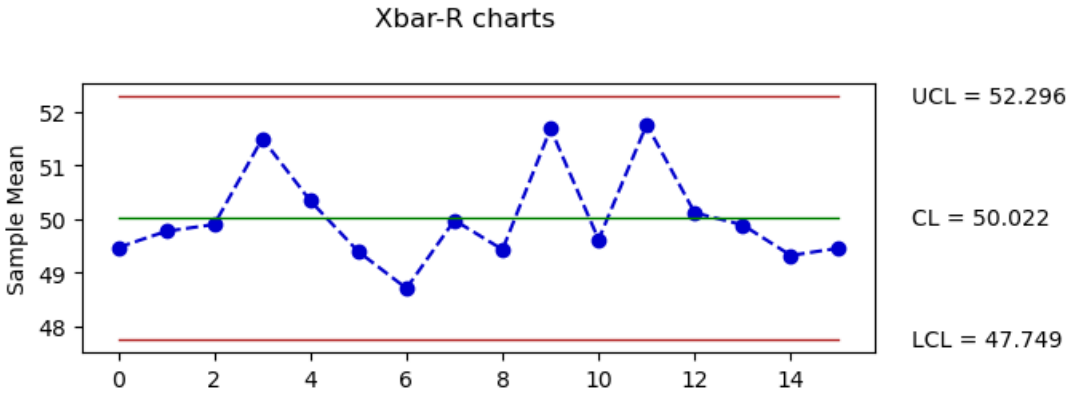
MECH ENG STUDENTS:

ARL0 is set to 500. Therefore, the corresponding family-wise error rate alpha_fam is 1/ARL0 = 0.002. To design multiple CCs we need to apply a correction (i.e., Bonferroni as the data are correlated) and estimate the control limits.

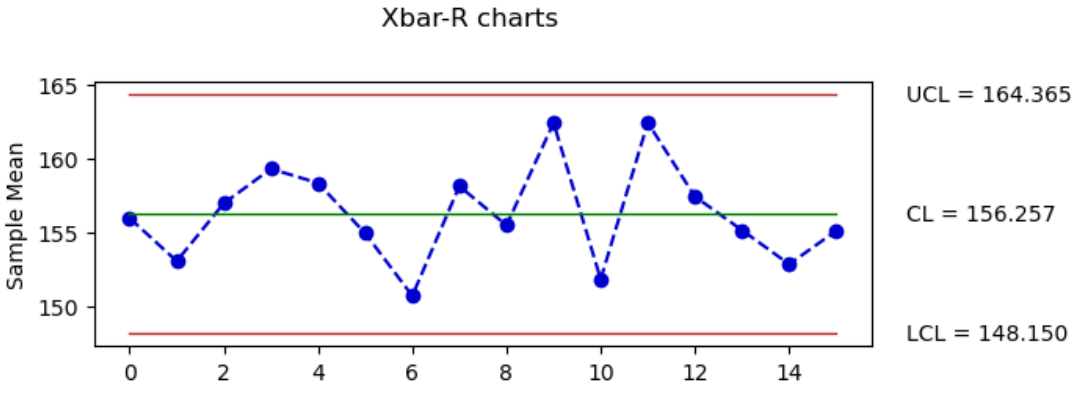
alpha = alpha_fam/2 = 0.001

K = 3.291

Xbar for temperature.



Xbar for pressure.



No out-of-control is present. The control charts are adequate, and we can conclude the design phase (phase 1).

OTHER STUDENTS

Since the data are correlated, we can design a T2 control chart.

First, we compute the grand mean (Xbarbar):

```
temp      50.022375
press    156.257500
```

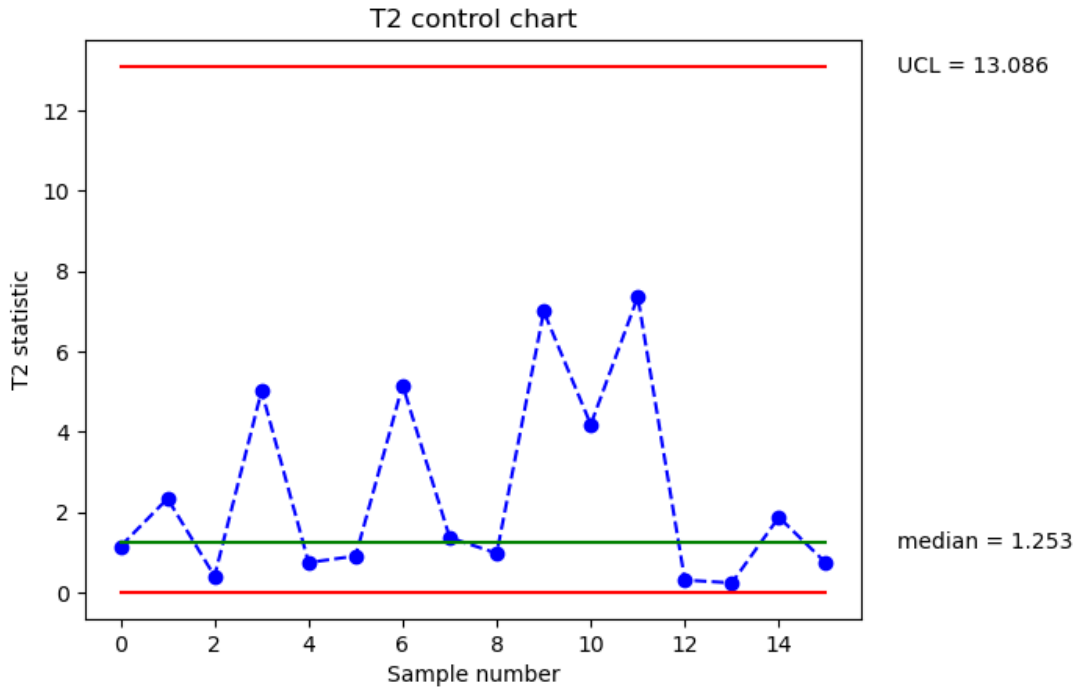
Then we compute the variance-covariance matrix (S):

```
            temp      press
temp    2.299776   6.143086
press   6.143086  30.990074
```

We can now compute the T2 statistic for each of the 16 samples and compare them with the UCL:

```
UCL = (p * (m-1) * (n-1)) / (m * (n-1) - (p-1)) * stats.f.ppf(1-alpha_fam, p, m*n
- m + 1 - p) = 13.086
```

Where p = 2, m = 16, n = 5.

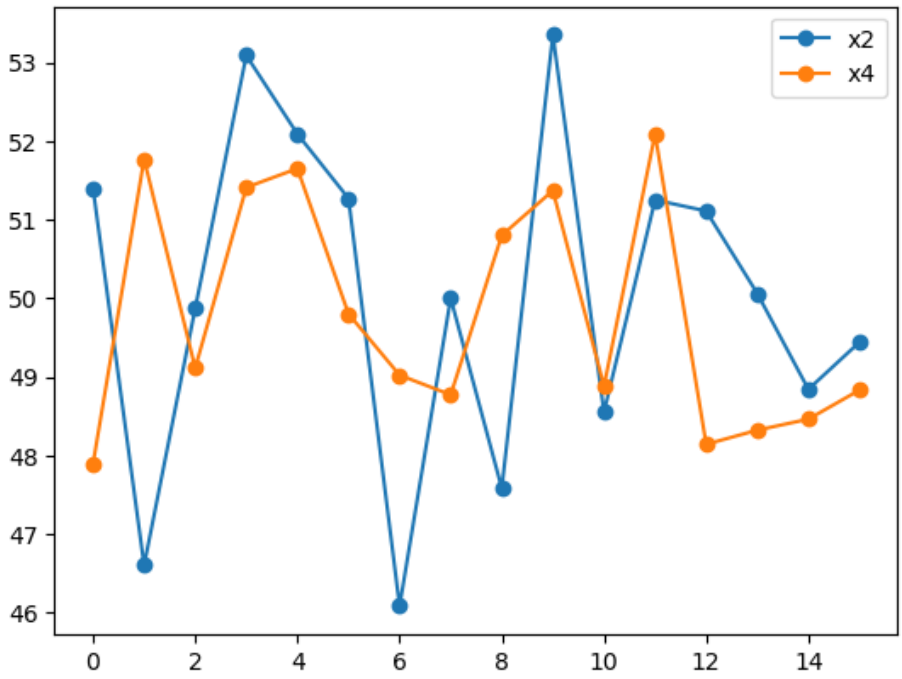


All the samples are in-control. Design phase is finished.

2)

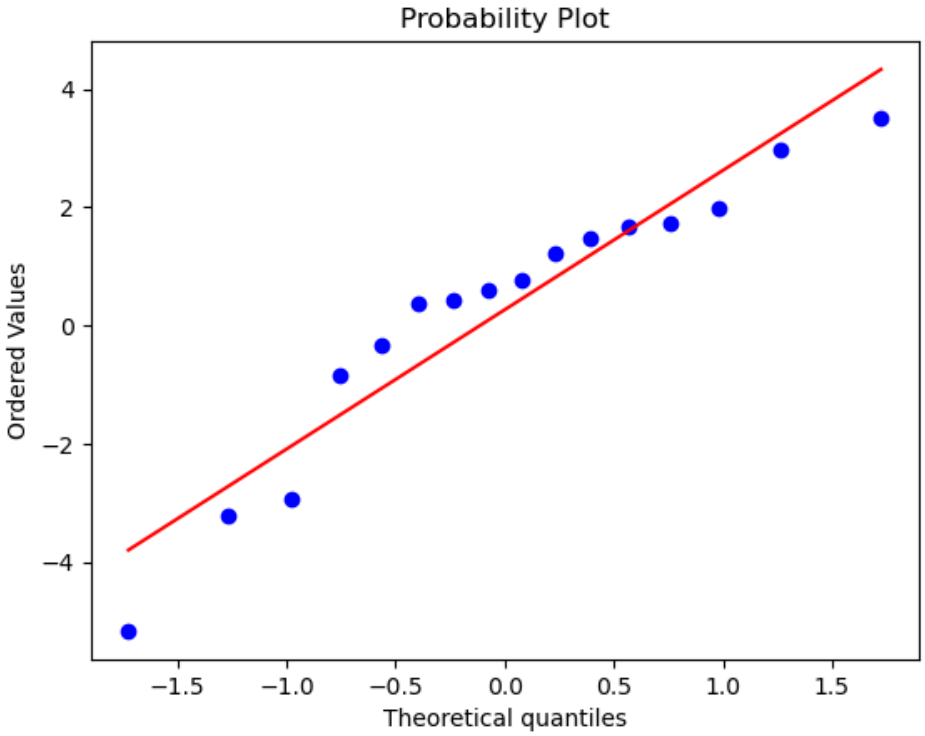
We can perform a paired t-test.

First, we can plot the temperature at the two locations:



Process seems stationary with no significant deviations.

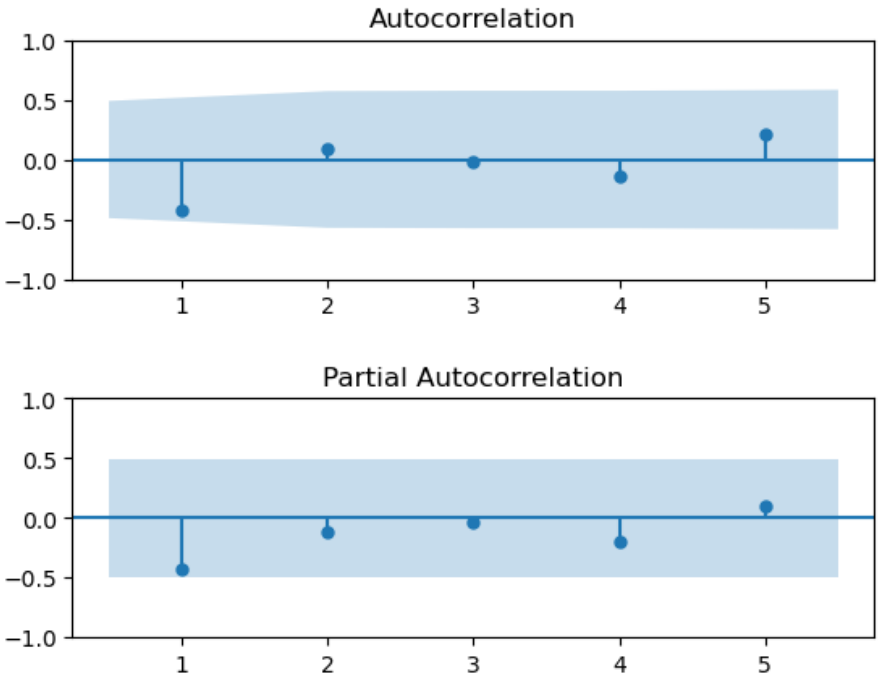
Now we can compute the difference between the paired data (d = x2 – x4) and check for normality and randomness of the difference variable.



We cannot reject the normality hypothesis (SW p-value = 0.131).

From the runs test the data appear to be random (p-value = 0.493).

Let’s compute the sample ACF and PACF.



Lag 1 seems to be borderline. Let's perform the Bartlett test at 5% significance level.

```
Test statistic r1 = 0.425751
Rejection region starts at 0.489991
```

The autocorrelation at lag 2 is not significant. The difference is normal and random.

We can now perform the paired t-test.

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d > 0$$

$$t_0 = \frac{\bar{d}}{s_d/\sqrt{n}}$$

```
t-statistic: 0.459
p-value: 0.327
```

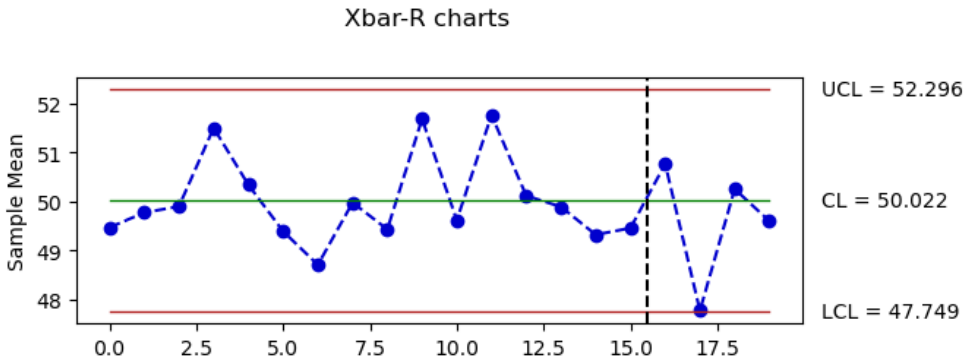
Based on the result we got from the paired t-test, there is no statistical evidence to reject H0.

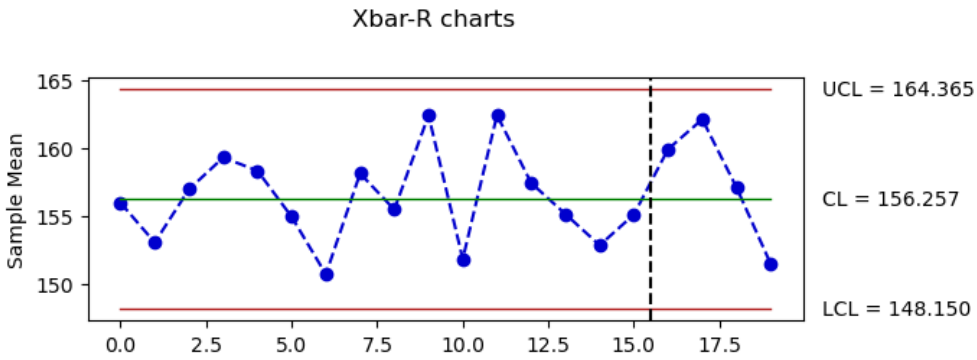
3)

Using the control limits, grand mean and variance-covariance computed before:

MECH ENG STUDENTS

We estimate the Xbar statistics for the new observations and compare them with the control limits.

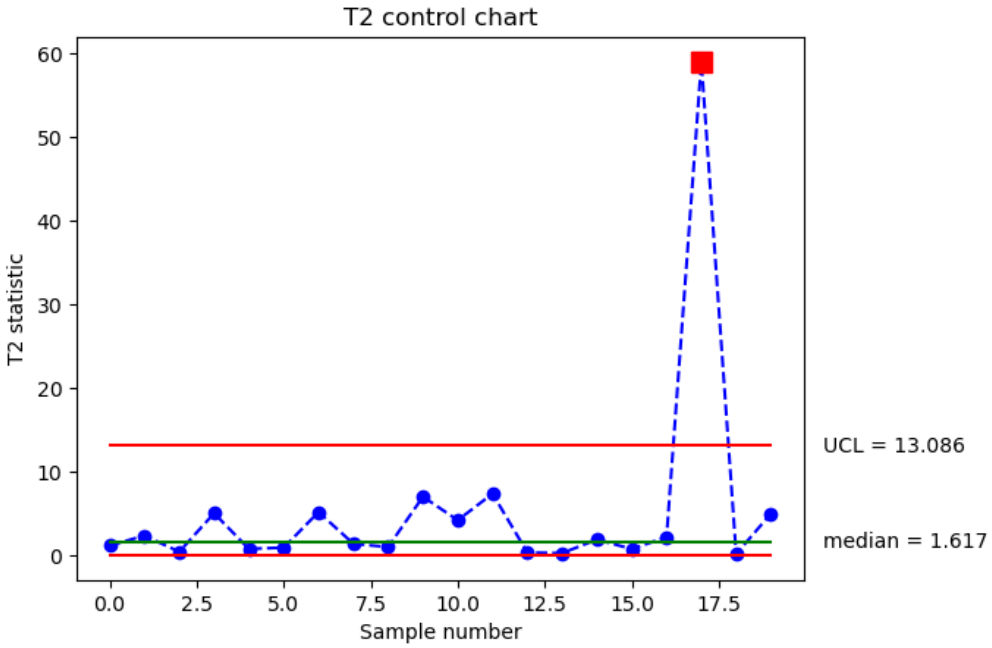




All the new observations appear to be in-control.

OTHER STUDENTS

We estimate the T2 statistics for the new observations and compare them with the upper control limit.

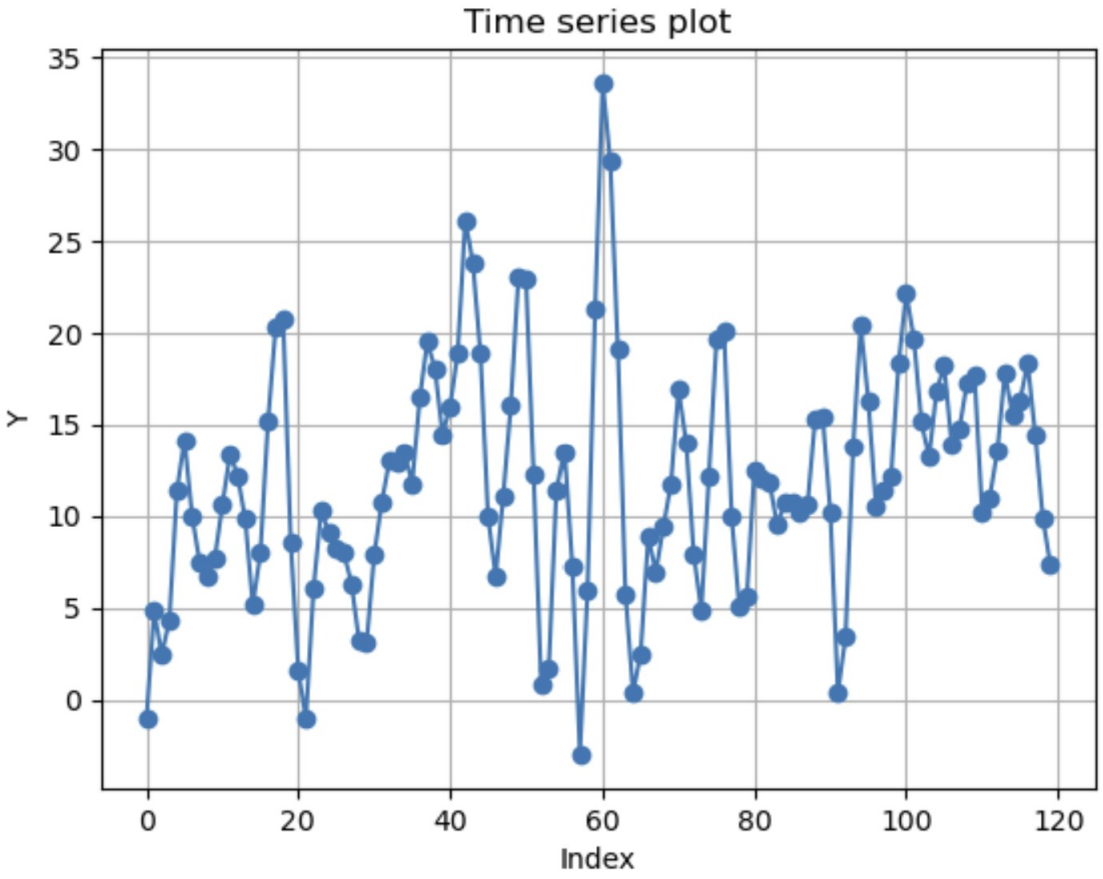


The second sample of phase 2 data is out-of-control. This could only be captured thanks to the multi-variate approach which takes into account the correlation structure between the two random variables.

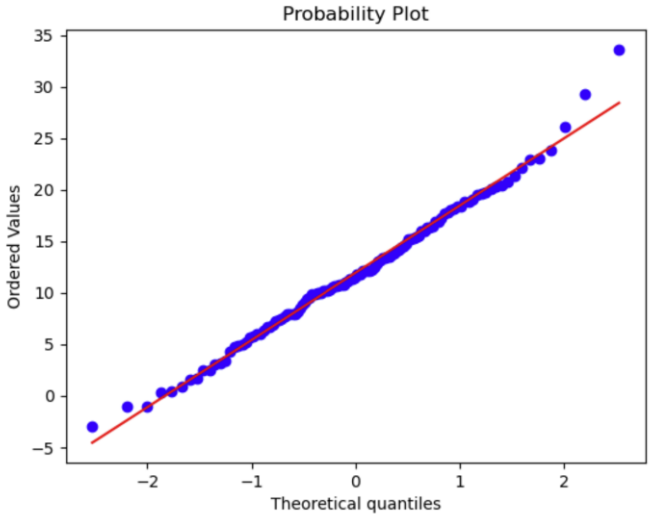
Exercise 2 solution

1)

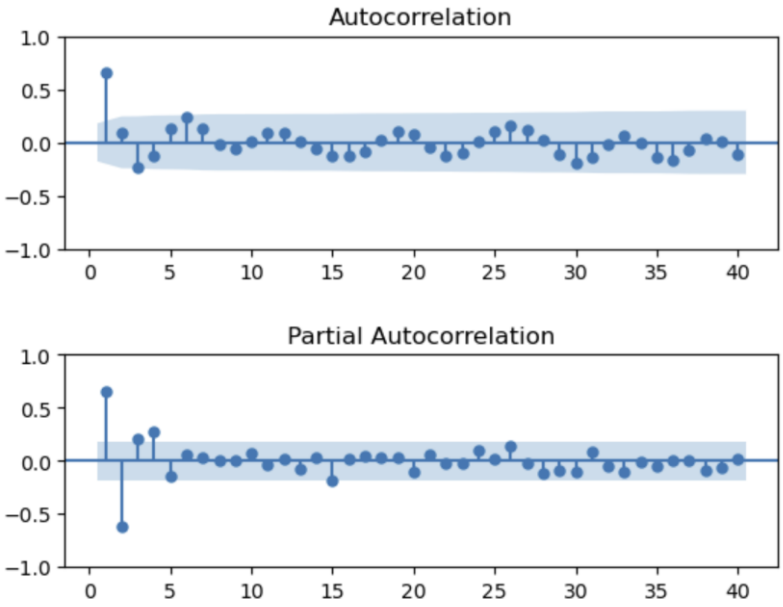
Data snooping: the time series of the performance indicator is the following:



A meandering pattern may be present. Indeed, statistical tests highlight that the data are normal but not independent:



Shapiro-Wilk test p-value = 0.606



LBQ test with L = 10:

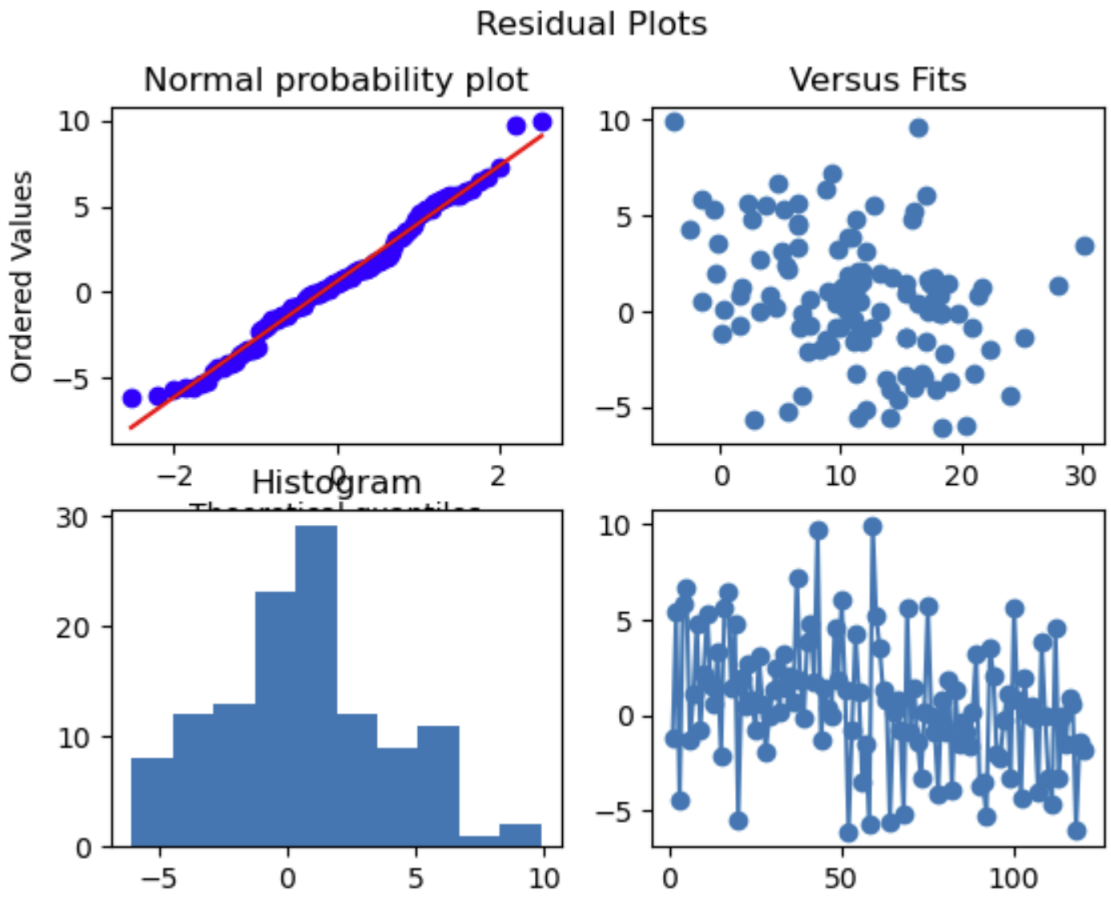
- Q0_LBQ = 75.172
- p-value = 0.000

The SACF and SPACF pattern may be interpreted in different ways, either an AR model of order ≥ 2 or an MA(1). However, neither an AR nor an MA model alone are suitable to fit this time series. An appropriate model is an ARMA model where both AR and MA terms are included to capture the temporal dependence of the data. An appropriate model results to be an ARMA(2,2):

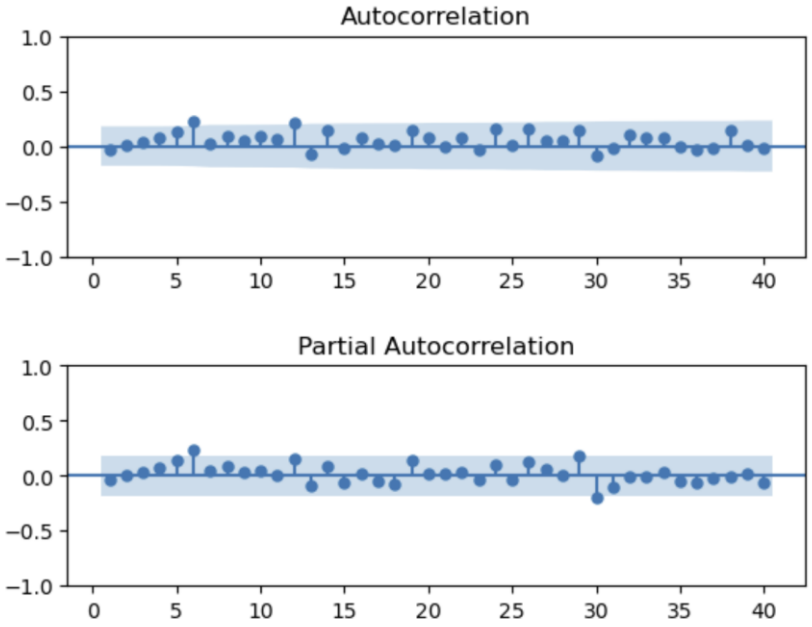
```
---------------------
ARIMA MODEL RESULTS
---------------------
ARIMA model order: p=2, d=0, q=2

FINAL ESTIMATES OF PARAMETERS
------------------------------
 Term    Coef  SE Coef  T-Value    P-Value
   x1  0.1586   0.0185   8.5622 1.1074e-17
ar.L1  0.6195   0.1207   5.1334 2.8454e-07
ar.L2 -0.3152   0.1149  -2.7433 6.0831e-03
ma.L1  0.9097   0.0925   9.8344 8.0050e-23
ma.L2  0.7382   0.0851   8.6775 4.0440e-18

RESIDUAL SUM OF SQUARES
-------------------------
   DF        SS      MS
115.0 1336.5537 11.6222
```



Shapiro-Wilk test p-value on the residuals = 0.161



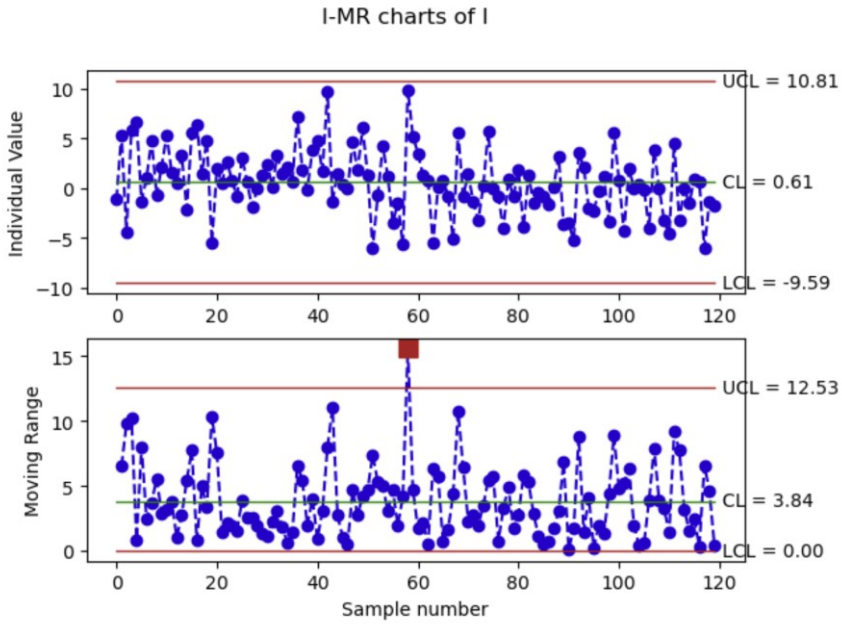
LQB test with L = 10:

- Q0_LBQ = 12.235
- p-value = 0.269

The model is adequate as residuals are normal and independent. All terms are significant.

2)

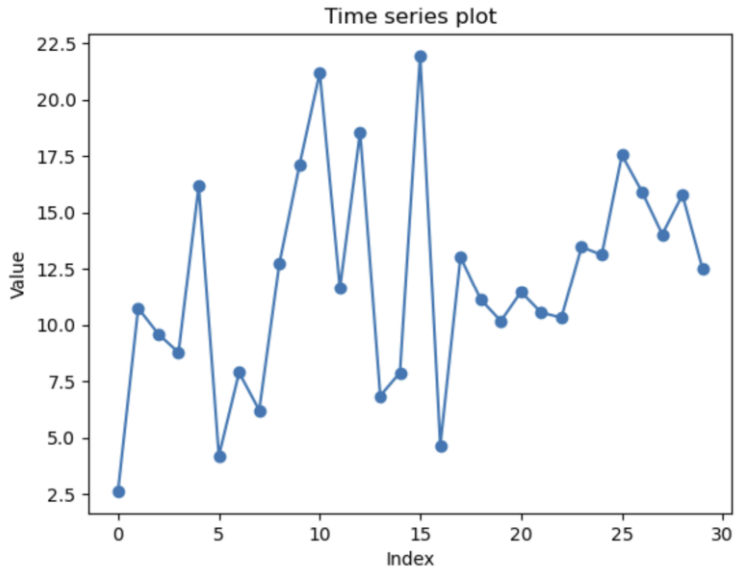
The special cause control chart applied to the model residuals with K = 3 is the following:



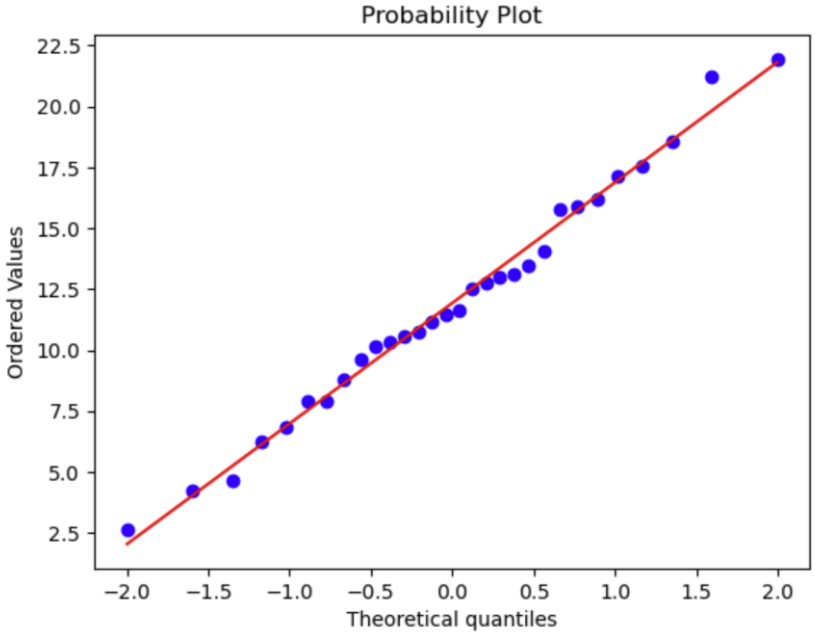
There is only one point (observation 58) that violates the upper control limit in the MR chart. Since no assignable cause is present, no further action is needed and the control chart design is over. Note that this violation might be also caused by an intrinsic violation of the normality assumption in the MR: a probabilistic control chart may be used.

3)

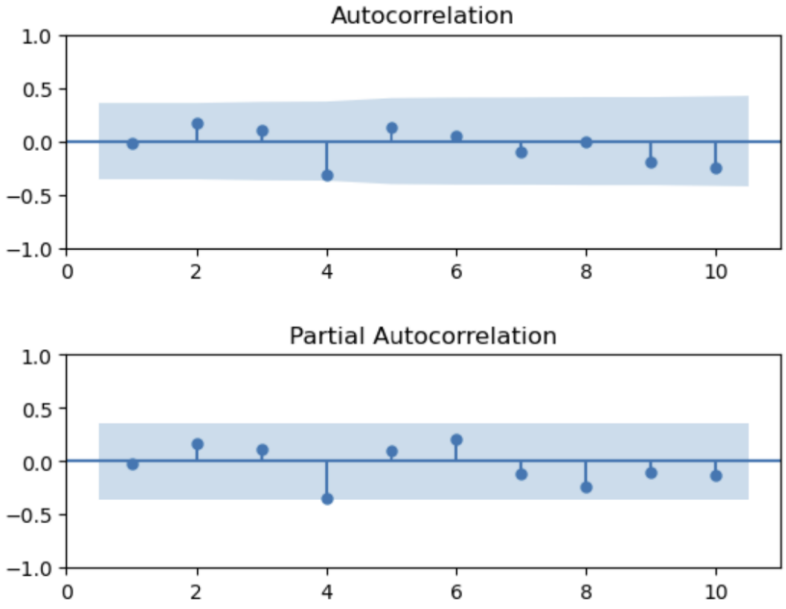
By applying a batching operation with batch size = 4, we get the following time series:



After batching data are normal and independent:



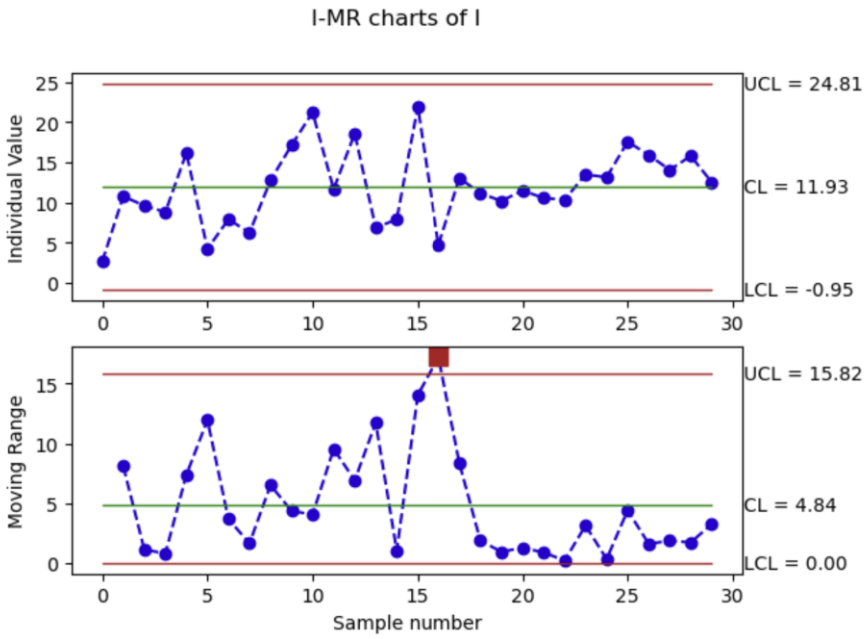
Shapiro-Wilk test p-value = 0.966



LBQ test with L = 5:

- Q0_LBQ = 5.585356
- p-value = 0.348677

Therefore, it is possible to design the following I-MR control chart:



Also in this case there is one violation of the upper control limit in the MR chart (batch number 16). There is also a stratification that starts after the 16th batch and, possibly, a slight increasing trend in the batched data. Since no assignable causes are present, the control chart design is over, but attention should be paid to these patterns that may indicate a change in the process. It is worth noticing that the batching operation entails a filtering of the data that, on the one hand, may be beneficial in capturing patterns like the ones mentioned above, which may be hardly visible in the original time series. On the other hand, there is also the risk of filtering out events and patterns of actual interest. The batching operation also implies a slower reaction to actual changes in the process, which may have possible detrimental effects on the final control chart performance. Pros and cons of data batching should be carefully taken into account.

Exercise 3 solution

Question 1

Answer: c

Explanation: (a) and (b) carry equivalent information (the latter is obtained from the former) and they provide what percentage of the total variance, the PCA approximation we can achieve, and so they are useful to decide how many components we will keep. Similarly (d) is a plot of either (a) or (b), where the "elbow" method can be used to decide for the number of components we will keep. For (c) we do know that the PCA results will be different depending on whether we apply PCA to original or scaled variables, but this is not related on the decision making of how many components we will need to keep in the approximation.

Question 2

Answer: b

Explanation: Since we have 5 variables, we know that the cumulative proportion of the explained variance from the 5 principal components needs to be 100%. Answers (c) and (d) have a sum over the first four components to be 1.01 and 1.05 respectively, i.e., they exceed 100% and so they are wrong. In (a) the sum is exactly 1 but since it was given that we did not have two perfectly correlated variables (i.e., $\left|r\left(X_i, X_j\right)\right| < 1$, for $i \neq j$) we cannot have a redundant component (i.e., a component with 0% explained variance) and so (a) is also invalid. In (b) the sum is 0.93 and this is a valid option (where the last component will explain the remaining 1-0.93=0.07, i.e., 7% of the total variation).