QUALITY DATA ANALYSIS

29/01/2024

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: you can skip Exercise 2, point 4), Exercise 3, question 2).**

Exercise 1 (14 points)

The closing prices of a stock for 60 consecutive trading days are reported in `stock_price_phase1.csv`.

1) Find an adequate model to fit the data.
2) Design the appropriate control charts to monitor the price of the stock such that the average number of trading days between two false alarms is 200. *Note: in case of violations of control limits, assume no assignable cause was found.*
3) Using the control chart(s) designed in point 1 (phase 1), check if the data collected during the following 20 trading days (stored in `stock_price_phase2.csv`) are in control. Report the index of the OOC points, if any.

Exercise 2 (15 points)

An oil & gas company has implemented a new quality assurance protocol to keep under control their welding operations. During a laser welding process, four critical quality characteristics are monitored. The data collected with the process in its regime condition are stored in `PCA_phase1.csv`. The time order of the data corresponds to the actual time order of process execution and data collection. The head of the quality department is interested in designing and testing a control chart based on Principal Component Analysis.

1) For these data, is it more appropriate to apply the PCA using the variance-covariance matrix of the data or their correlation matrix? Motivate your answer.
2) Based on the outcome of point 1, apply PCA to the available data and determine the number of principal components that should be retained to capture at least 60% of the total variance (report the eigenvectors and the eigenvalues of the retained components). Discuss the results trying to interpret the retained PCs.
3) Based on the result of point 2, design multiple univariate control charts to monitor the laser welding process with a family-wise $ARL_0 = 350$. In case of violations of control limits, assume the existence of assignable causes.
4) Assume the company is interested in monitoring only the first PC using an I chart, with $ARL_0 = 350$. What is the probability of not detecting a shift of the process along the first PC whose size is 2.5 standard deviation units?

Exercise 3 (4 points)

Question 1)

$X_t, t = 1, 2, 3, \ldots$ is a stationary AR(1) process with ρ_1 being its autocorrelation of order 1. If we will perform the transformation: $X_t^* = c * X_t, t = 1, 2, 3, \ldots$ with $c \neq 0$ being a constant, then for the autocorrelation of order 1 for X_t^*, i.e., ρ_1^* which of the following is valid?

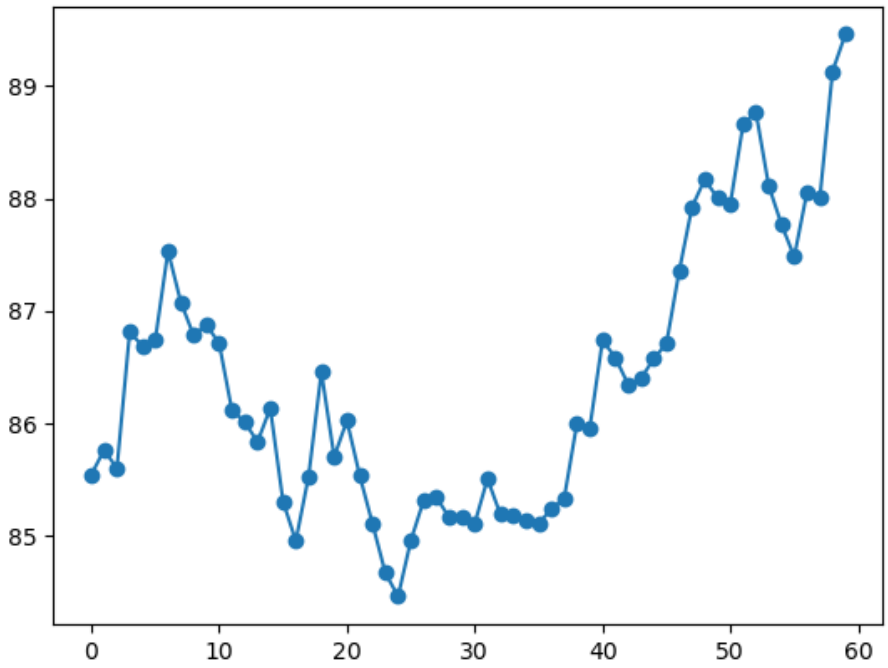
a. $\rho_1^* = \rho_1$
b. $\rho_1^* = c * \rho_1$
c. $\rho_1^* = c^2 * \rho_1$
d. $\rho_1^* = c + \rho_1$

Question 2)
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 the probability to get a point outside of the region $[LCL, UCL]$ is $\alpha = 0.05$. If we will decide to change α and select $\alpha = 0.01$, then which of the following statements will **not** be valid?
a. The type I error will decrease.
b. The power will decrease.
c. The in-control Average Run Length (ARL_0) will increase.
d. The out-of-control Average Run Length (ARL_1) will decrease.

Exercise 1 solution

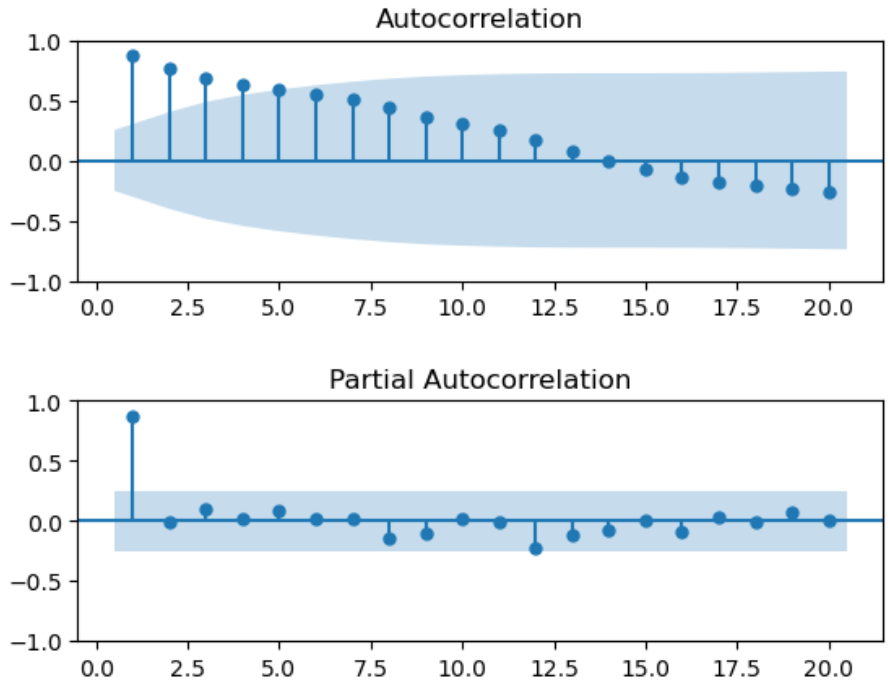
1)

Import the data and plot them.



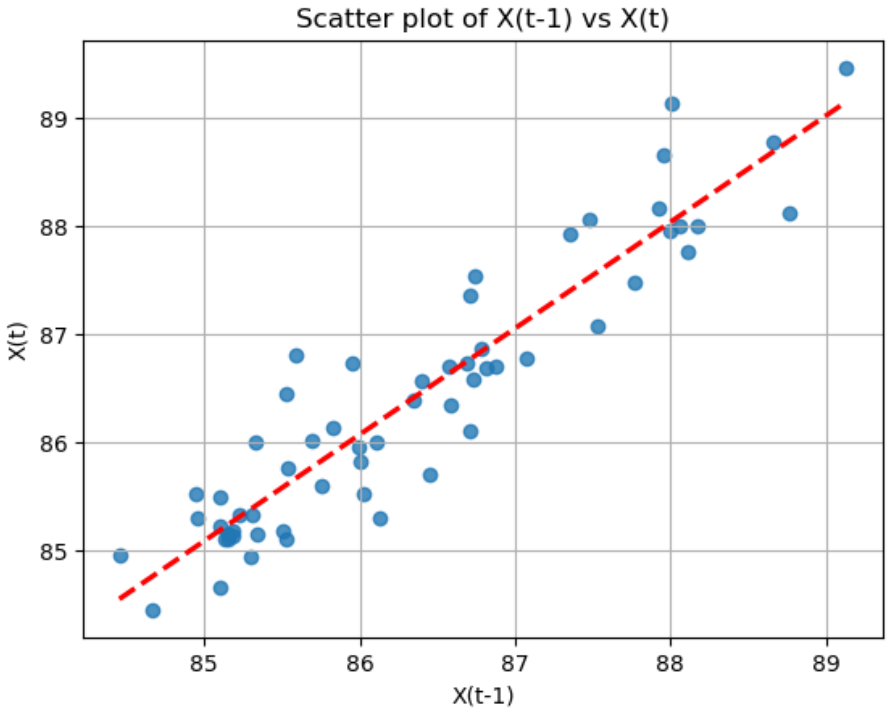
A meandering patters seems to be present in the data and the process does not seem to be stationary.

Let's check the randomness.

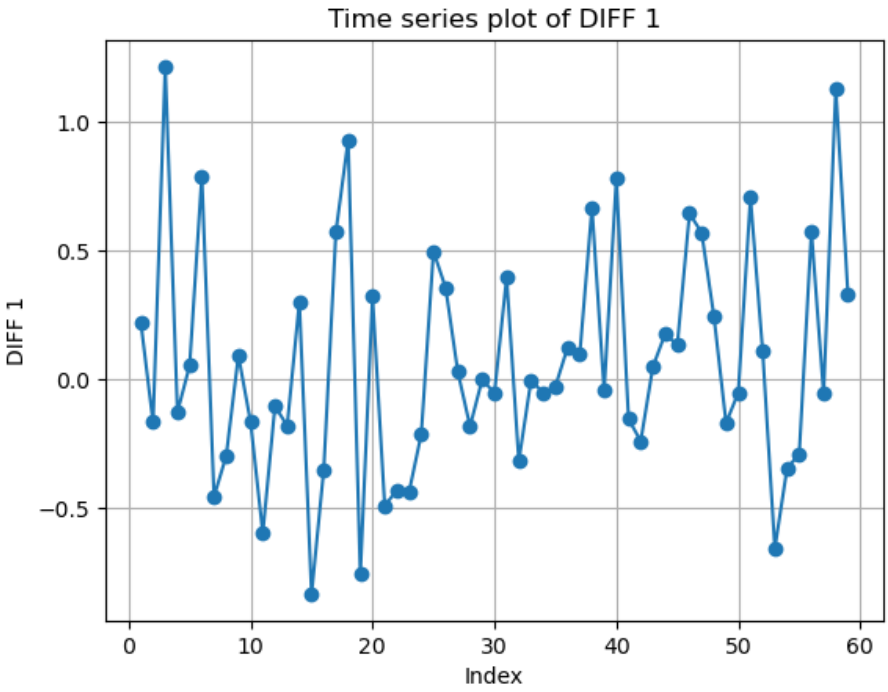


The runs test returns a p-value of 0.000, and the sample ACF shows a linearly decaying trend, which is typical of non-stationary time series. Based on ACF analysis, we can state that the process is non-stationary

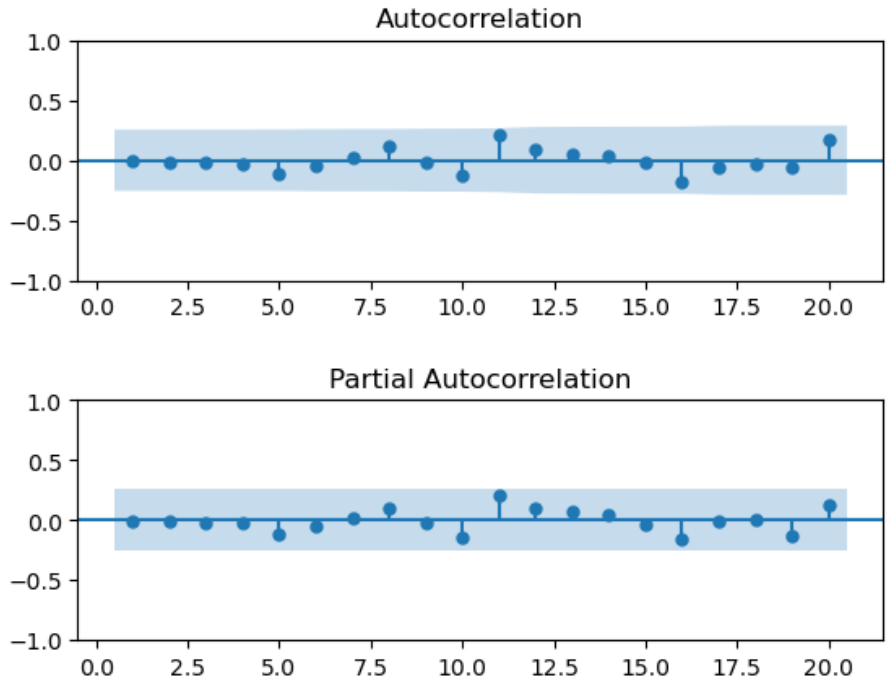
Moreover, we can observe with a scatterplot the correlation between X(t) and X(t-1).



The lagged time series is very correlated with the original time series. Let’s get rid of the non-stationarity by differencing the time series.

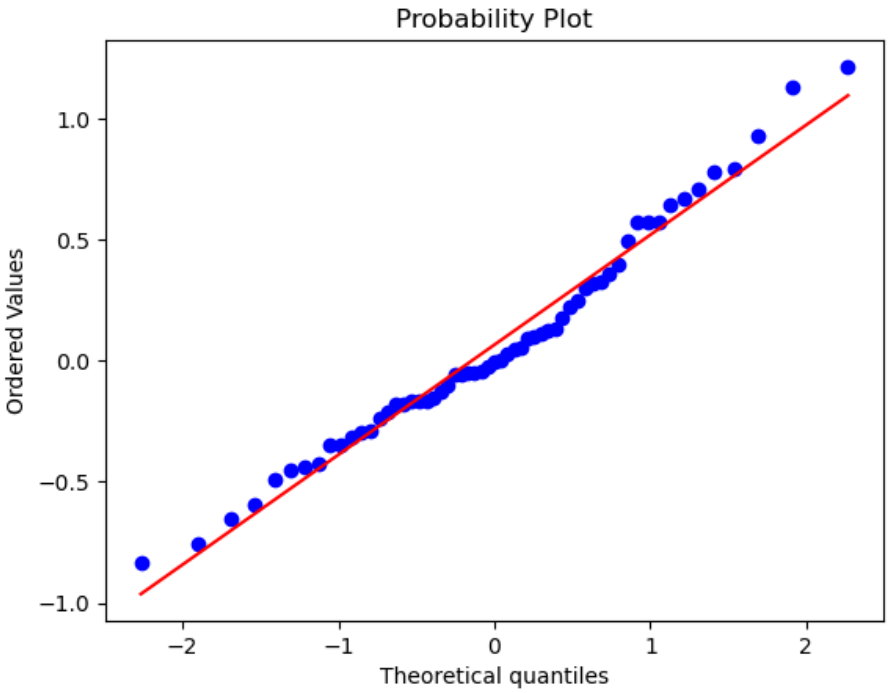


Now everything seems stationary. Let's check randomness.



The runs test returns 0.827 and no strange pattern appears in the sample ACF/PACF.

Let's check the normality.



The Shapiro-Wilk test returns a p-value of 0.311.

The data, after applying the differencing operator, are random and normally distributed. The data can be modeled as a random walk:

$Y_t = Y_{t-1} + \epsilon_t$

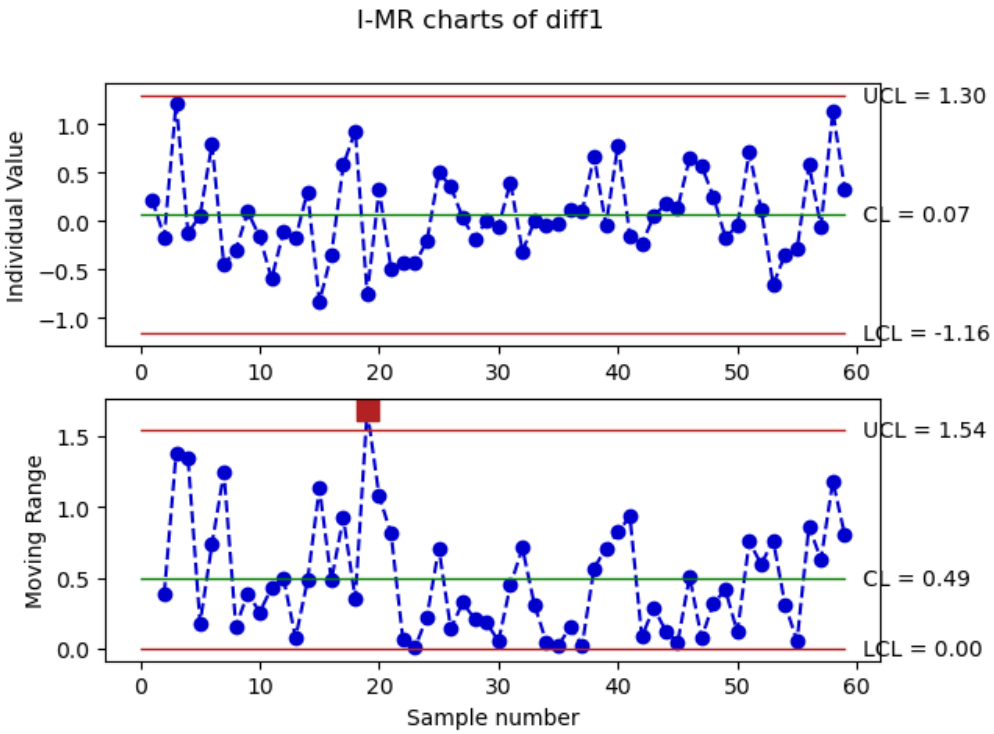
2)

We can design an I-MR control chart on the residuals.

The ARL0 is set to 200.

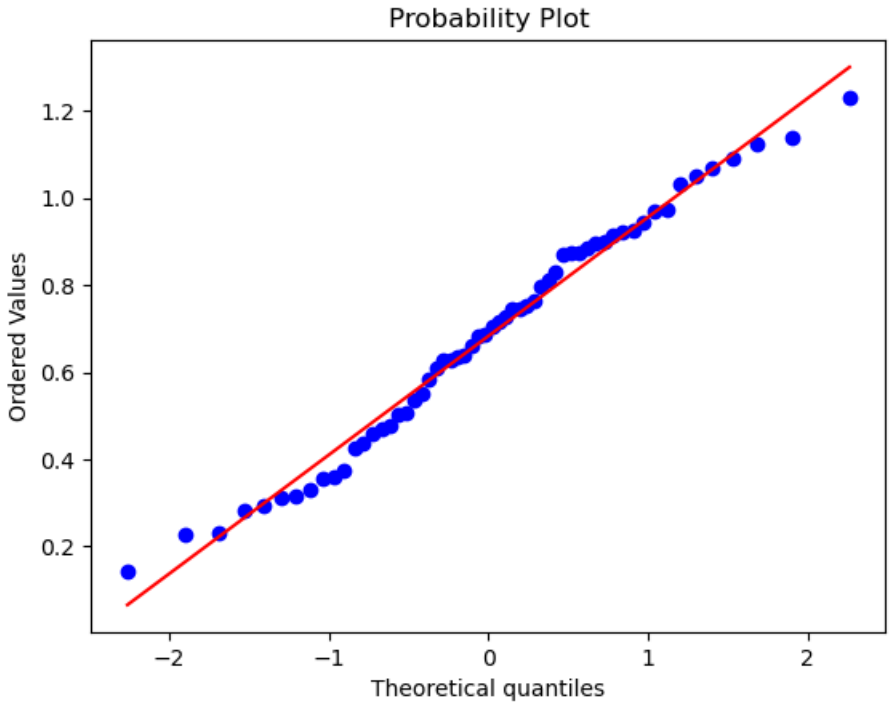
alpha = 1/ARL0 = 0.005

K = 2.807

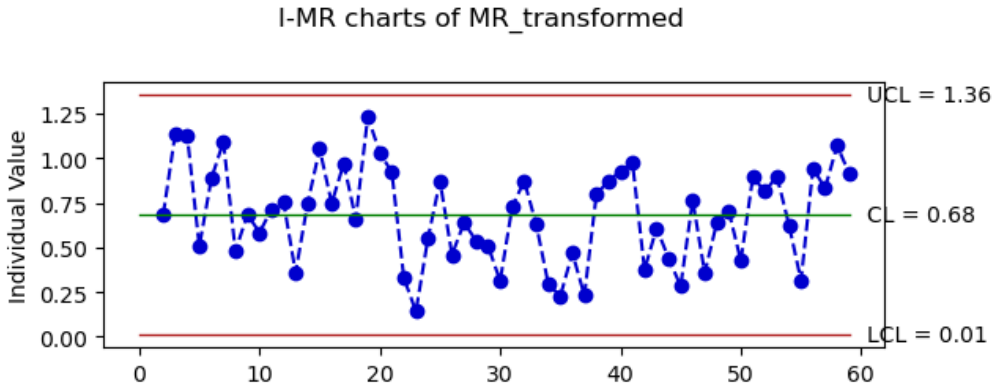


One moving range value seems to be OOC.

Let's use the normality transformation on the MR data to make them normal. We know we can apply a power transformation with lambda = 0.4. After transformation, the MR series follows a normal distribution (SW p-value = 0.427).



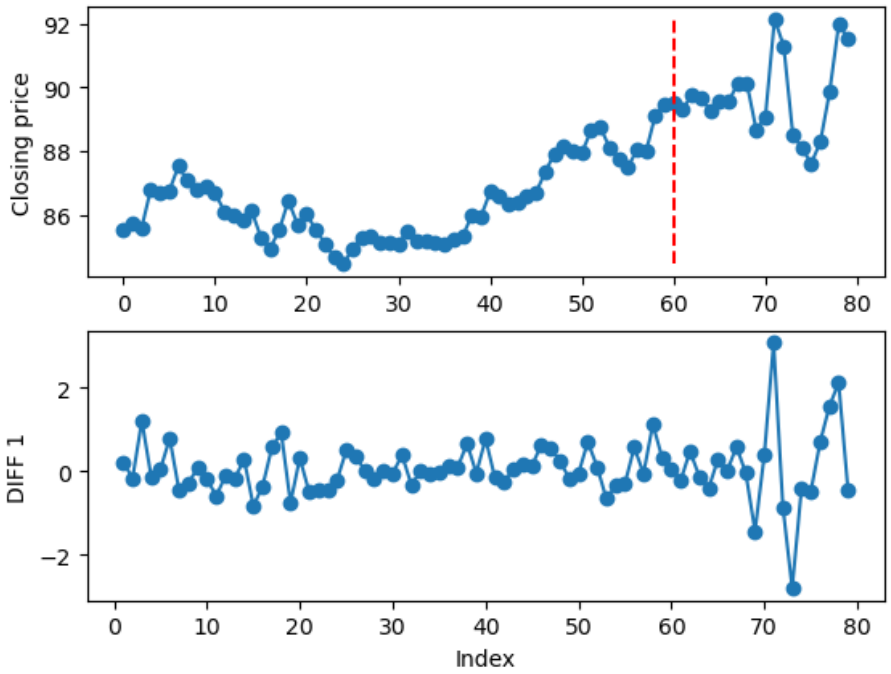
The individual control chart on the transformed variable can be designed.



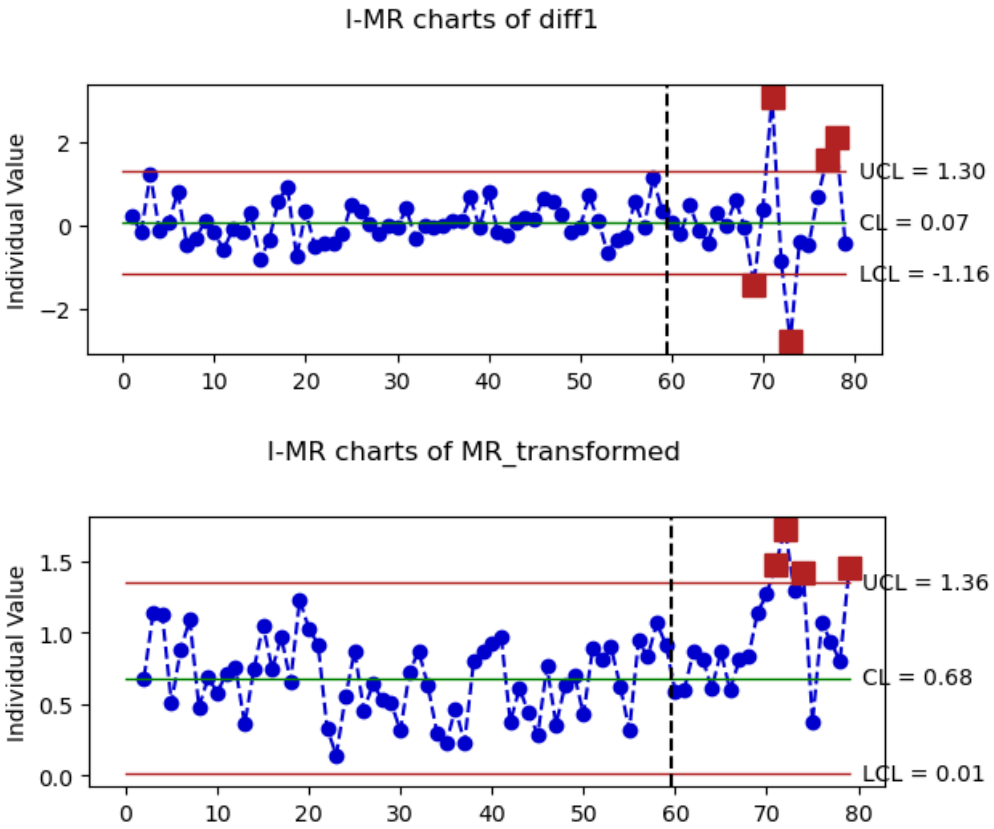
No points are out of control, the design phase is complete.

3)

Let’s plot the new data next to the phase 1 data.

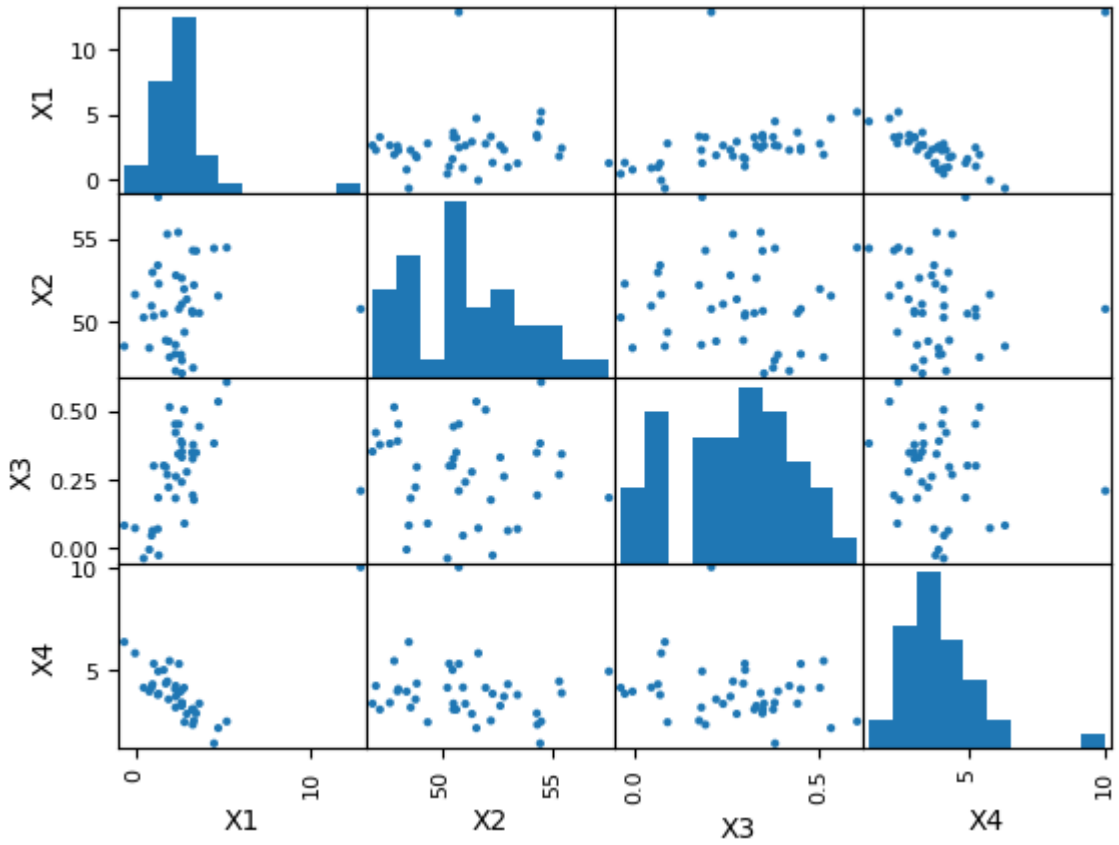


The new data seem to show a higher volatility, with larger fluctuations compared to the phase 1 data. The control chart confirms this and highlights a large number of OOC points in the I-chart (69, 71, 73, 77, 78) and in the MR chart after transformation (71, 72, 74, 79).



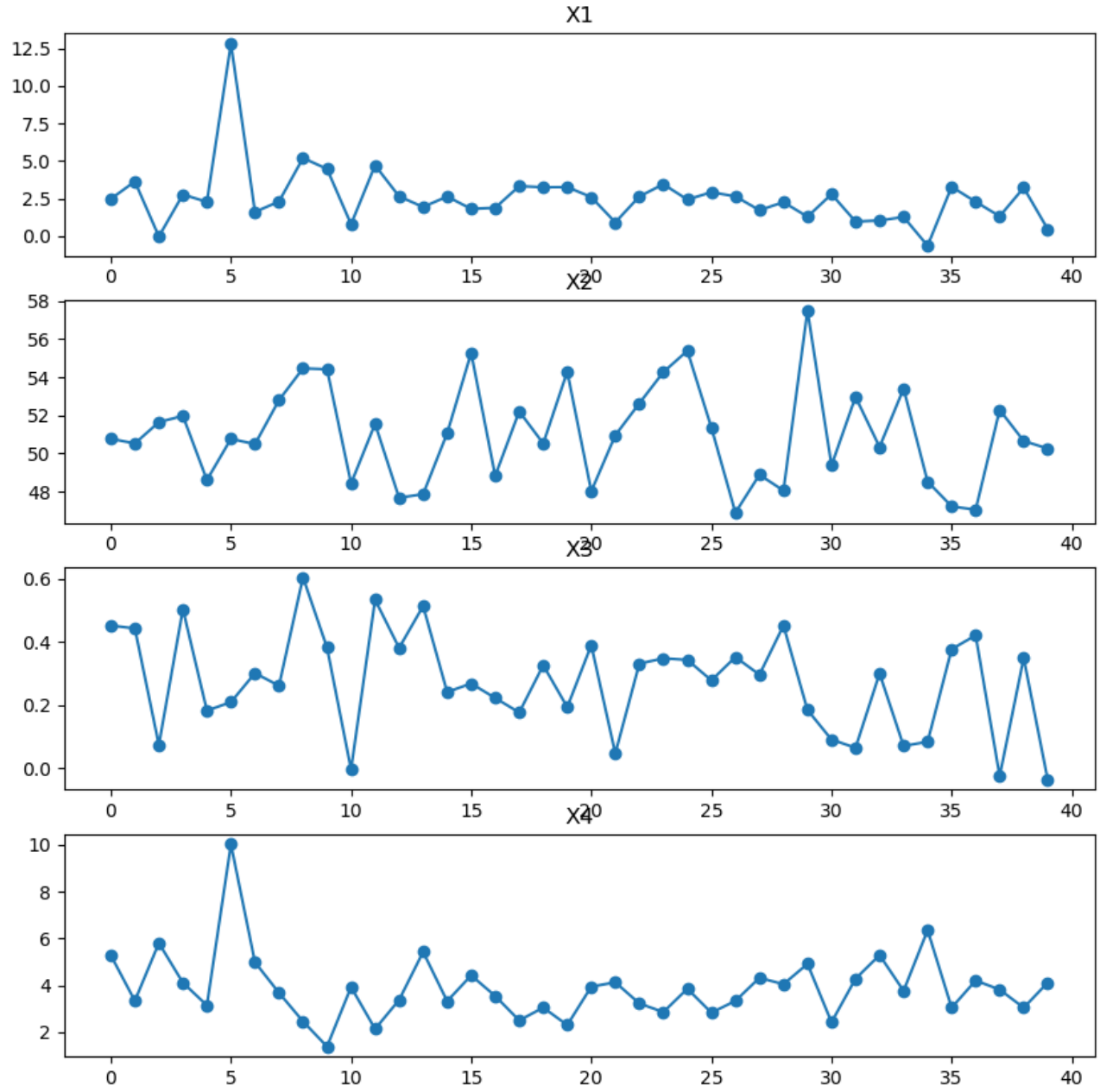
Exercise 2 solution

1) Let's inspect the data first.



An outlier is possibly present (outlying value of variables 1 and 4). We may take care of this in the following analysis.

We may also look at time series plots.



Apart from the possible presence of an outlier in variables 1 and 4, no other systematic pattern seems to be present.

Let's estimate the sample mean and the variance/covariance matrix

```
Sample Mean:
X1     2.571600
X2    51.005850
X3     0.274325
X4     3.899400
dtype: float64

Sample Variance-Covariance Matrix:
           X1        X2        X3        X4
X1  4.242852  0.453980  0.121228  0.621575
```

```
X2  0.453980  6.756714 -0.034737 -0.518304
X3  0.121228 -0.034737  0.026465 -0.045179
X4  0.621575 -0.518304 -0.045179  2.090987
```

The variables have quite different marginal variances. Thus, it is more appropriate to estimate the PCA by using the correlation matrix of the original data, which is equivalent to standardize the data and estimate the PCA for the standardized variables.

2)

Let's apply the PCA using the correlation matrix.

Explained variance ratio:

[0.34058614 0.29590993 0.25559052 0.10791341]

Cumulative explained variance ratio

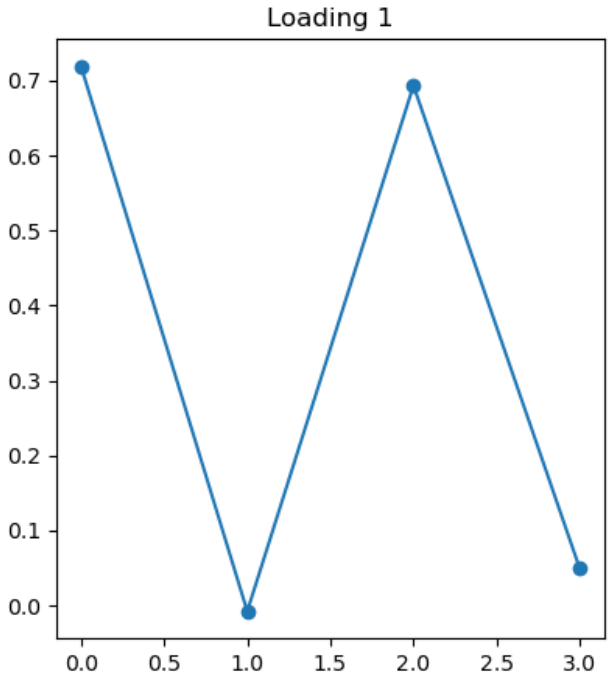
[0.34058614 0.63649607 0.89208659 1.]

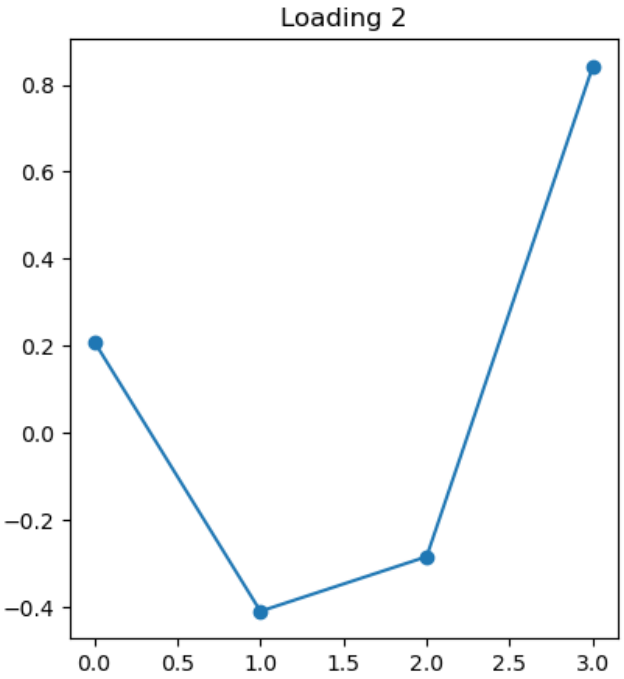
In order to capture at least 60% of the total variance, the first 2 PCs shall be retained.

```
Eigenvalues
 [1.36234456 1.18363973]

Eigenvectors
 [[ 0.71890941 -0.00780065  0.69329239  0.04953852]
 [ 0.21053354 -0.40834727 -0.2830647   0.8419041 ]]
```

Let's plot the loadings:

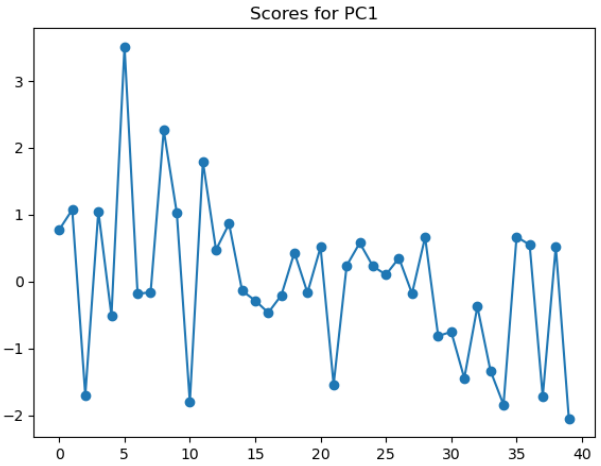


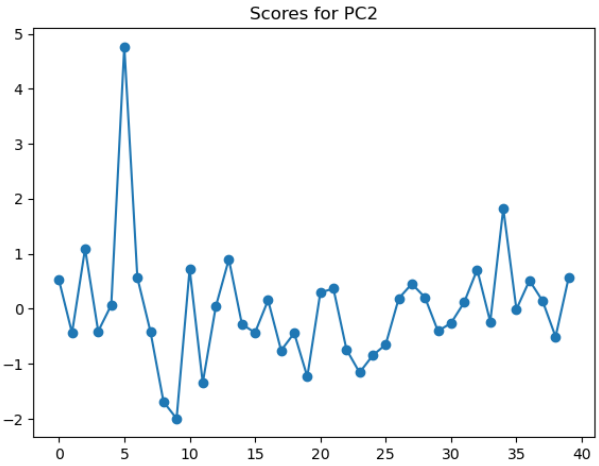


The first PC is mainly influenced by variables 1 and 3 (large positive weights), while variables 2 and 4 have a weight close to 0. The second PC is a contrast between variables 1 and 4 (positive weights) and variables 2 and 3 (negative weights), and variable 4 is the one with the largest weight.

3)

We can design two I-MR control charts for the first 2 PCs. But, first, we need to check the assumptions.

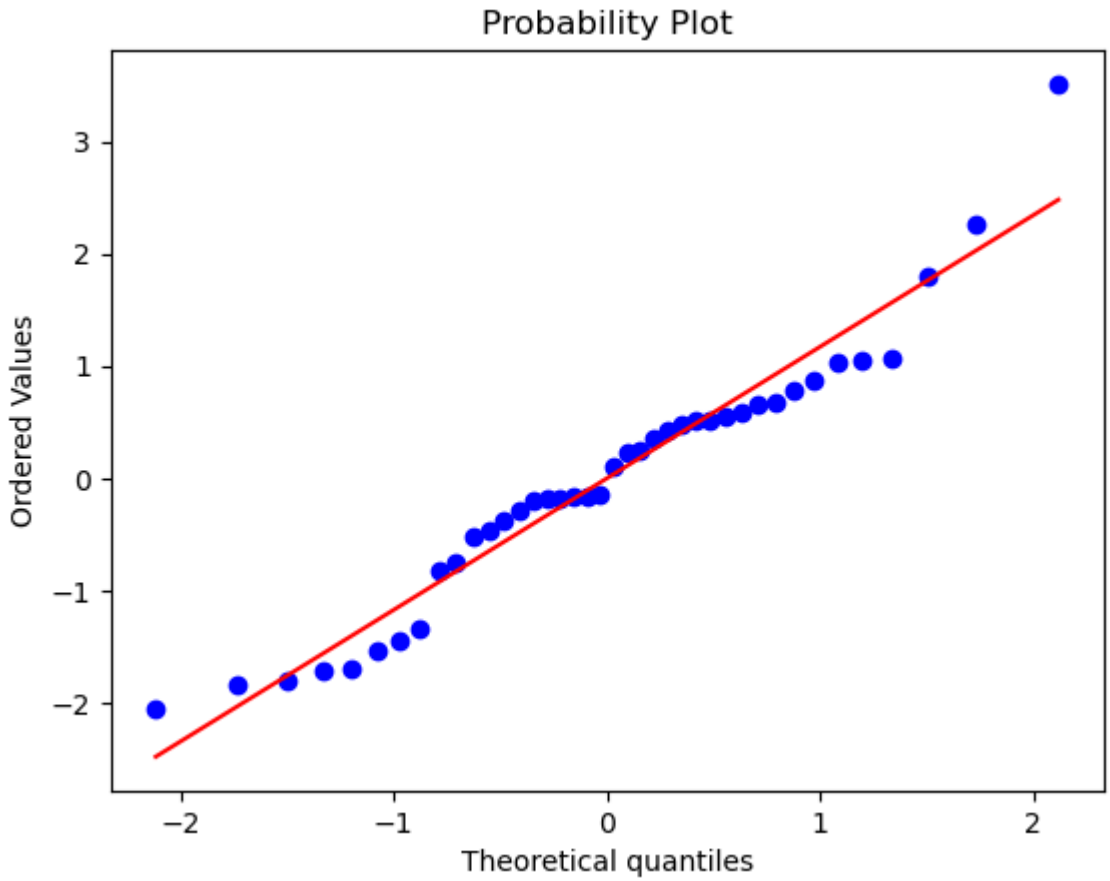




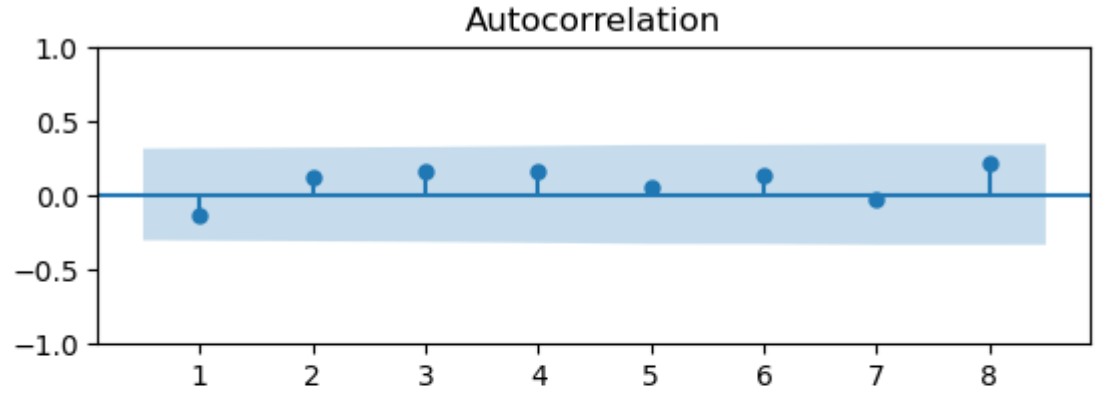
An outlier is possibly present along the second principal component. It is the same datapoint that was observed in the scatterplots of variables 1 and 4. This outlier mainly influences the second PC due to the linear combinations discussed above. Let's test for normality and independence.

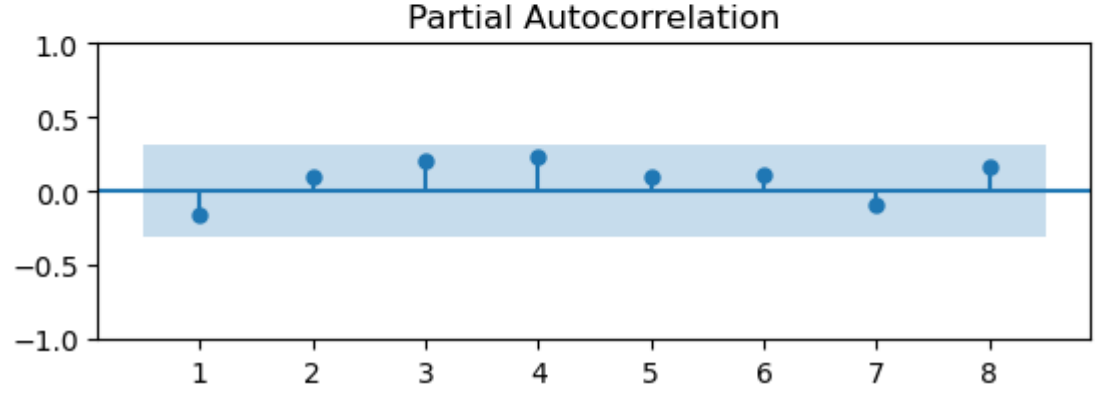
Tests for PC1.

```
Shapiro-Wilk test p-value = 0.098
```




```
Runs test p-value = 0.873
```

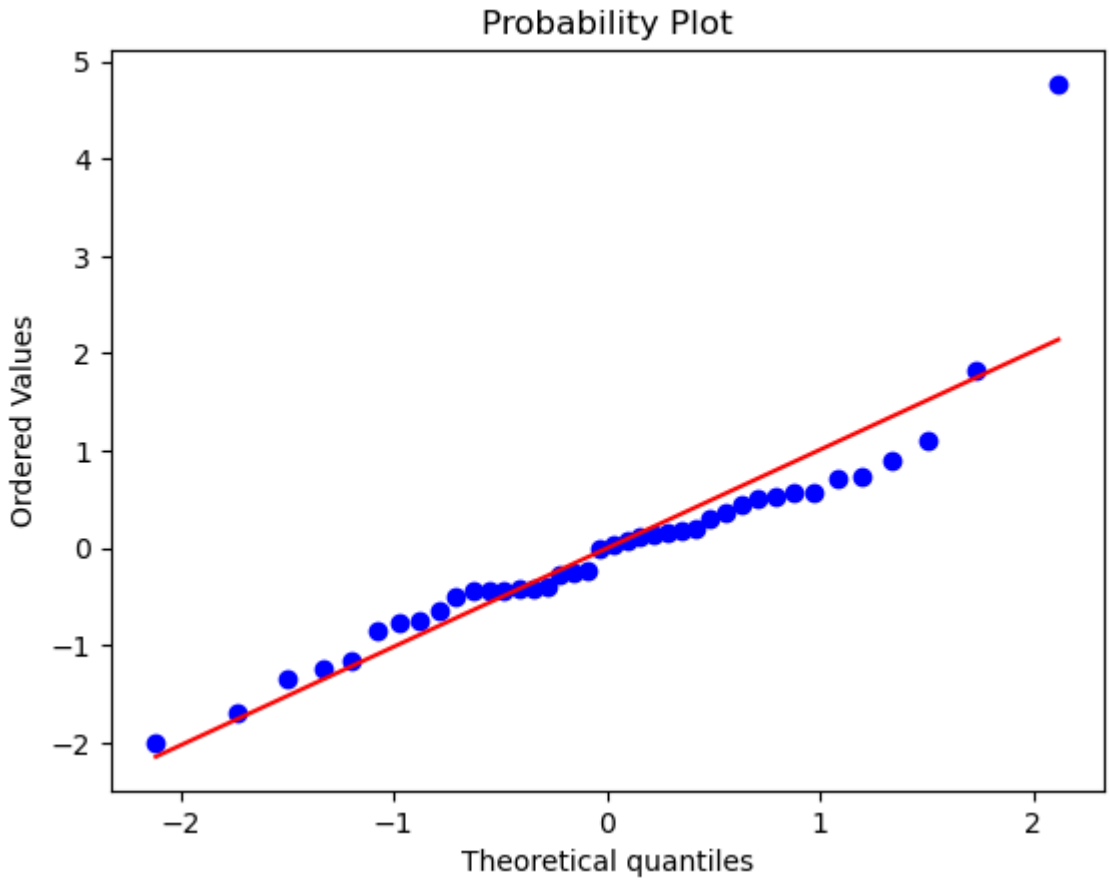




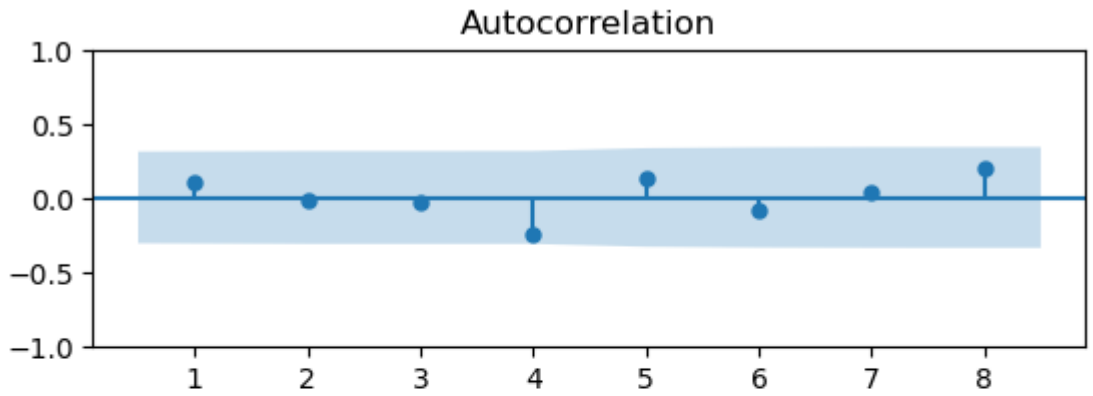
PC1 is normal and independent.

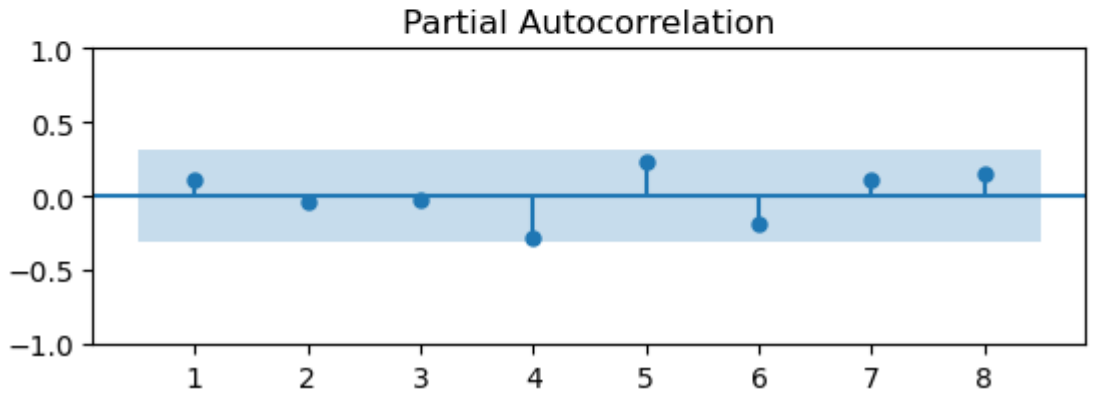
Tests for PC2:

```
Shapiro-Wilk test p-value = 0.000
```



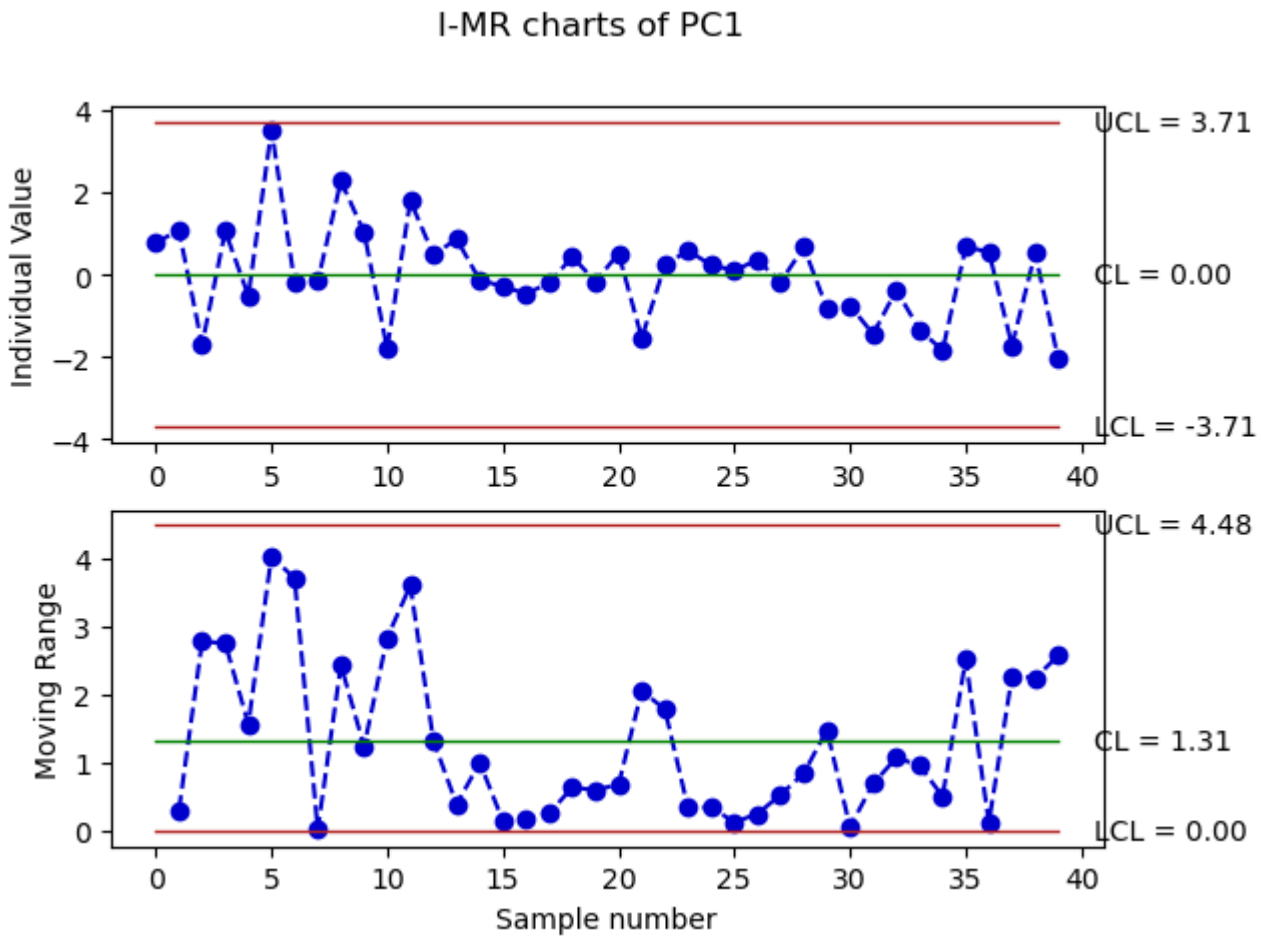

```
Runs test p-value = 0.631
```

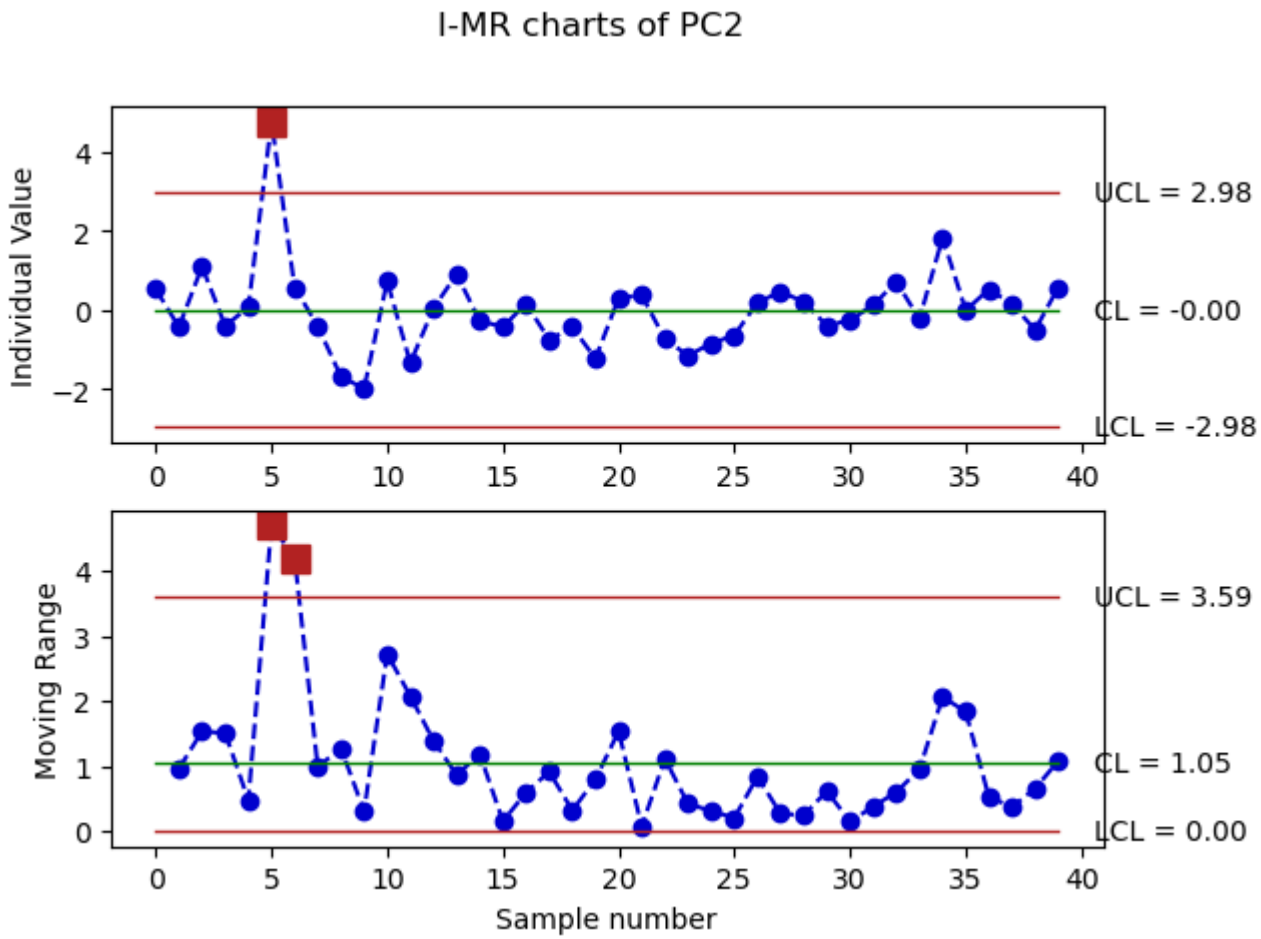




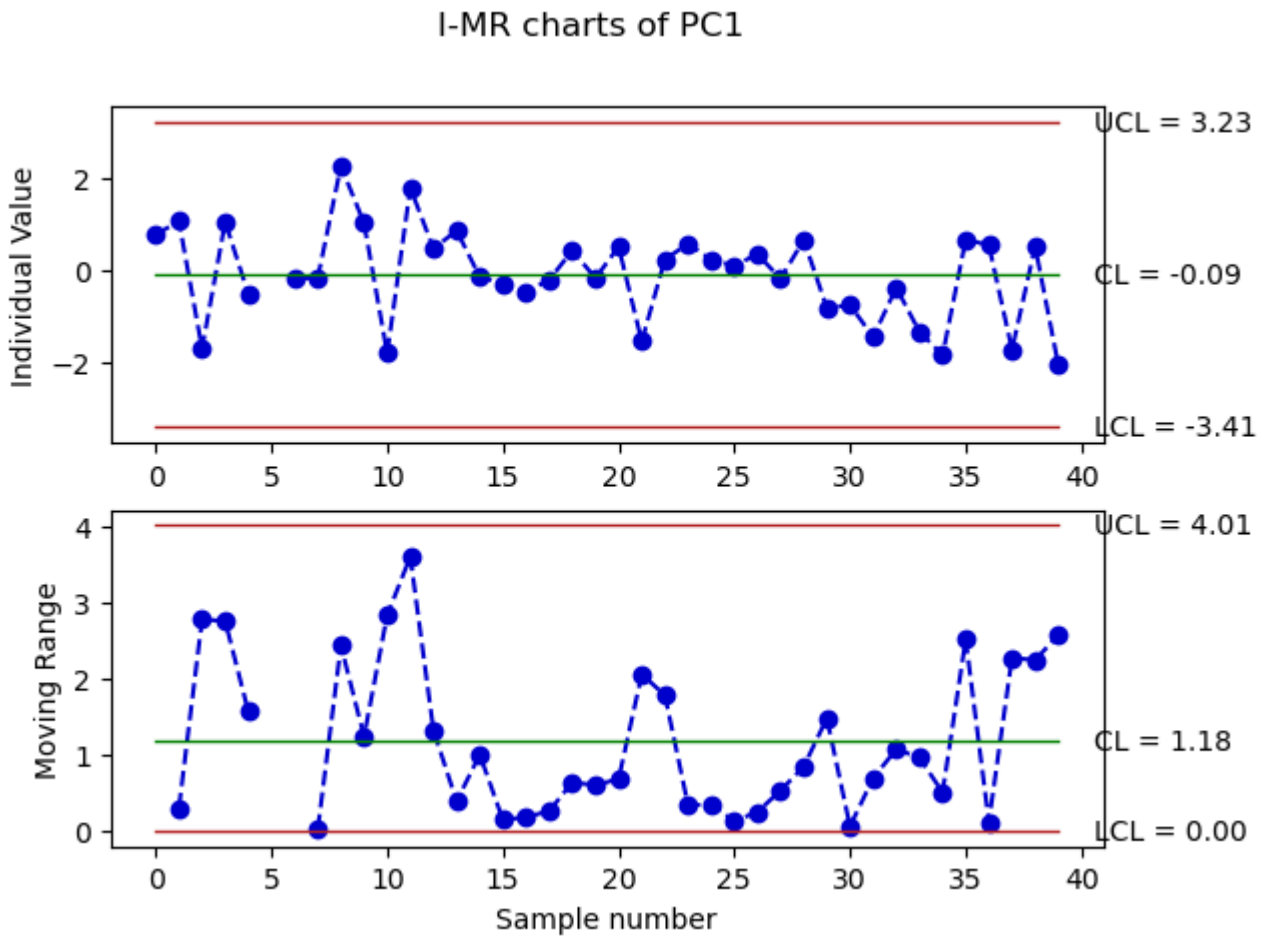
The second PC violates the normality assumption, but such violation is caused by the presence of the outlier. Indeed, it can be easily verified that removing the outlier, normality is met. Since this is the only violation of assumptions, we may design the I-MR control charts for the two PCs and verify whether they signal any alarm.

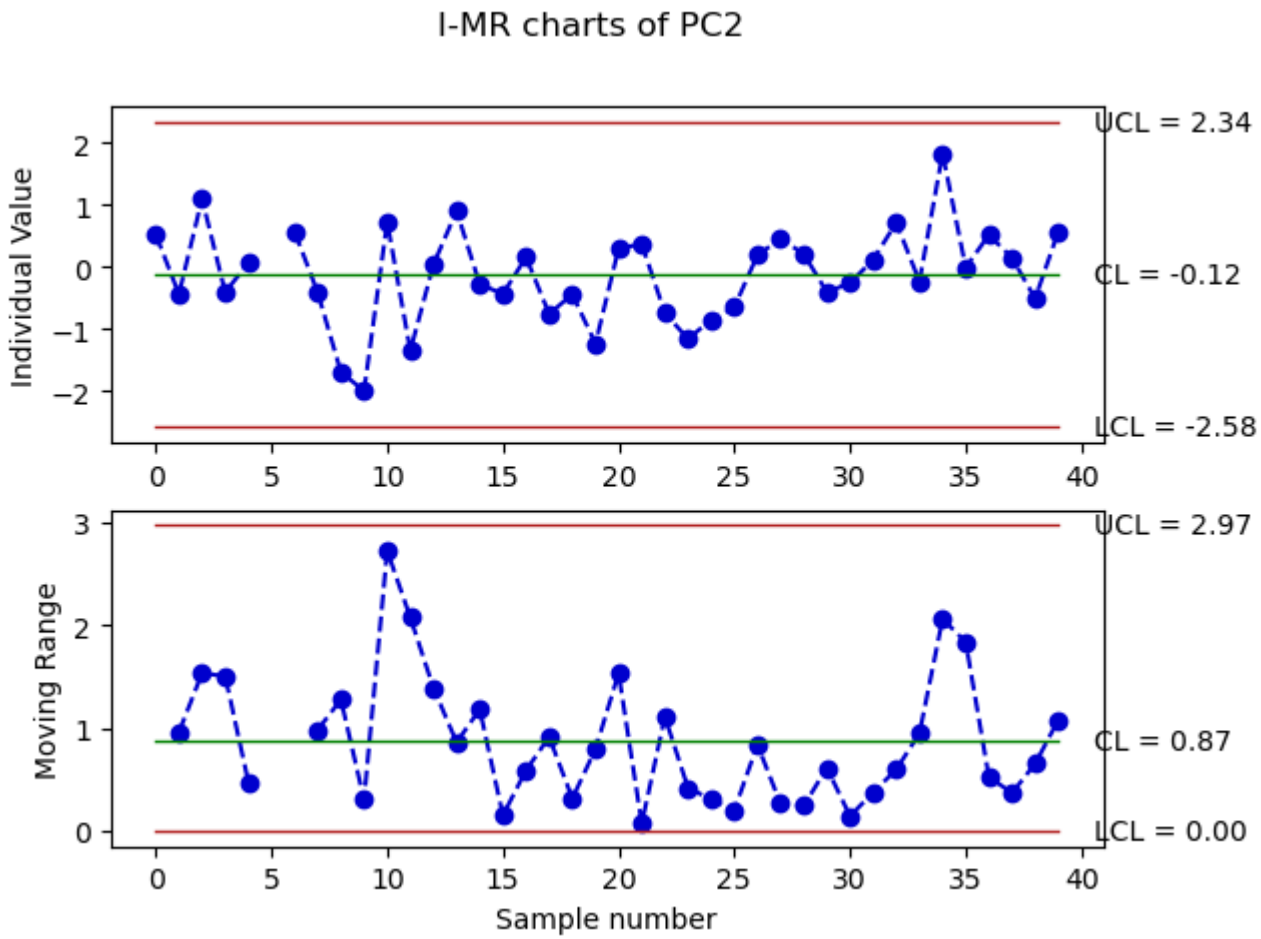
Design the I-MR control charts. Since we have two independent control variables we shall use the family-wise correction for independent data, which leads to K = 3.189.


I-MR charts of PC1
Individual Value
UCL = 3.71
CL = 0.00
LCL = -3.71
Moving Range
UCL = 4.48
CL = 1.31
LCL = 0.00
Sample number



The control chart for PC1 exhibits no violation of limits, although some hugging effect is possibly present in the central portion of the chart. The control chart for PC2 signals an alarm corresponding to the outlier discussed above. This confirms the presence of a contamination within the Phase I dataset. Assuming the existence of an assignable cause, we can remove the out-of-control observation and re-estimate the control limits. The new control charts are the following.


I-MR charts of PC1
Individual Value
UCL = 3.23
CL = -0.09
LCL = -3.41
Moving Range
UCL = 4.01
CL = 1.18
LCL = 0.00
Sample number



No other violation of control limits is found. The process is in-control. The design phase is over.

4)

We need to compute the probability of β for the given shift expressed in standard deviation units.

Since we are considering an I chart, we are testing the null hypothesis H_0 that X is normally distributed with mean μ_0 and variance σ^2.

$$H_0 : X \sim N(\mu_0, \sigma^2)$$

The alternative hypothesis is:

$$H_1 : X \sim N(\mu_1, \sigma^2)$$

So β is the probability

$$\beta = P(LCL \leq X \leq UCL | H_1)$$

If we define:

$$\delta = (\mu_1 - \mu_0)/\sigma$$

We can estimate β as:

$$\beta = P(Z \leq K - \delta) - P(Z \leq -K - \delta)$$

In this case we have:

$$\delta = 2.5$$

$$K = z_{\alpha/2}$$

Where:

$$\alpha = 1/350$$

Thus K = 2.983, and $\beta = 0.685$.

Exercise 3) Solutiond

Question 1)

Answer: a
Explanation: It is known (easy to show) that:
$E(aX) = aE(X)$, $Cov(aX, bY) = abCov(X, Y)$, $V(aX) = a^2E(X)$
Then using the above formulas in the definition of the autocorrelation we have:

$$\rho_1^* = \frac{\gamma_1^*}{\gamma_0^*} = \frac{Cov(X_t^*, X_{t-1}^*)}{V(X_t^*)} = \frac{Cov(c * X_t, c * X_{t-1})}{V(c * X_t)} = \frac{c^2 Cov(X_t, X_{t-1})}{c^2 V(X_t)} = \rho_1$$

Question 2)

Answer: d
Explanation: The type I error is α and since we will decrease it from 0.05 to 0.01, (a) is valid. We also know that as type I error, α, decreases, then the type II error, β, will increase and given that $\text{power} = 1 - \beta$, we will have that power will also decrease, so (b) is correct as well. We also know that $ARL_0 = 1/a$ and since α decreases ARL_0 will increase, i.e. (c) is also valid. Finally, as $ARL_1 = 1/(1 - \beta)$ and since β increases, we will get that ARL_1 will also increase, i.e., statement (d) is not valid.