QUALITY DATA ANALYSIS

06/02/2023

General recommendations:

- For exams in presence: to access the software on the provided laptops, go on browser → Favourites → Managed favourites → Virtual Desktop and enter your Polimi credentials.
- 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 10min

Exercise 1 (15 points)

In a Coca Cola plant, the head of the quality department has implemented a quality inspection system to keep under control the level of liquid within the bottle along the production line. The measurement device determines the deviation (in mm) from a target level. The measurements collected for 40 consecutive measurements are shown in Table 1.

Table 1

Sample	X	Sample	X	Sample	X	Sample	X
1	0,00	11	-0,08	21	-0,11	31	-0,17
2	0,03	12	-0,05	22	-0,26	32	-0,21
3	0,05	13	0,04	23	-0,17	33	-0,02
4	0,23	14	0,33	24	-0,25	34	-0,19
5	0,07	15	0,15	25	-0,16	35	0,00
6	0,19	16	0,21	26	-0,24	36	-0,04
7	0,23	17	0,15	27	-0,13	37	0,16
8	0,17	18	-0,08	28	-0,05	38	-0,14
9	0,19	19	-0,08	29	0,15	39	-0,07
10	0,02	20	-0,3	30	0,07	40	-0,24

a) Fit a suitable model to the data in Table 1
b) Based on the result of point a), design a suitable control chart and determine if the process is in-control or not (use K = 3). Discuss the result.
c) The head of the quality department decides to test a different control charting method, which consists of batching the data with a batch size equal to 2. Design a suitable control chart based on this approach (with K = 3).
d) After some tests with different batch sizes on a more extended dataset, the quality department has finally found a way to get rid of the temporal dependence in the measurements. After the batching operation, the data are normal and independent, with mean $\mu = 0$ and standard deviation σ. In the presence of an out-of-control mean shift $\delta\sigma$, what is the minimum value of δ that can be detected with a power of 90%? (use K = 3).

Exercise 2 (15 points)

In a plant that produces thrusters for satellites to be used in a low Earth orbit constellation of satellites, the performance indexes are measured during fire tests. The measured values for 20 consecutive tests are reported in Table 2. The project manager is interested in testing different process monitoring methods to determine if the performance indexes are in control or not.

Table 2

X1	X2
3,94	177,68
4,51	614
5,14	1380,22
4,69	1422,26
5,32	1176,15
5,42	4536,9
4,02	354,25
4,81	1176,15
2,76	275,89
6,2	3102,61
4,34	1164,45
3,74	487,85
4,73	518,01
3,5	93,69
4,98	3827,63
4,41	450,34
7,28	7631,2
6,19	4272,69
7,02	138,38
5,46	862,64

a) Design two traditional univariate control charts for data in Table 2 with a familywise ARL0 = 250.
b) Design a multivariate control chart to monitor the same data, with the same ARL0 used in point a). Compare this control chart with the ones designed at point a) and discuss the result.
c) The project manager wants to evaluate a third approach. He wants to apply the PCA to performance index values in Table 2 and monitor the first PC. Is it better to use the sample variance-covariance matrix or the correlation matrix to estimate the PCA model? Motivate the answer and apply the PCA (report the eigenvalues and eigenvectors, as well as the percentage of variance explained by the first PC).
d) Based on the result of point c), design a control chart on the first PC and compare the result with the ones obtained in point a) and b) (using the same ARL0 adopted in previous points). Discuss the result.

Exercise 3 (3 points)

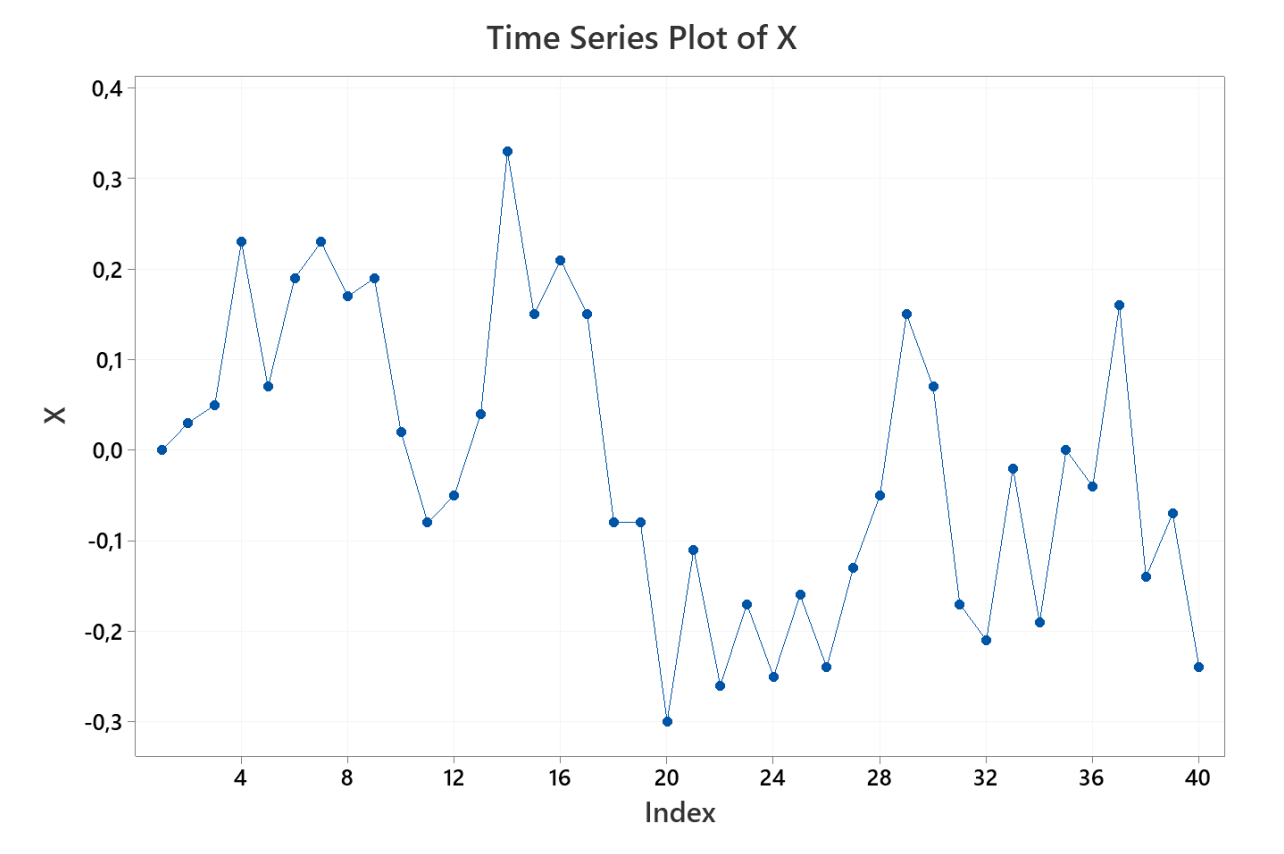
A manufacturing process is modelled by means of an AR(2) model. In case of a sudden shift of the process mean with entity $\delta\sigma_x$ (where σ_x is the standard deviation of the process data), what is the new average of the model residuals? Express the result as a function of σ_ε and of model parameters ϕ_1 and ϕ_2.

Solutions

Exercise 1

a)

The time series plot reveals a meandering pattern.



The runs test confirms the non randomness of the data.

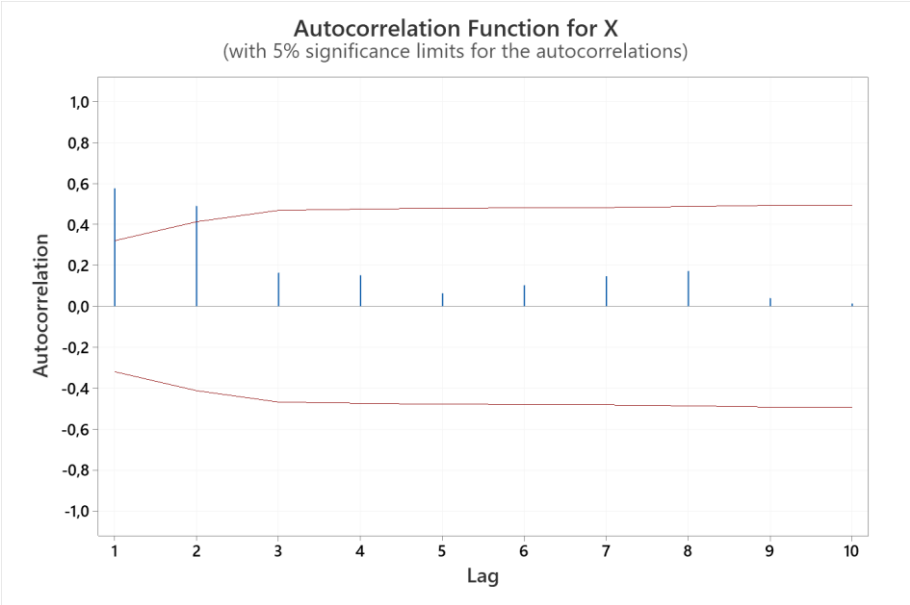
Test

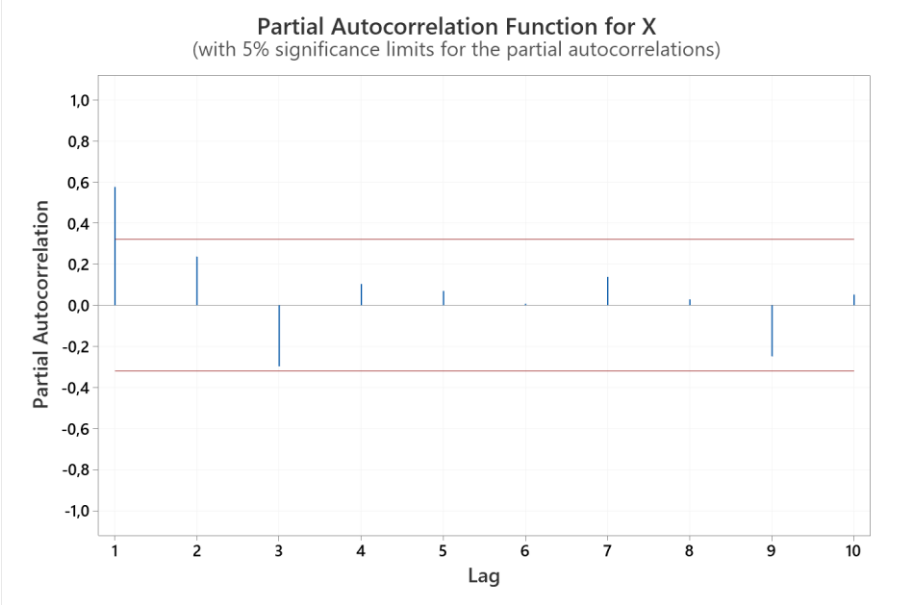
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs

Observed	Expected	P-Value
10	20,95	0,000

Sample ACF and PACF:





Based on the sample ACF and PACF, a suitable model may be an MA(2).

EXE1_VERSION1

ARIMA Model: X

Estimates at Each Iteration

Iteration	SSE	Parameters		
0	1,92521	0,100	0,100	0,085
1	1,40015	-0,050	0,071	0,076
2	0,96952	-0,196	-0,079	0,063
3	0,77509	-0,297	-0,229	0,050
4	0,65996	-0,398	-0,379	0,033
5	0,59991	-0,516	-0,529	0,003
6	0,59360	-0,561	-0,548	-0,015
7	0,59332	-0,575	-0,557	-0,017
8	0,59328	-0,581	-0,559	-0,017
9	0,59327	-0,583	-0,560	-0,017
10	0,59327	-0,584	-0,560	-0,017
11	0,59327	-0,584	-0,561	-0,017

Relative change in each estimate less than 0,001

Final Estimates of Parameters

Type		Coef	SE Coef	T-Value	P-Value
MA	1	-0,584	0,138	-4,23	0,000
MA	2	-0,561	0,142	-3,96	0,000
Constant		-0,0167	0,0424	-0,39	0,696
Mean		-0,0167	0,0424		

Number of observations: 40

Residual Sums of Squares

DF	SS	MS
37	0,591601	0,0159892

Back forecasts excluded

The constant term is not significant, thus we may remove it and refit the model.

EXE1_VERSION1

ARIMA Model: X

Estimates at Each Iteration

Iteration	SSE	Parameters	
0	1,33580	0,100	0,100
1	1,07098	-0,050	0,076
2	0,82339	-0,198	-0,074
3	0,70279	-0,296	-0,224
4	0,63161	-0,396	-0,374
5	0,59936	-0,510	-0,524
6	0,59620	-0,561	-0,551

Unable to reduce sum of squares any further

Final Estimates of Parameters

Type		Coef	SE Coef	T-Value	P-Value
MA	1	-0,561	0,138	-4,07	0,000
MA	2	-0,551	0,141	-3,91	0,000

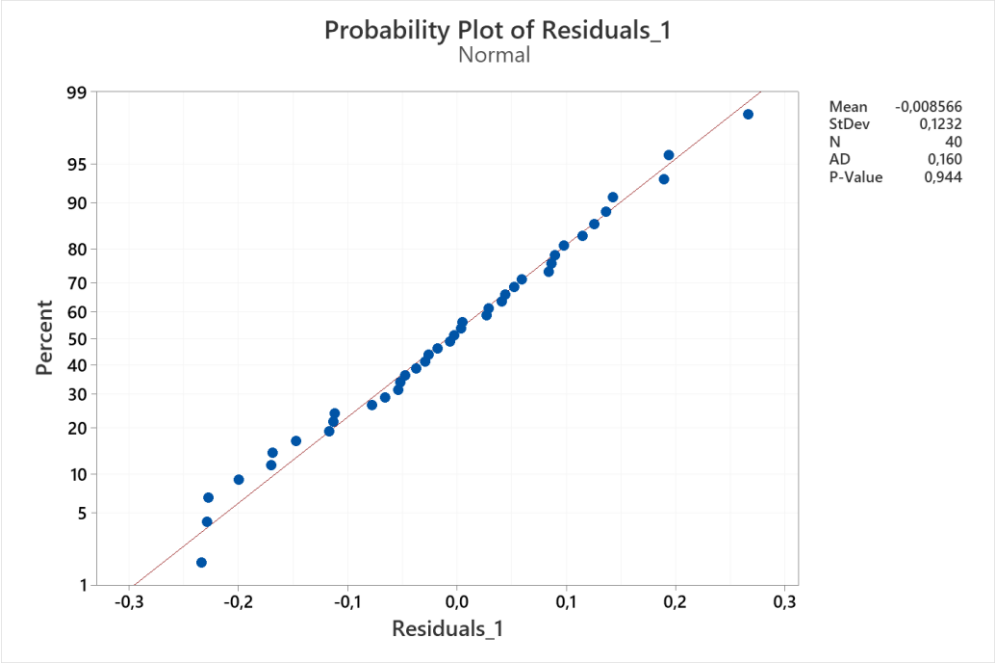
Number of observations: 40

Residual Sums of Squares

DF	SS	MS
38	0,594809	0,0156529

Back forecasts excluded

The residuals are normal and independent, thus the model is appropriate.

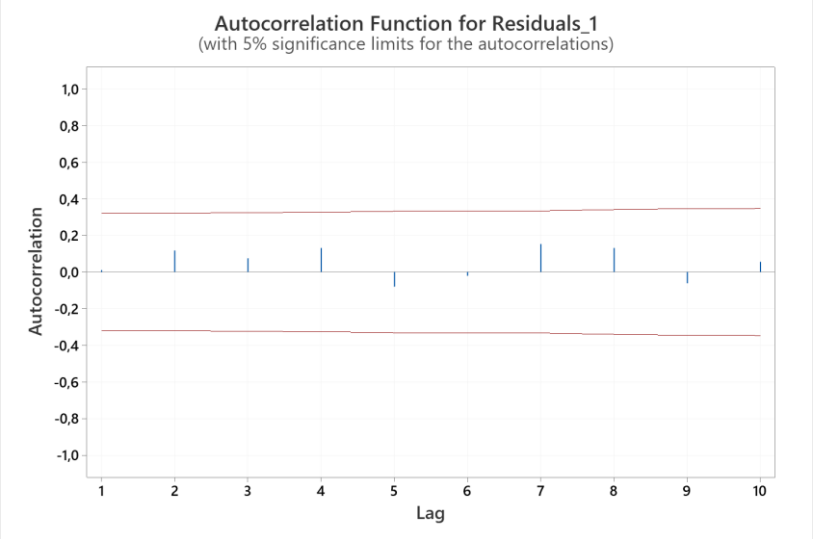


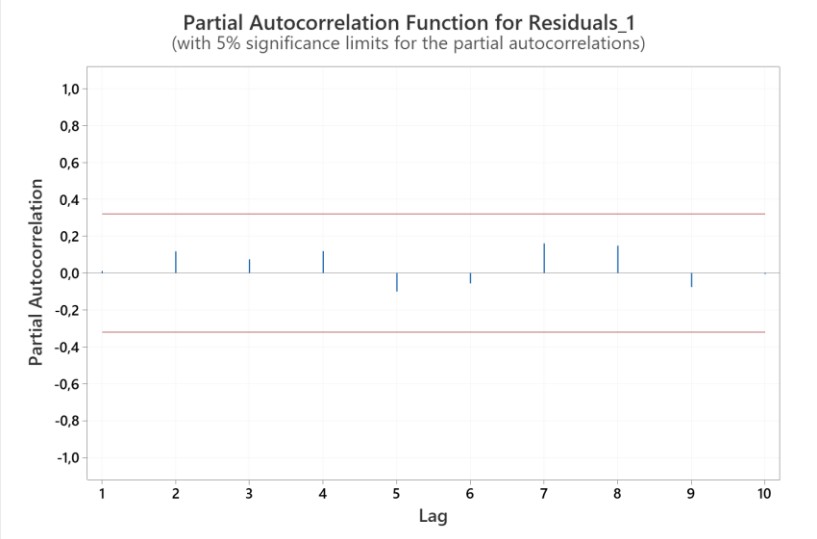
Test

Null hypothesis H_0: The order of the data is random
Alternative hypothesis H_1: The order of the data is not random

Number of Runs

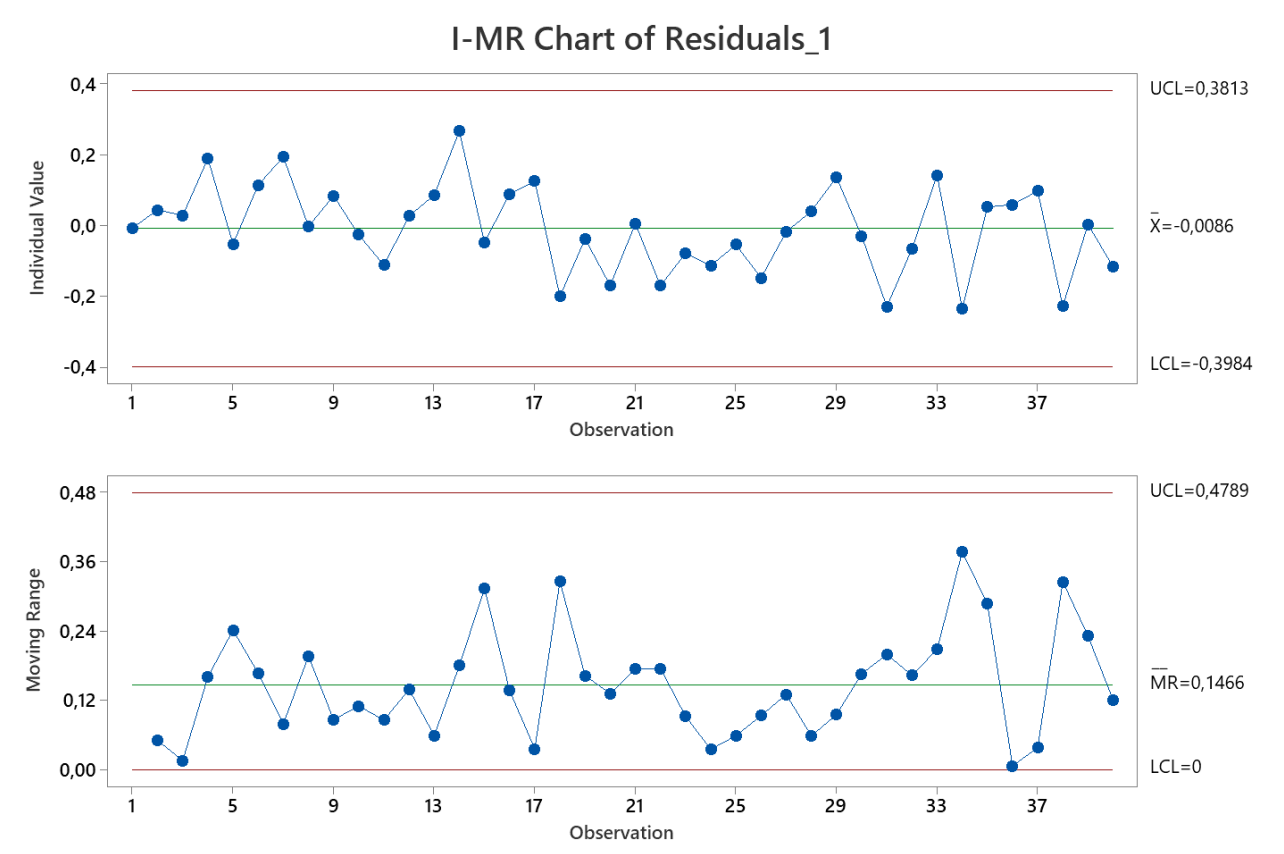
Observed	Expected	P-Value
18	20,95	0,343





b)

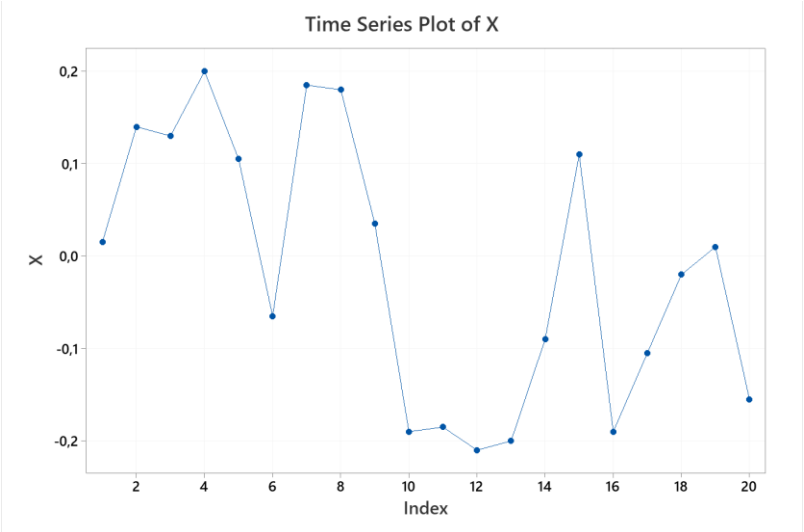
The control chart on the model residuals is the following:



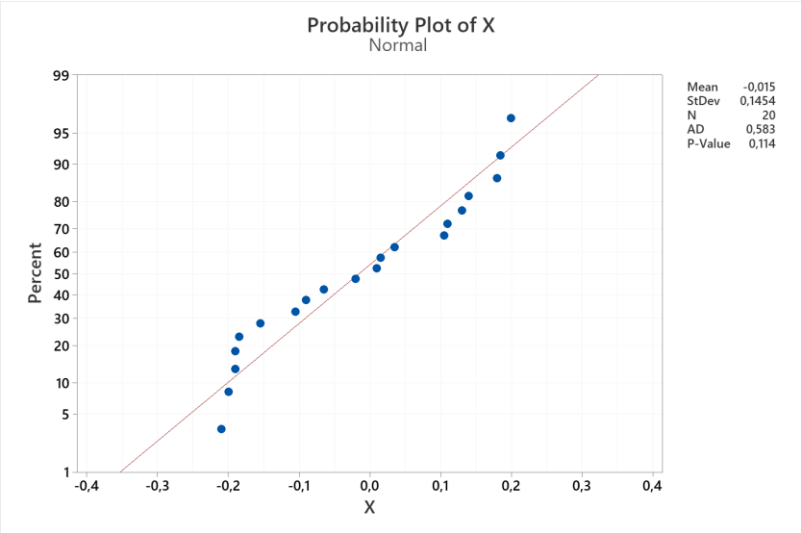
No violation of the limit is signalled, although a small shift may have occurred, since the residuals of the first half of samples are predominantly above the center line in the I chart, whereas the second half is predominantly below. However, this shift is too small to generate an alarm in Shewhart control charts. Different types of control charts are specifically thought to enhance the capability of detecting small shifts. They are called “time-weighted control charts”.

c)

After batching, we have the following time series:



Check of assumptions:



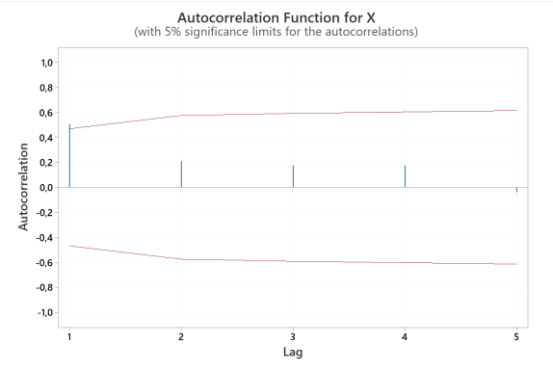
Test

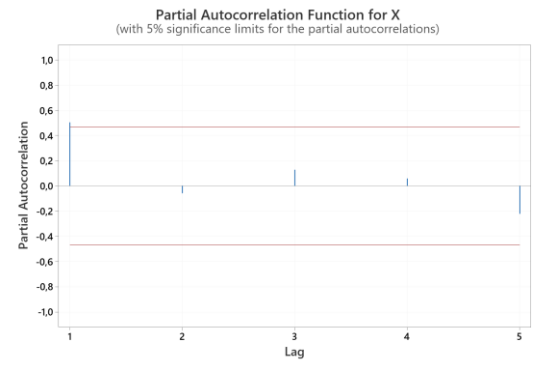
Null hypothesis H_0: The order of the data is random
Alternative hypothesis H_1: The order of the data is not random

Number of Runs

Observed	Expected	P-Value
8	11,00	0,168

The p-value may not be accurate for samples with fewer than 11 observations above K or fewer than 11 below.





Bartlett tets at lag = 1 (95% confidence):

$$|r_k| = 0.5037$$

$$\frac{z_{\alpha/2}}{\sqrt{n}} = 0.438$$

The autocorrelation at lag 1 is significant.
Therefore, we can fit an AR(1) model (also in this case the constant term is not significant and hence we can remove it):

BATCHING

Regression Analysis: X versus AR1

Method

Rows unused 1

Regression Equation

X = 0,536 AR1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
AR1	0,536	0,207	2,59	0,019	1,00

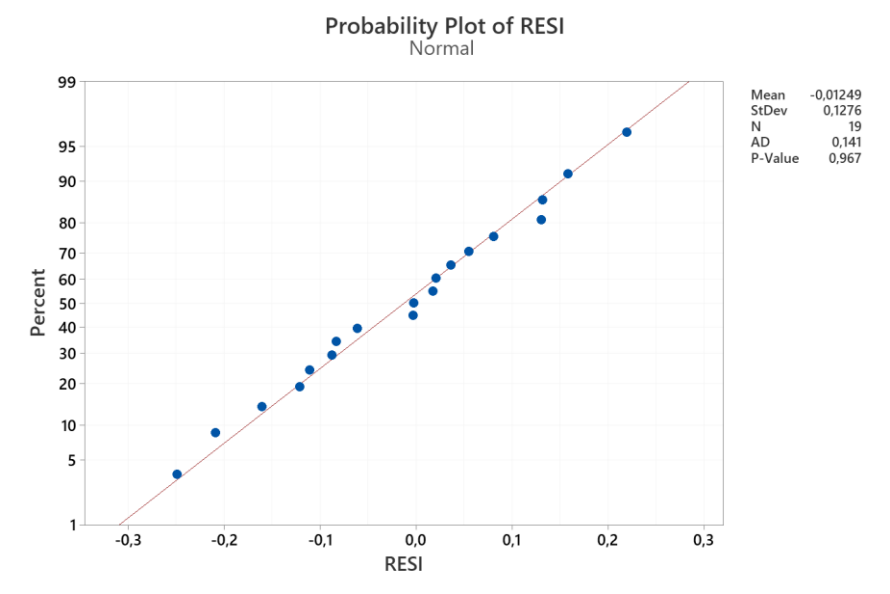
Model Summary

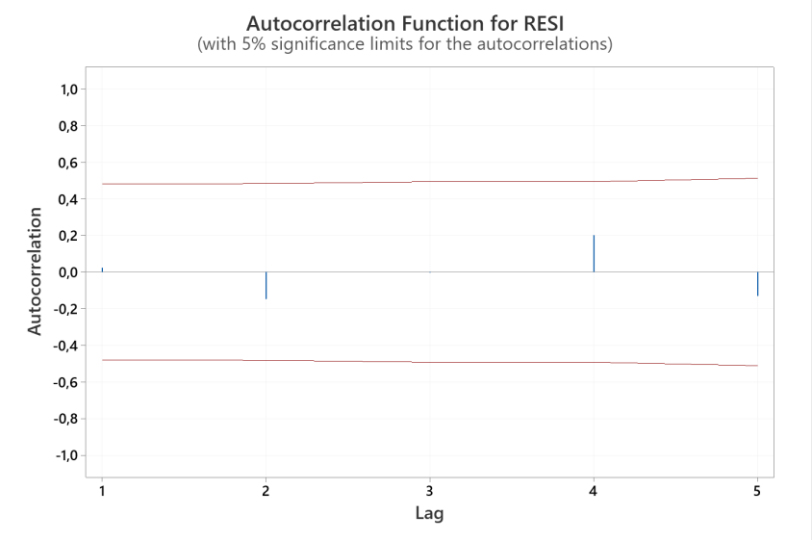
S	R-sq	R-sq(adj)	R-sq(pred)
0,128234	27,07%	23,02%	22,67%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0,109884	0,109884	6,68	0,019
AR1	1	0,109884	0,109884	6,68	0,019
Error	18	0,295991	0,016444		
Lack-of-Fit	17	0,292791	0,017223	5,38	0,328
Pure Error	1	0,003200	0,003200		
Total	19	0,405875			

Residuals are normal and independent:





Test

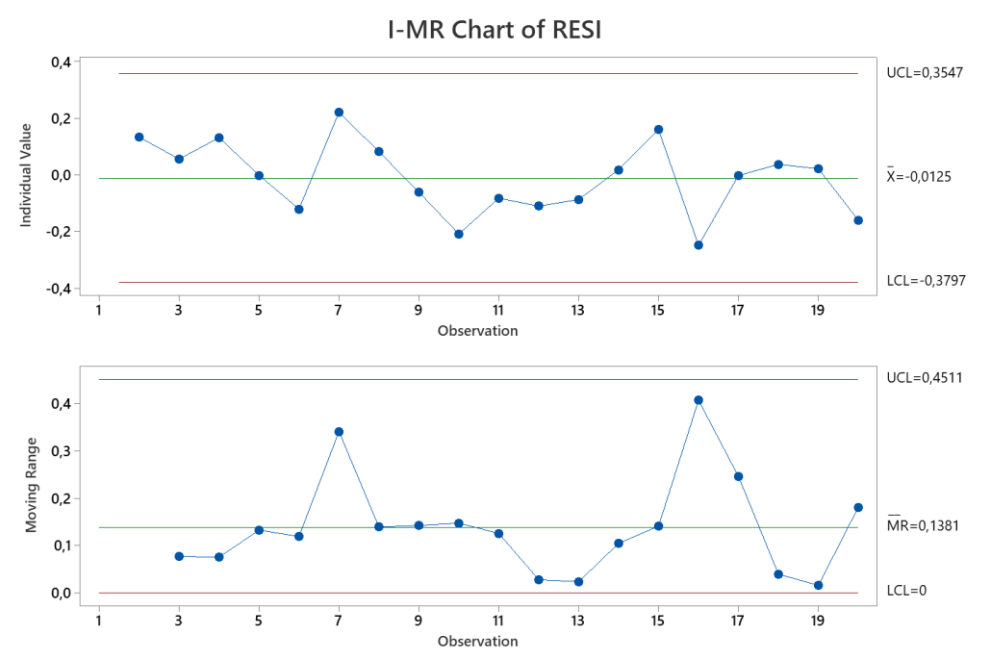
Null hypothesis H_0: The order of the data is random
Alternative hypothesis H_1: The order of the data is not random

Number of Runs

Observed	Expected	P-Value
8	10,26	0,272

The p-value may not be accurate for samples with fewer than 11 observations above K or fewer than 11 below.

The resulting control chart is the following:



The batching operation has not removed the temporal dependence. Using a larger batch size may be more effective to this aim, but a larger Phase I dataset would be needed. After batching, a possible small shift is still visible, but the control chart on model residuals is still not effective in revealing it.

d)

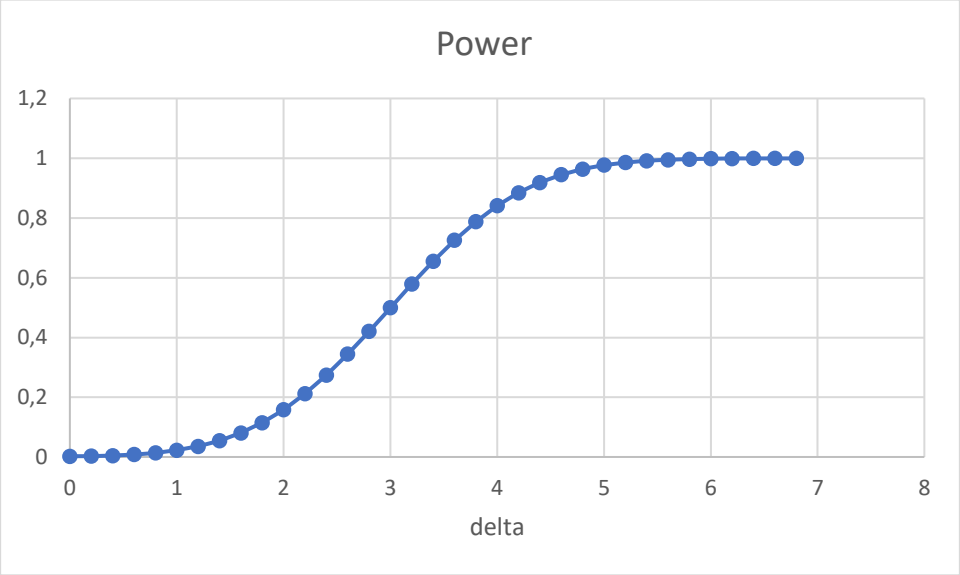
The type II error is:

$$\beta = P(x_t < UCL|H_1) - P(x_t < LCL|H_1) =$$

$$= P\left(\frac{x_t - \delta\sigma}{\sigma} < \frac{K\sigma - \delta\sigma}{\sigma}\right) - P\left(\frac{x_t - \delta\sigma}{\sigma} < \frac{-K\sigma - \delta\sigma}{\sigma}\right) =$$

$$= \Phi(K - \delta) - \Phi(-K - \delta)$$

The power $P(\delta) = 1 - \beta(\delta)$ is:

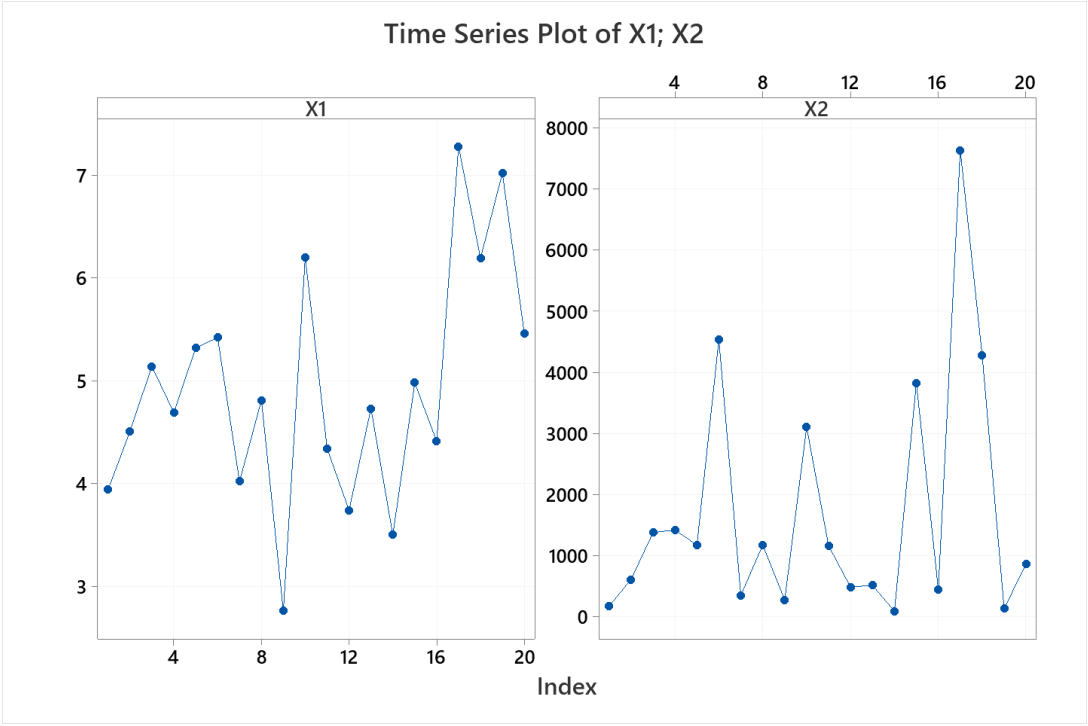


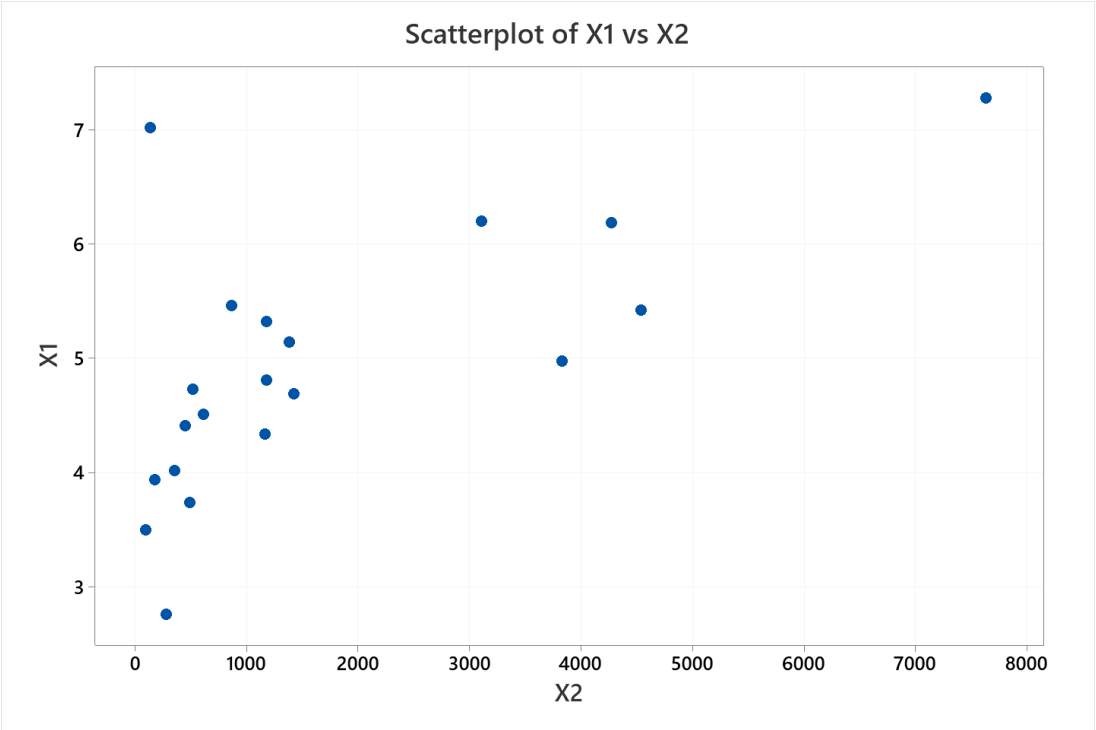
Therefore, the minimum value of δ that is detected with a power of 90% is $\delta = 4.4$.

Exercise 2

a)

Data snooping.





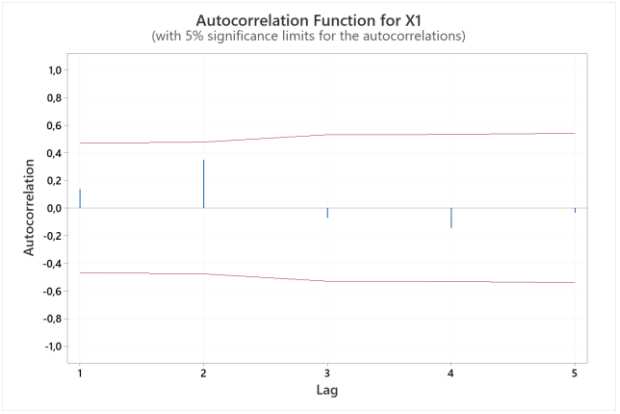
The second variable looks quite skewed. Check of assumptions:

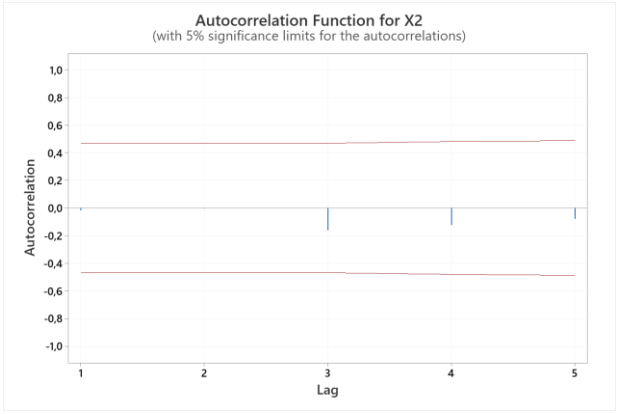
Test

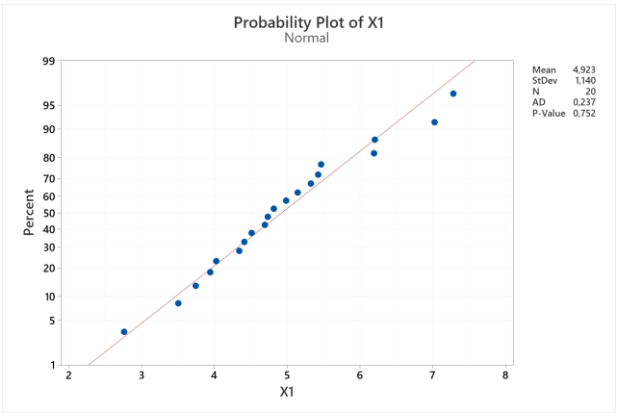
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

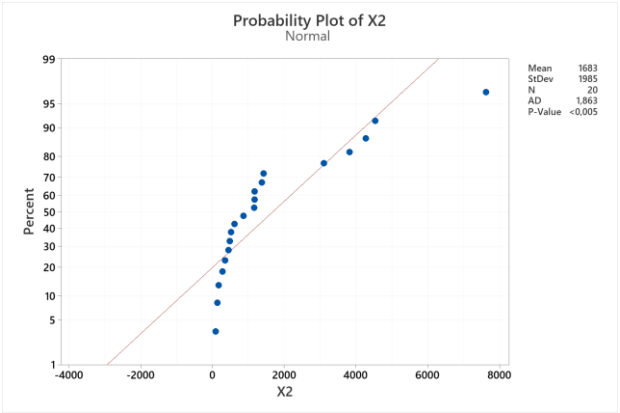
	Number of Runs		
Variable	Observed	Expected	P-Value
X1	10	10,90	0,676
X2	9	8,50	0,755

The p-value may not be accurate for samples with fewer than 11 observations above K or fewer than 11 below.

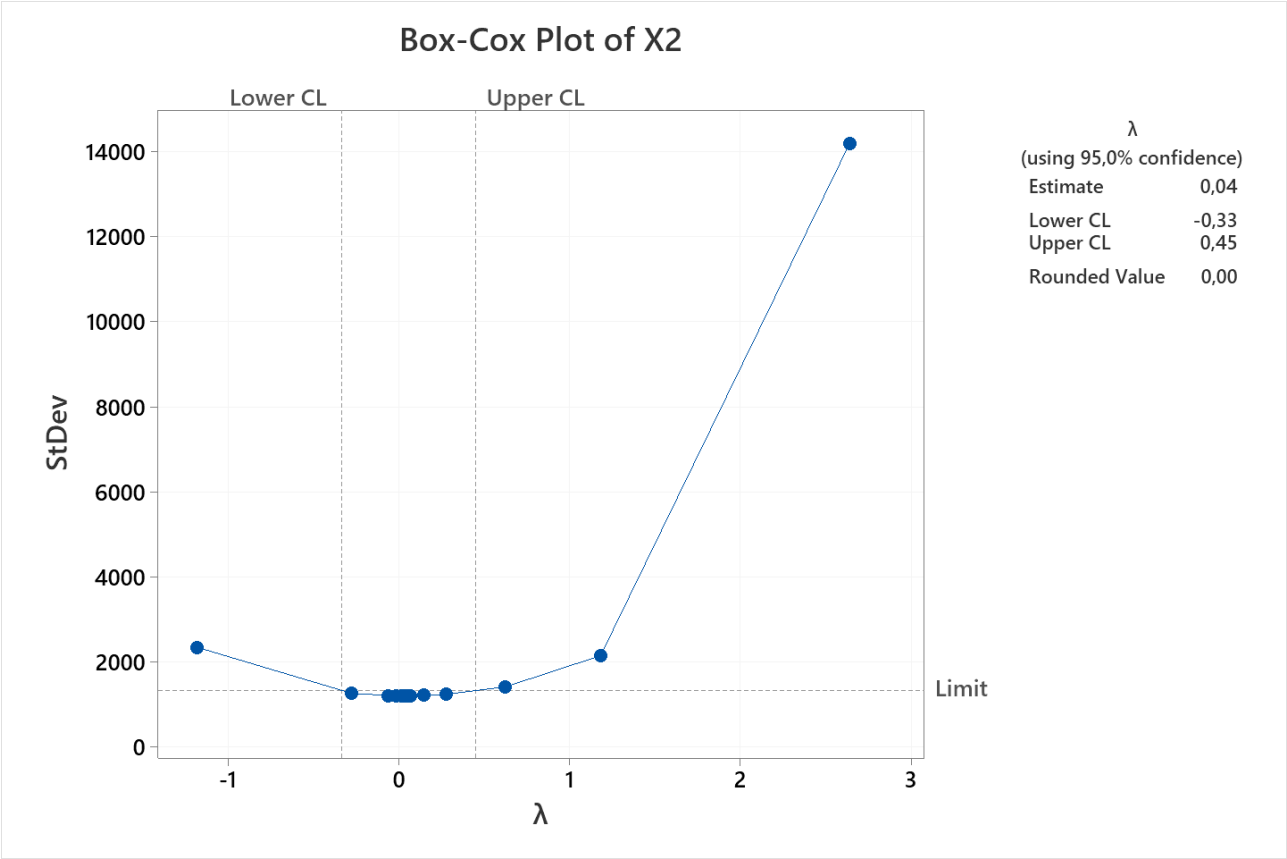




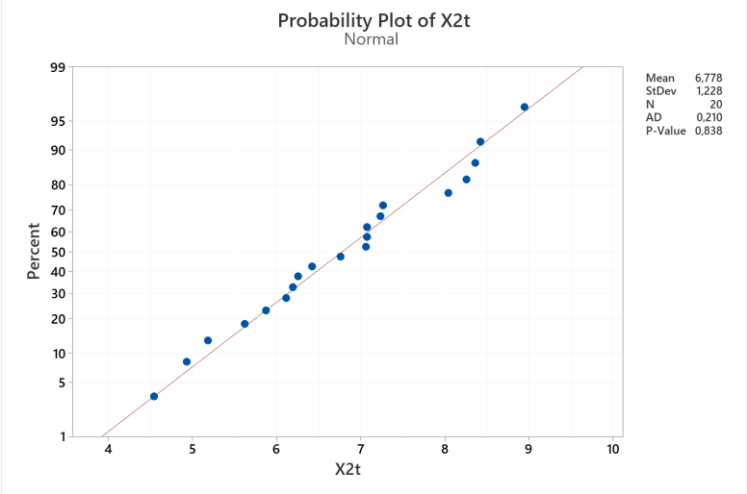




The first variable is normal and independent. The second variable meets the randomness assumption but it violates the normality assumption. We can try to apply the Box-Cox transformation.



The second variable can be transformed by applying a natural logarithm. Let's check normally after the transformation.



Randomness is still met too.

Test

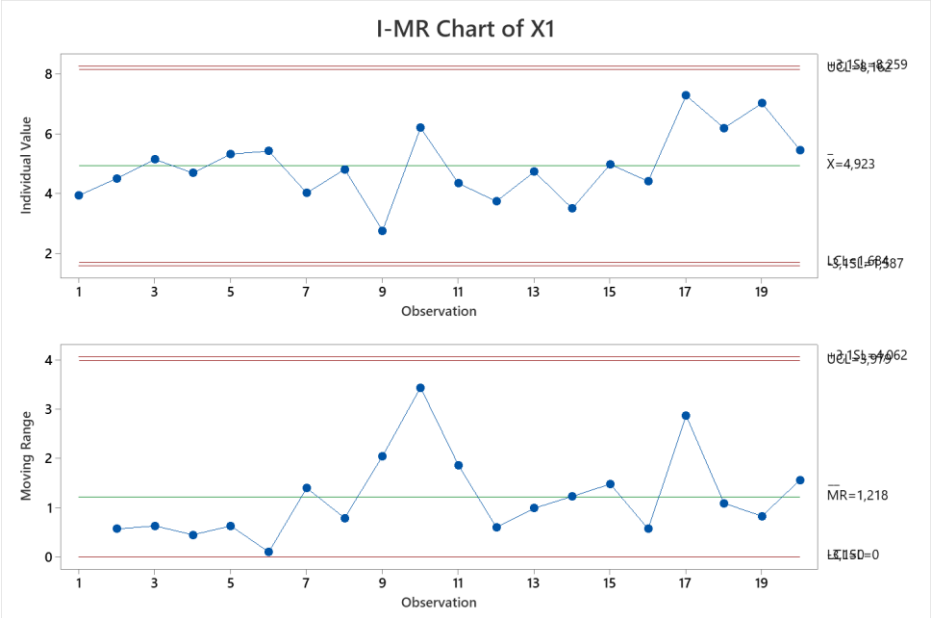
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

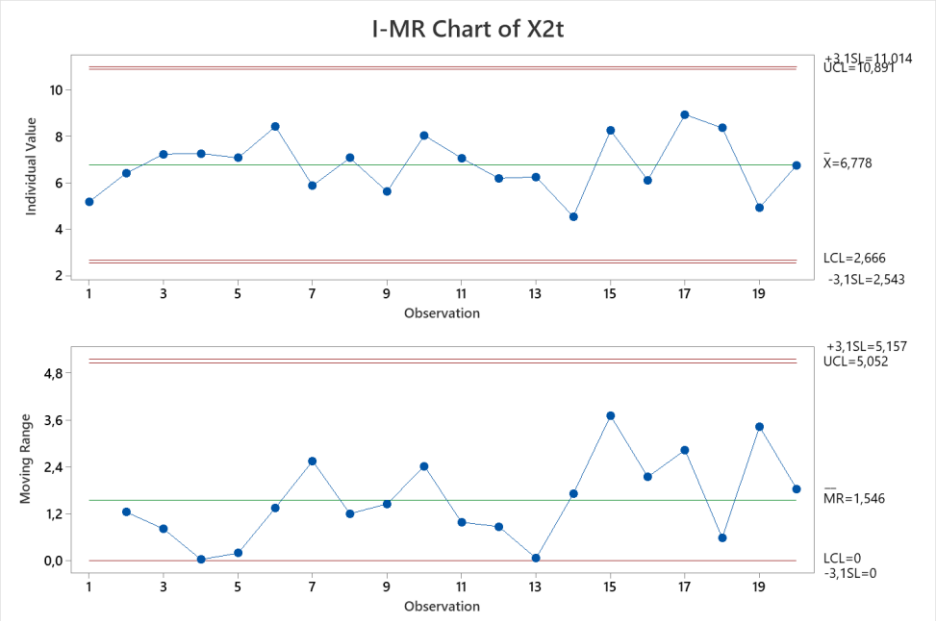
Number of Runs

Observed	Expected	P-Value
11	11,00	1,000

The p-value may not be accurate for samples with fewer than 11 observations above K or fewer than 11 below.

With familywise $ARL_0 = 250$, $z_{\alpha/2} = 3.09$. The I-MR control charts for model residuals are the following (do not consider the limits at k=3 in the figure).

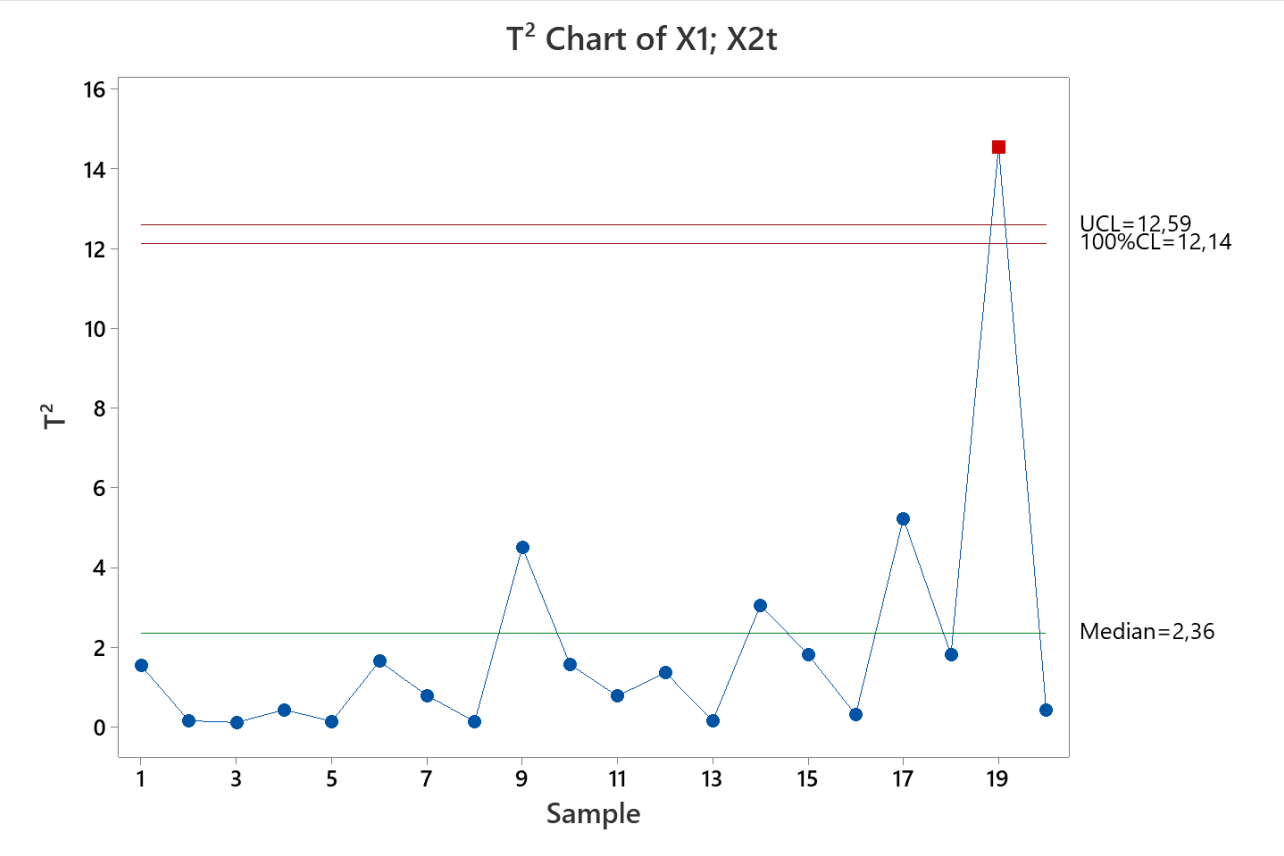




Apart from a small sustained shift in the last four observations of the first variable (which may possibly deserve some attention), no violation of the control limits is present.

b)

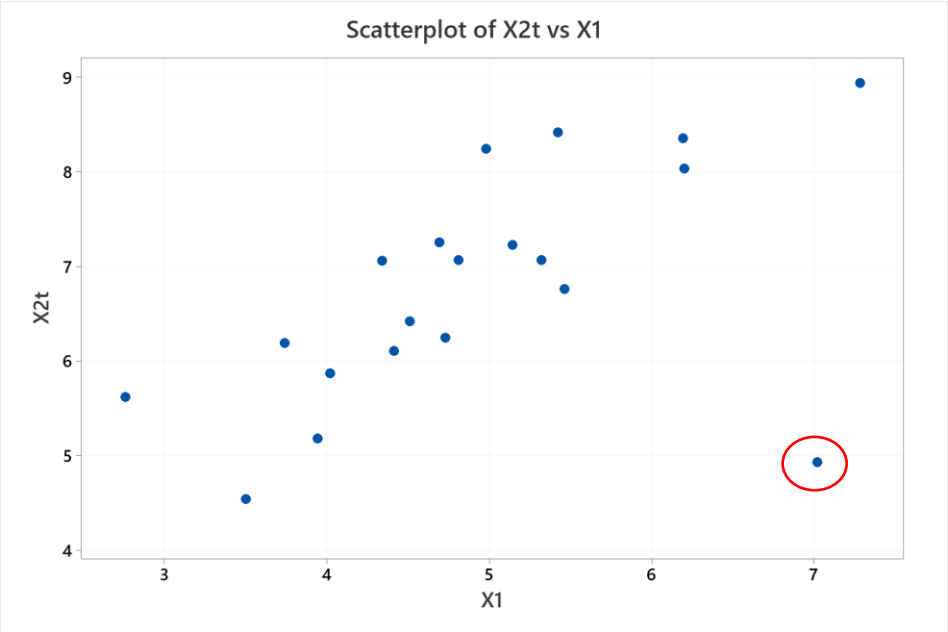
It is possible to design a T2 control chart with $ARL_0 = 250$, and hence $\alpha = 0.004$. The control chart is the following:



Differently from the two univariate control charts, a violation of the limit is present at sample 19.

By looking at the scatter plot of the two variables after the logarithm transformation on the second, it is possible to see that there is a positive correlation among them. The observation in sample 19 (highlighted in red) is quite far away from the bivariate scatter of all other observations. Since the control region of the T2 control chart corresponds to an ellipse in the bivariate space around the data, the T2 is effective in signalling that sample as anomalous with respect to the given dataset.

No alarm was raised by the univariate control charts because the values of the two variables in sample 19 are within the range of the data used to design the control chart. Using two univariate control charts implies a rectangular control region in the bivariate space, and sample 19 is inside that region.



c)

PCA does not require the normality assumptions, but the control chart to be designed on the first PC requires both normality and randomness assumption. Thus we may apply the PCA on the data after the logarithm transformation of the second variable. The two variables have a similar standard deviation, thus either using the sample variance-covariance matrix or the sample correlation matrix would be ok. Let's use the variance-covariance matrix.

Statistics

Variable	StDev
X1	1,140
X2t	1,228

The result of the PCA is the following:

EXE2 RIGHT

Principal Component Analysis: X1; X2t

Eigenanalysis of the Covariance Matrix

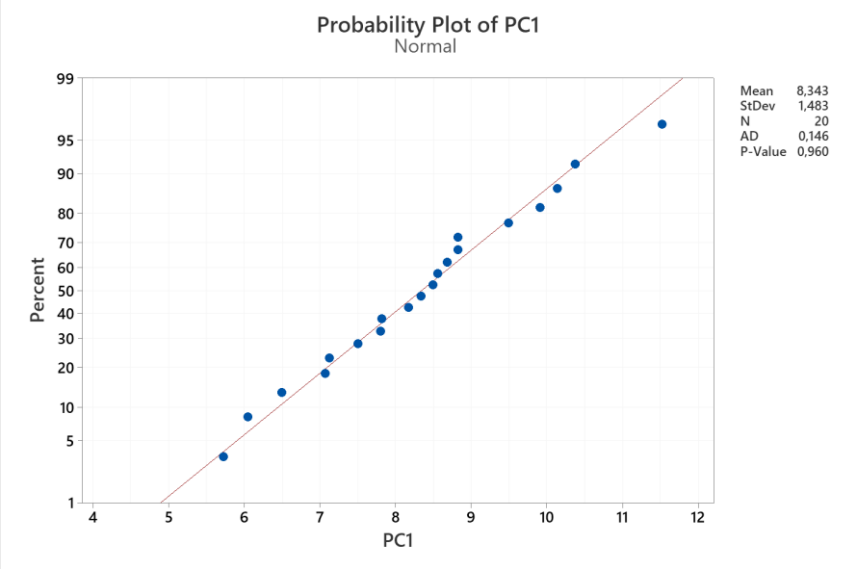
Eigenvalue	2,2004	0,6078
Proportion	0,784	0,216
Cumulative	0,784	1,000

Eigenvectors

Variable	PC1	PC2
X1	0,659	0,752
X2t	0,752	-0,659

The first PC explains about 78% of the overall variability. It associates similar weights to the two variables.

We may check the assumptions on the scores of the first PC before applying the chart.



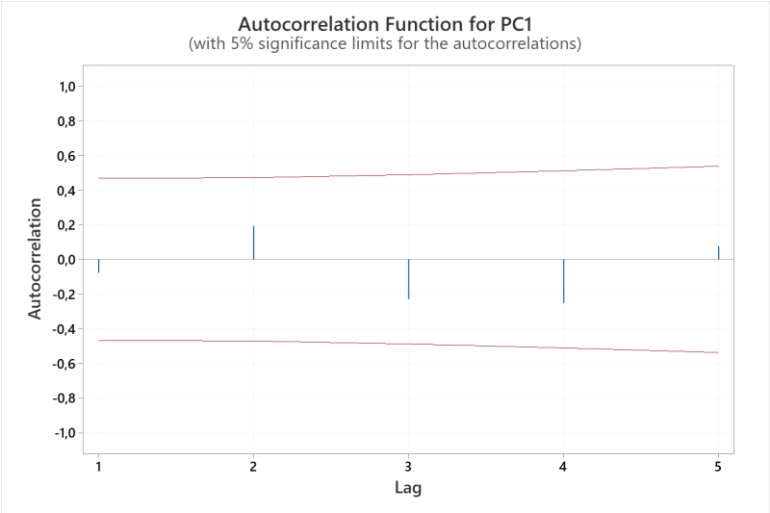
Test

Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

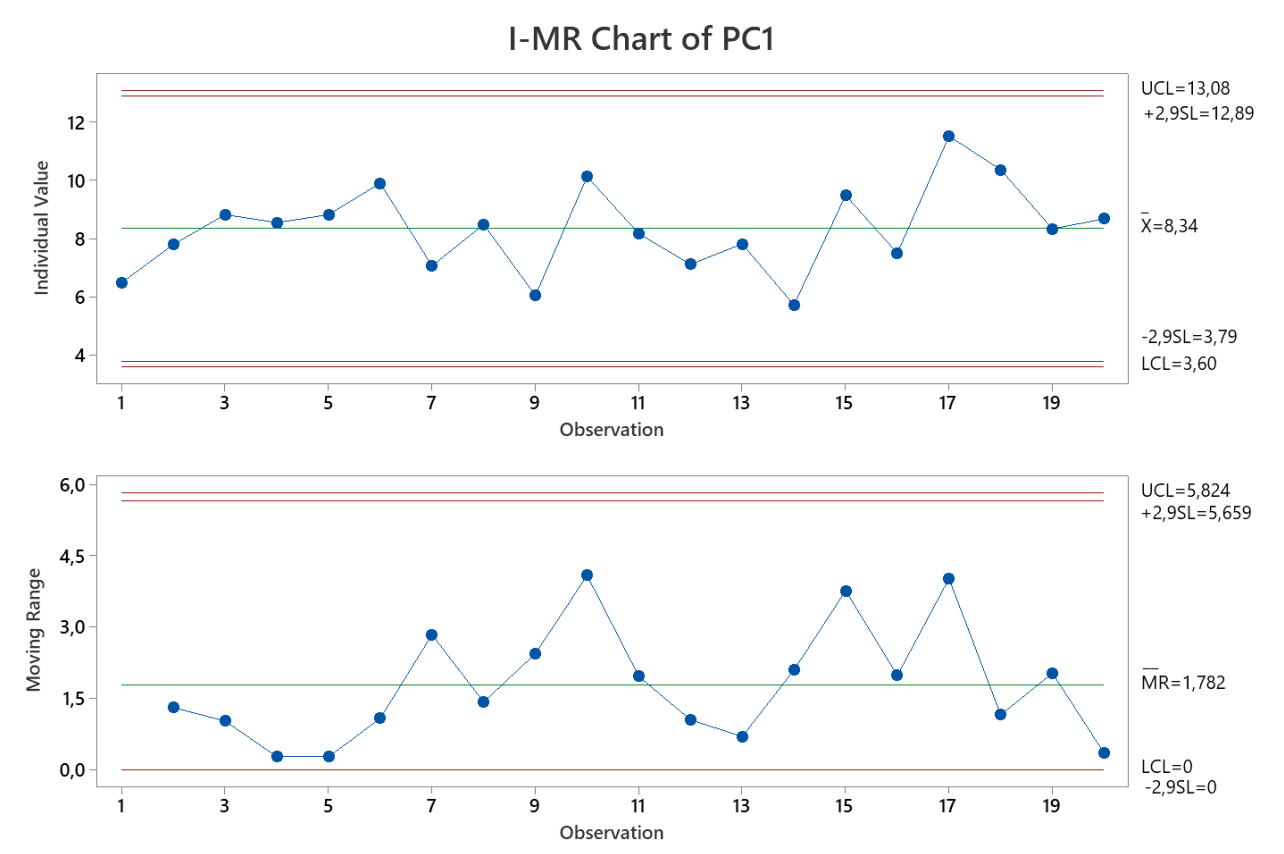
Number of Runs

Observed	Expected	P-Value
12	11,00	0,646

The p-value may not be accurate for samples with fewer than 11 observations above K or fewer than 11 below.

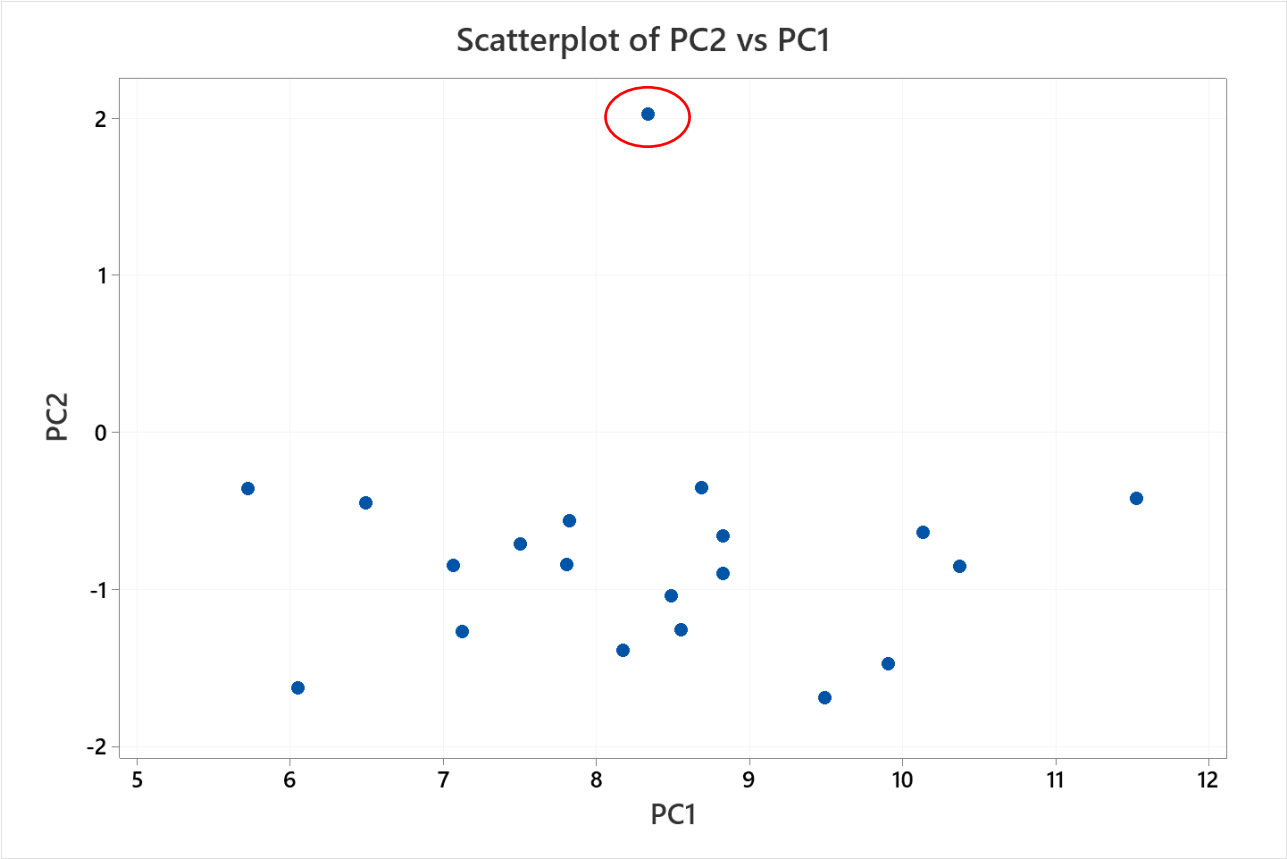


The I-MR control chart on the first PC with $ARL_0 = 250$, and hence $z_{\alpha/2} = 2.878$, is the following (ignore the limits at K=3):



No violation of control limits is present. The anomaly signalled by the T2 control chart is not signalled by monitoring the first PC. By looking at the scatterplot between PC1 and PC2, it is possible to see that the anomaly in sample 19 affects only PC2 (highlighted in red). PC2 is a contrast between the variables, and the anomaly actually affects this contrast. As shown above, the positive correlation between the two variables implies that high values of X1 correspond to high values of X2 and viceversa. In sample 19, instead, a high values of X1 corresponds to a low values of X2.

When monitoring a process in the PC space, it is a common practice to combine a control chart on first retained PCs with a control charts on the PCA model residuals, which is helpful to detect anomalies that do not affect the first PCs but only the remaining ones, preventing any information loss.



Exercise 3

For an AR(2) process, the following expression applies:

$$\tilde{X}_t = \phi_1 \tilde{X}_{t-1} + \phi_2 \tilde{X}_{t-2} + \varepsilon_t$$

Moreover:

$$\begin{cases} \rho_1 = \phi_1 + \phi_2 \rho_1 \\ \rho_2 = \phi_2 \rho_1 + \phi_2 \end{cases} \Rightarrow \begin{cases} \rho_1 = \frac{\phi_1}{1-\phi_2} \\ \rho_2 = \frac{\phi_1^2}{1-\phi_2} + \phi_2 \end{cases}.$$

In the presence of a shift with entity $\delta\sigma_X$we get:

$$\tilde{X}'_t = \tilde{X}_t + \delta\sigma_X$$

therefore:

$$\varepsilon'_t = \tilde{X}'_t - \phi_1 \tilde{X}'_{t-1} - \phi_2 \tilde{X}'_{t-2}$$

Reminding that $\sigma_X = \frac{\sigma_\varepsilon}{\sqrt{1-\phi_1\rho_1-\phi_2\rho_2}}$, the mean of ε'_t can then be computed as follows:

$$\mu_{\varepsilon'_t} = \delta\sigma_X - \phi_1\delta\sigma_X - \phi_2\delta\sigma_X = \delta(1-\phi_1-\phi_2)\frac{\sigma_\varepsilon}{\sqrt{1-\phi_1\rho_1-\phi_2\rho_2}} =$$

$$= \delta(1-\phi_1-\phi_2)\frac{\sigma_\varepsilon}{\sqrt{1-\frac{\phi_1^2}{1-\phi_2}-\frac{\phi_1^2-\phi_2^2+\phi_2}{1-\phi_2}\phi_2}}$$