: quelli che segno così sono quelli che faccio/ho fatto

QUALITY DATA ANALYSIS

08/07/2022

General recommendations:

a) write the solutions in CLEAR and READABLE way on paper and show (qualitatively) all the relevant plots;
b) avoid (if not required) theoretical introductions or explanations covered during the course;
c) always state the assumptions and report all relevant steps/discussion/formulas/expression to present and motivate your solution;
d) when using hypothesis tests provide the numerical value of the test statistic and the test conclusion in terms of p-value.
e) 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.
f) Exam duration: 2h 10min
g) <u>Multichance students should skip: point d) of exercise 1, point b) of exercise 2.</u>

<u>Exercise 1 (15 points)</u>

In a thermal process, the cooling profile has a direct effect on the final quality and performance of the material. For Ti6Al4V components, it is known that the temperature of the material during the cooling process follows an exponential decay in the form $temp_t = \beta_1 \cdot e^{-0.1t} + \varepsilon_t$, where time t is expressed in seconds. Table 1 shows the material temperature during the cooling phase for one thermally treated part.

Table 1

Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)
1	361,9	11	123,3	21	50,5	31	16
2	322,7	12	114,7	22	46,7	32	18,8
3	298,1	13	104,8	23	38,5	33	10,5
4	265,6	14	104,3	24	29	34	3,4
5	240,7	15	89,8	25	35	35	8,8
6	213,7	16	80	26	39	36	10,9
7	194,5	17	66,4	27	32	37	10,2
8	164,8	18	60,9	28	22,9	38	11,1
9	150,8	19	51,9	29	26,1	39	13,5
10	135,5	20	50,3	30	19,6	40	13,9

a) Is the exponential decay model appropriate for designing a control chart procedure? If not, what is the appropriate model for fitting data in Table 1?
b) Based on the model selected in point a), design a suitable control chart with $ARL_0 = 250$.
c) Using the control chart designed at point b), determine if the cooling process of a different component of the same material (Table 2) is in-control or not.

Table 2

Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)
1	373	11	189,1	21	95,8	31	46,6
2	345,7	12	177,8	22	83,2	32	51,6
3	318,4	13	167,8	23	73	33	46,4
4	302	14	150,5	24	73,9	34	34,2
5	280,3	15	136,4	25	66,1	35	32,3
6	262,2	16	123,7	26	64,4	36	29,1
7	241,6	17	120,3	27	65,8	37	30
8	220,5	18	123,9	28	60,6	38	29,6
9	209,8	19	113,9	29	58,4	39	28,5
10	196,7	20	99,6	30	45,5	40	24,1

d) Design and implement a statistical test of hypothesis to check whether the first and second components (referring to Table 1 and Table 2, respectively) have a cooling history that is statistically different or not.

Exercise 2 (15 points)

In a finishing process for the production of an oil & gas component, two critical hole diameters are measured and monitored by randomly sampling $n = 3$ parts every hour. Under in-control conditions, it is known that the two diameters are independent and normally distributed with mean and standard deviation:

$$\mu_1 = 20.5\ mm,\ \mu_2 = 25.5\ mm,\ \sigma_1 = 0.2\ mm,\ \sigma_2 = 0.28\ mm$$

a) Design two univariate control charts for the mean with a familywise Type I error $\alpha = 0.01$ and determine if the data shown in Table 3 are in-control or not.

Table 3

Sample	Diameter hole 1 (mm)			Diameter hole 2 (mm)		
1	20,78	19,89	20,22	24,88	25,19	25,16
2	20,49	20,63	20,35	25,34	25,59	25,42
3	20,59	20,44	20,73	25,56	25,17	25,25
4	20,43	20,46	20,62	25,76	25,23	25,42
5	20,3	20,58	20,24	25,53	25,32	25,37

b) Determine the ARL_0 value if no familywise correction on the Type I error is applied and compare it with the ARL_0 of the chart designed at point a). Discuss the result.

c) In case both the diameters exhibit a shift of the mean $\Delta\mu_1 = \Delta\mu_2 = 0.3\ mm$, determine the probability of detecting it at the first sample after the shift using the control chart designed at point a).

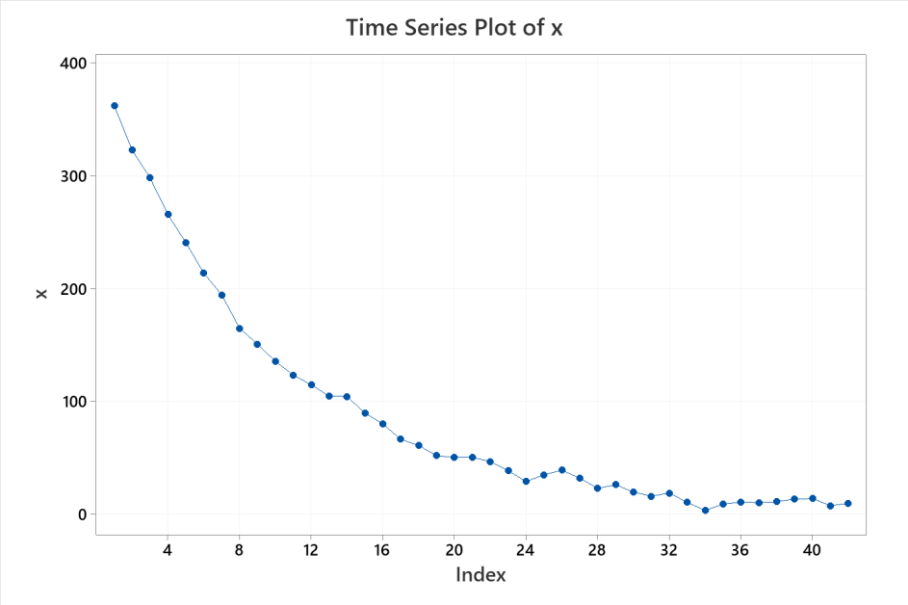
d) What is the minimum sample size to be used to detect a simultaneous shift of the means $\Delta\mu_1 = \Delta\mu_2 = 0.3\ mm$ with a probability $P > 90\%$?

Exercise 3 (3 points)

A quality characteristic X_t follows a stationary AR(1) model $X_t = \xi + \phi_1 X_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ with positive autocorrelation coefficient and known σ_ε^2. Let $E(X_t) = \mu$ and $V(X_t) = \sigma^2$. Compute the expressions of ξ and ϕ_1 as functions of μ, σ^2 and σ_ε^2.

Exercise 1 (solution)

a) The cooling process for data in Table 1 is:



By fitting a model in the form $temp_t = \beta_1 \cdot e^{-0.1t} + \varepsilon_t$, we get:

WORKSHEET 1

Regression Analysis: x versus exp2

Regression Equation

x = 392,37 exp2

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	392,37	2,46	159,32	0,000	1,00

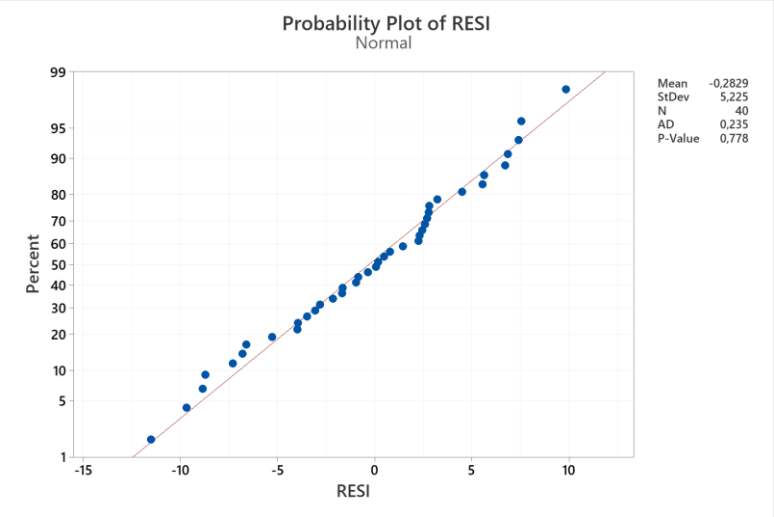
Model Summary

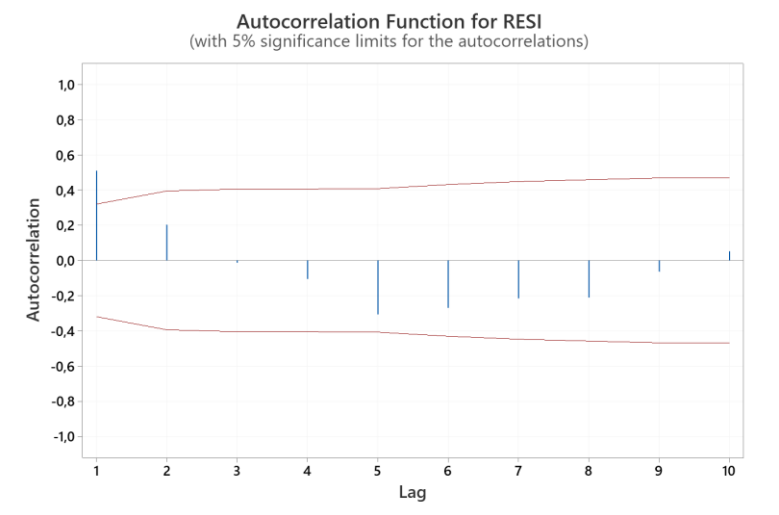
S	R-sq	R-sq(adj)	R-sq(pred)
5,23300	99,85%	99,84%	99,84%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	695111	695111	25383,52	0,000
exp2	1	695111	695111	25383,52	0,000
Error	39	1068	27		
Total	40	696179			

The residuals are normal but not independent:





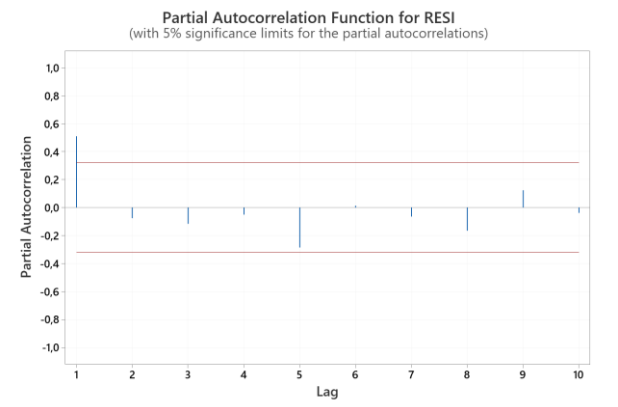
Bartlett's test at lag 1 (95% confidence):

$$|r_k| = 0.509$$

$$\frac{z_{\alpha/2}}{\sqrt{n}} = 0.354$$

The autocorrelation at lag 1 is significant.

PACF:



A more appropriate model should include an AR(1) term, i.e.: $temp_t = \beta_1 \cdot e^{-0.1t} + \beta_2 temp_{t-1} + \varepsilon_t$.

By fitting this model, we get:

WORKSHEET 1

Regression Analysis: x versus exp2; AR1

Method

Rows unused 1

Regression Equation

x = 157,8 exp2 + 0,537 AR1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	157,8	59,7	2,65	0,012	680,40
AR1	0,537	0,137	3,91	0,000	680,40

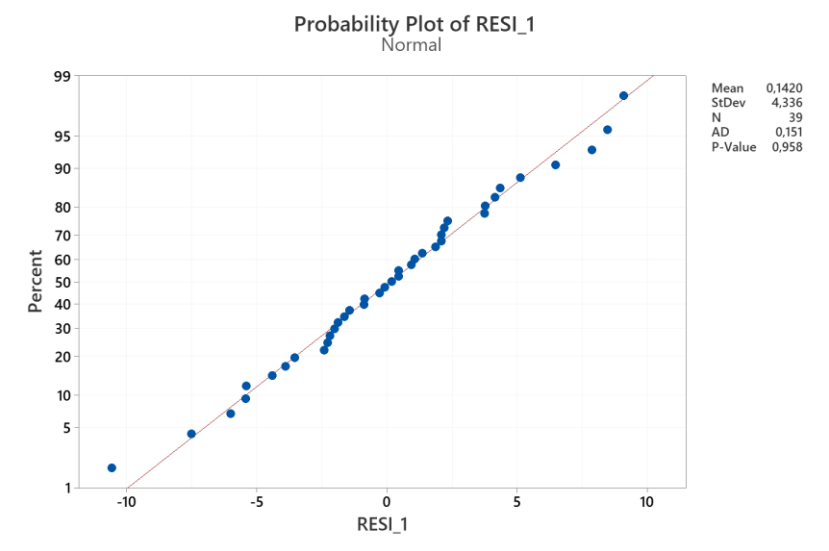
Model Summary

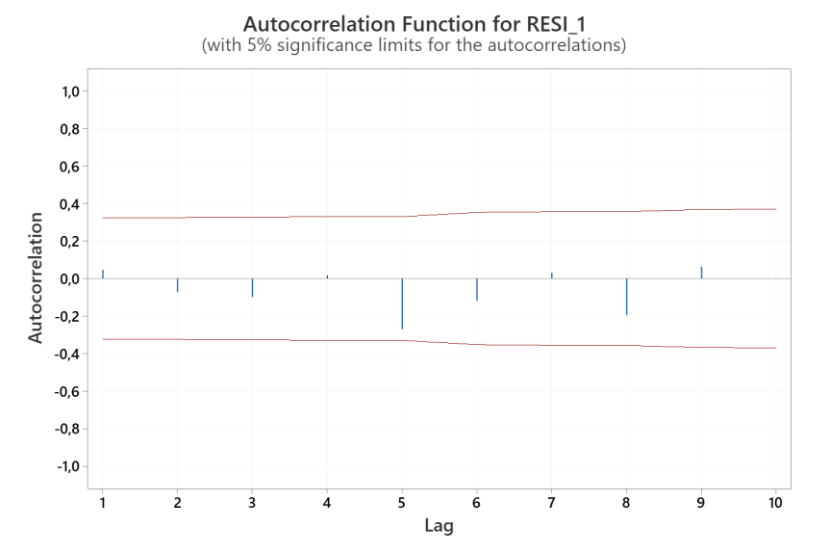
S	R-sq	R-sq(adj)	R-sq(pred)
4,39697	99,87%	99,87%	99,86%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	564492	282246	14598,89	0,000
exp2	1	135	135	7,00	0,012
AR1	1	295	295	15,26	0,000
Error	37	715	19		
Total	39	565207			

The residuals of this model are normal and independent, thus the model is adequate.





Test

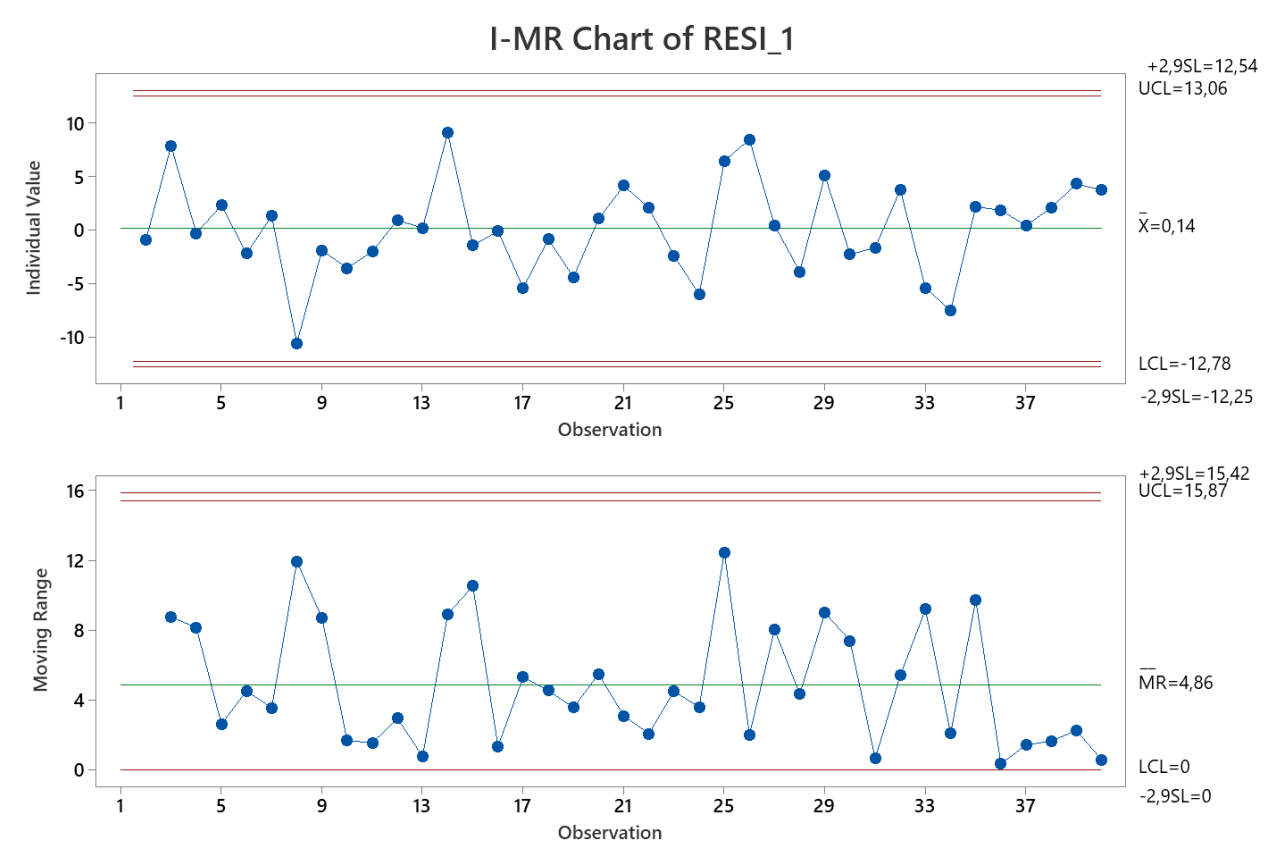
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs

Observed	Expected	P-Value
18	20,49	0,419

b) Given $ARL_0 = 250$, the Type I error is $\alpha = 0.004$, thus $K = z_{\alpha/2} = 2.878$.

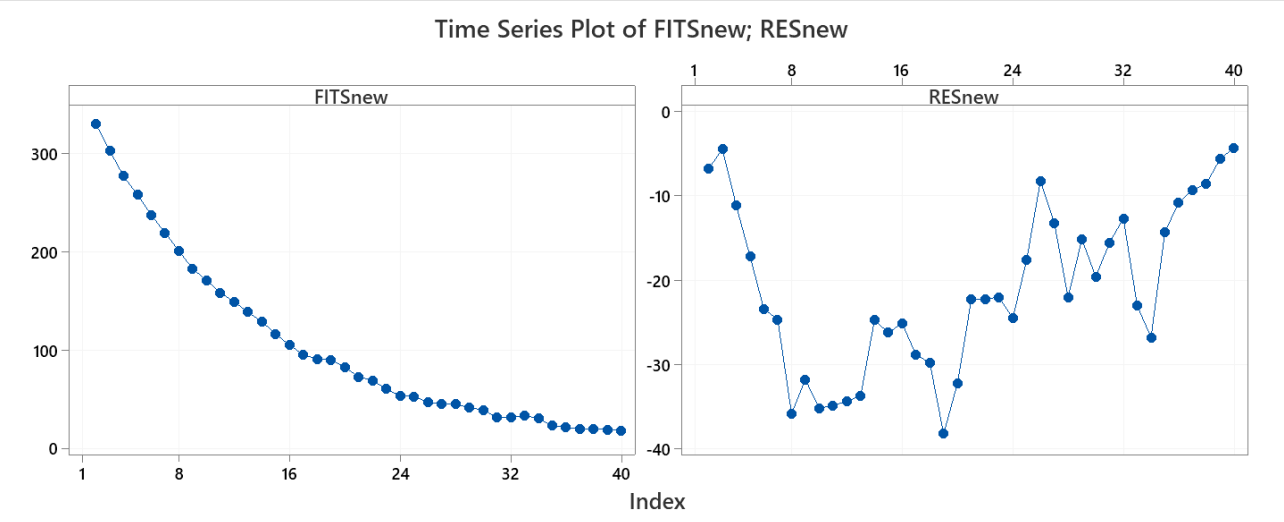
The special cause control chart for the process is the following:



The process is in-control.

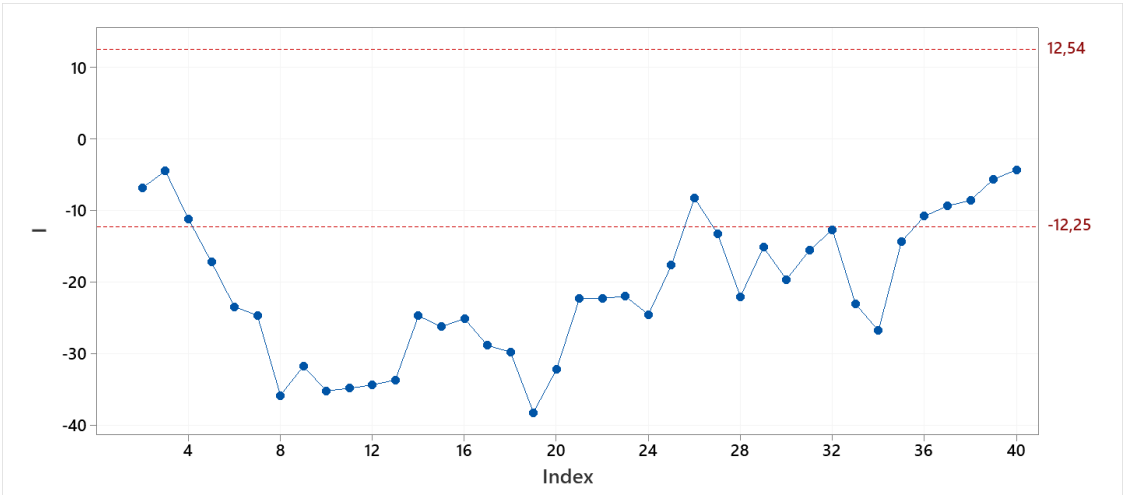
c) The same model shall be applied to the new data.

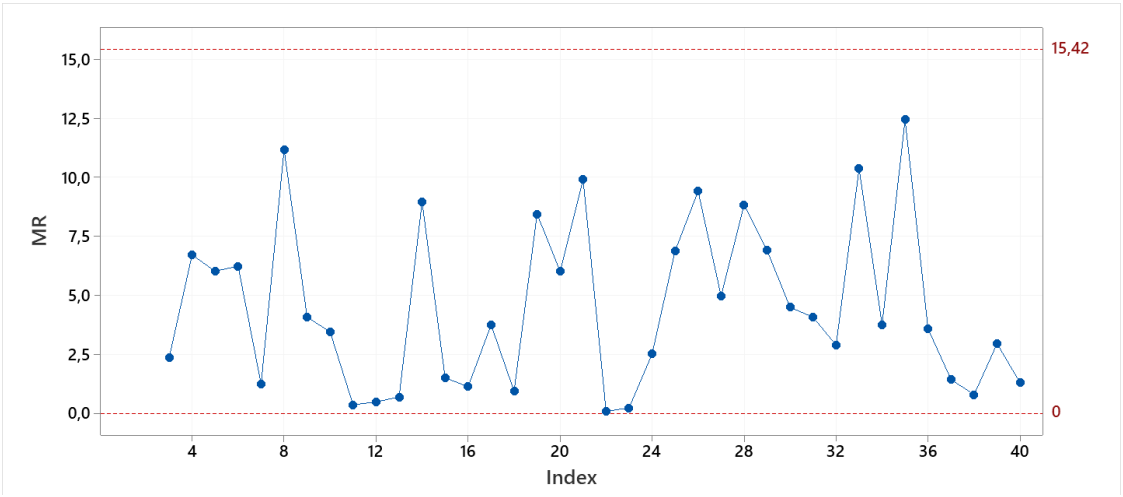
The resulting fits and residuals are the following:



FITSnew	RESnew
329,4967	-6,79671
302,542	-4,44202
276,7573	-11,1573
257,8845	-17,1845
237,1236	-23,4236
219,1626	-24,6626
200,6433	-35,8433
182,5652	-31,7652
170,714	-35,214
158,155	-34,855
149,0751	-34,3751
138,4841	-33,6841
129,0216	-24,7216
116,0284	-26,2284
105,1061	-25,1061
95,25436	-28,8544
90,68526	-29,7853
90,13623	-38,2362
82,52021	-32,2202
72,80882	-22,3088
68,92934	-22,2293
60,49925	-21,9992
53,51629	-24,5163
52,63731	-17,6373
47,21607	-8,21607
45,18783	-13,1878
44,93043	-22,0304
41,22486	-15,1249
39,2172	-19,6172
31,54226	-15,5423
31,45648	-12,6565
33,52936	-23,0294
30,1831	-26,7831
23,13055	-14,3305
21,65678	-10,7568
19,52807	-9,32807
19,64011	-8,54011
19,08937	-5,58937
18,19471	-4,29471

Using the previously designed control chart, the new cooling pattern results to be out-of-control.





Indeed, as shown in the following figure, the new cooling process is characterized by a slower decay than the previous (in-control) one.

d) A suitable way to check whether the first and second components (referring to Table 1 and Table 2, respectively) have a cooling time series that is statistically different is to fit the same model to the two time series and check if process parameters are statistically different.

Time series 1

Regression Equation

x = 157,8 exp2 + 0,537 AR1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	157,8	59,7	2,65	0,012	680,40
AR1	0,537	0,137	3,91	0,000	680,40

Time series 2

Regression Equation

xnew = 1,2 exp2 + 0,9294 ARnew

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	1,2	13,8	0,09	0,931	31,22
ARnew	0,9294	0,0258	36,07	0,000	31,22

Both models have normal and independent residuals. The results of two 2-sample t tests on the model coefficients (with different variances) with a familywise confidence of 95% are:

Test on β_1:

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Sample 1	40	157,8	59,7	9,4
Sample 2	40	1,2	13,8	2,2

Estimation for Difference

Difference	97,5% CI for Difference
156,60	(134,10; 179,10)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
16,16	43	0,000

Test on β_2:

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Sample 1	40	0,537	0,137	0,022
Sample 2	40	0,9294	0,0258	0,0041

Estimation for Difference

Difference	97,5% CI for Difference
-0,3924	(-0,4437; -0,3411)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-17,80	41	0,000

The two cooling histories are statistically significant.

Exercise 2 (solution)

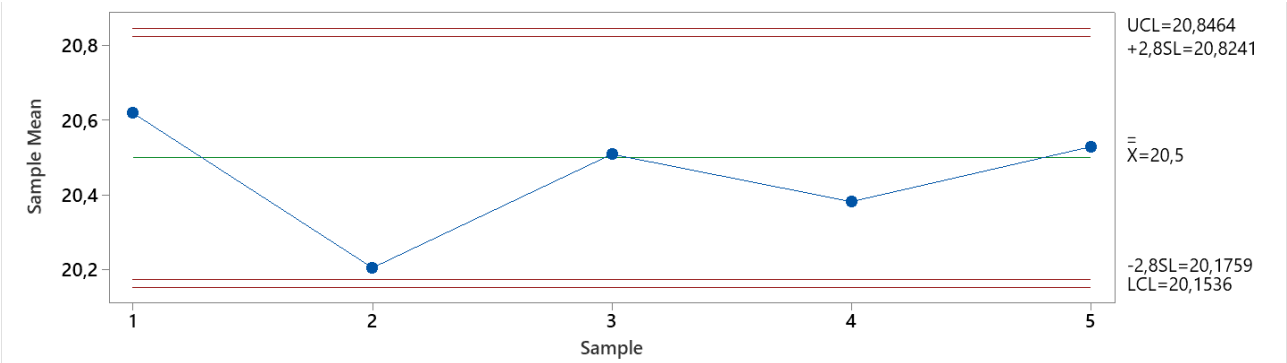
a) The two quality characteristics are independent. Therefore, the appropriate familywise correction is $\alpha^* = 1-(1-\alpha)^{1/2} = 0{,}005013$.

The control charts for the mean with $K = z_{\alpha^*/2} = 2.807$ have the following control limits:

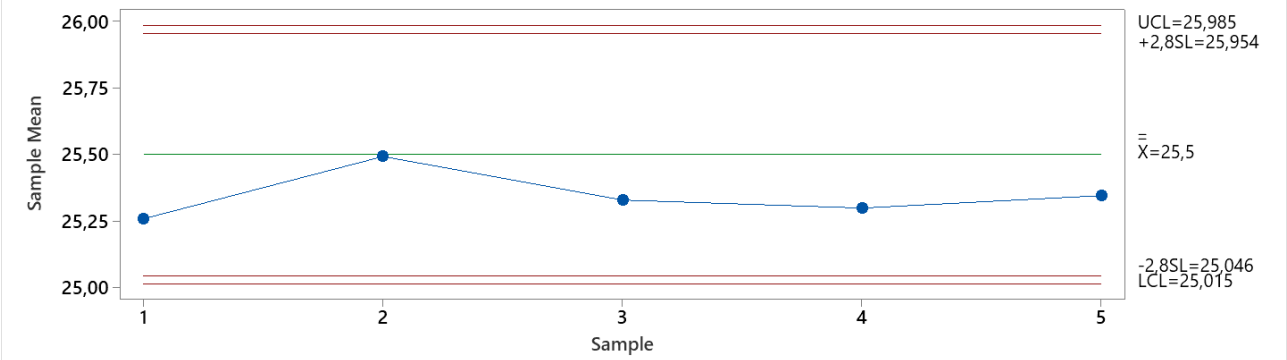
Diameter 1	Diameter 2
Xbar	Xbar
LCL =20.1536, UCL =20.8241	LCL =25.015, UCL =25.954

By applying these control charts to the provided data, no alarm is signalled, but all sample mean values of diameter 2 are below the center line, which may possibly indicate a small shift in the mean has occurred.

Diameter 1



Diameter 2



b) If no familywise correction was applied, $\alpha^* = 0.01$, and hence $\alpha = 1-(1-\alpha^*)^2 = 0{,}0199$. The corresponding Average Run Length is $ARL_0 = 50{,}25$. The expected one (with familywise correction) was $ARL_0 = 100$. Failing in using the proper correction would result is a much lower ARL.

c) The probability of detecting the shift is $P = 1 - \beta_{\bar{x},1} \cdot \beta_{\bar{x},2}$.

Let $\Delta\mu_1 = \Delta\mu_2 = 0.3\ mm$, then:

$$\beta_{\bar{x},1} = \phi\left(\frac{UCL_1 - \mu_1 - \Delta\mu_1}{\sigma_1/\sqrt{n}}\right) - \phi\left(\frac{LCL_1 - \mu_1 - \Delta\mu_1}{\sigma_1/\sqrt{n}}\right) =$$

$$\phi\left(\frac{20.8241-20.5-0.3}{0.2/\sqrt{3}}\right)-\phi\left(\frac{20.1536-20.5-0.3}{0.2/\sqrt{3}}\right)=0.5266$$

$$\beta_{\bar{x},2}=\phi\left(\frac{UCL_2-\mu_2-\Delta\mu_2}{\sigma_2/\sqrt{n}}\right)-\phi\left(\frac{LCL_2-\mu_2-\Delta\mu_2}{\sigma_2/\sqrt{n}}\right)=$$
$$\phi\left(\frac{25.954-25.5-0.3}{0.28/\sqrt{3}}\right)-\phi\left(\frac{25.015-25.5-0.3}{0.28/\sqrt{3}}\right)=0.8296$$

The resulting power is P =0.5166.

d) It is possible to express the power $P(n)=1-\beta_{\bar{x},1}(n)\cdot\beta_{\bar{x},2}(n)$ as a function of the sample size, keeping in mind that also control limits are functions of the sample size n.

By increasing the sample size, we get:

n	UCL1	LCL1	UCL2	LCL2	beta1	beta2	P
3	20,82412444	20,17588	25,95377	25,04623	0,582746	0,829254948	0,516755
4	20,7807	20,2193	25,89298	25,10702	0,423479	0,746700187	0,683788
5	20,75106571	20,24893	25,85149	25,14851	0,292154	0,65954168	0,807312
6	20,72919059	20,27081	25,82087	25,17913	0,192907	0,572423119	0,889576
7	20,71218926	20,28781	25,79706	25,20294	0,122694	0,488937325	0,940011
8	20,69848487	20,30152	25,77788	25,22212	0,075552	0,411589674	0,968903
9	20,68713333	20,31287	25,76199	25,23801	0,045228	0,341899076	0,984537
10	20,67753027	20,32247	25,74854	25,25146	0,026408	0,280568119	0,992591

The minimum sample size to have a power P>90% is $n=7$.

Exercise 3 (solution)

Given a stationary AR(1) model $X_t=\xi+\phi_1X_{t-1}+\varepsilon_t,\ \varepsilon_t\sim N(0,\sigma_\varepsilon^2)$, it is known that:

$$\mu=\frac{\xi}{1-\phi_1}$$
$$\sigma^2=\frac{\sigma_\varepsilon^2}{1-{\phi_1}^2}$$

Therefore:

$$1-\phi_1=\frac{\xi}{\mu}$$
$$1-{\phi_1}^2=\frac{\sigma_\varepsilon^2}{\sigma^2}$$

By solving the two equations with two unknowns:

$$\phi_1=\sqrt{1-\frac{\sigma_\varepsilon^2}{\sigma^2}}$$

$$\xi = \mu\left(1 - \sqrt{1 - \frac{\sigma_\varepsilon^2}{\sigma^2}}\right)$$

QUALITY DATA ANALYSIS

08/07/2022

General recommendations:

a) write the solutions in CLEAR and READABLE way on paper and show (qualitatively) all the relevant plots;
b) avoid (if not required) theoretical introductions or explanations covered during the course;
c) always state the assumptions and report all relevant steps/discussion/formulas/expression to present and motivate your solution;
d) when using hypothesis tests provide the numerical value of the test statistic and the test conclusion in terms of p-value.
e) 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.
f) Exam duration: 2h 10min
g) <u>Multichance students should skip: point b) of exercise 1, point d) of exercise 3.</u>

<u>Exercise 1 (15 points)</u>
In a finishing process for the production of an oil & gas component, two critical hole diameters are measured and monitored by randomly sampling $n = 3$ parts every hour. Under in-control conditions, it is known that the two diameters are independent and normally distributed with mean and standard deviation:

$$\mu_1 = 13.5\ mm,\ \mu_2 = 20.0\ mm,\ \sigma_1 = 0.2\ mm,\ \sigma_2 = 0.28\ mm$$

a) Design two univariate control charts for the mean with a familywise Type I error $\alpha = 0.01$ and determine if the data shown in Table 1 are in-control or not.

Table 1

Sample	Diameter hole 1 (mm)			Diameter hole 2 (mm)		
1	13,78	13,49	13,59	19,38	19,84	20,06
2	13,43	13,30	12,89	20,26	20,03	19,69
3	13,63	13,44	13,46	20,09	19,67	19,73
4	13,58	13,22	13,35	19,82	19,66	19,92
5	13,93	14,02	14,05	19,75	19,92	19,87

b) Determine the ARL_0 value if no familywise correction on the Type I error is applied and compare it with the ARL_0 of the chart designed at point a). Discuss the result.
c) In case both the diameters exhibit a shift of the mean $\Delta\mu_1 = \Delta\mu_2 = 0.25\ mm$, determine the probability of detecting it at the first sample after the shift using the control chart designed at point a).
d) What is the minimum sample size to be used to detect a simultaneous shift of the means $\Delta\mu_1 = \Delta\mu_2 = 0.25\ mm$ with a probability $P > 90\%$?

<u>Exercise 2 (3 points)</u>
A quality characteristic X_t follows a stationary AR(1) model $X_t = \xi + \phi_1 X_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ with positive autocorrelation coefficient and known σ_ε^2. Let $E(X_t) = \mu$ and $V(X_t) = \sigma^2$. Compute the expressions of ξ and ϕ_1 as functions of μ, σ^2 and σ_ε^2.

<u>Exercise 3 (15 points)</u>

In a thermal process, the cooling profile has a direct effect on the final quality and performance of the material. For Ti6Al4V components, it is known that the temperature of the material during the cooling

process follows an exponential decay in the form $temp_t = \beta_1 \cdot e^{-0.07t} + \varepsilon_t$, where time t is expressed in seconds. Table 2 shows the material temperature during the cooling phase for one thermally treated part.

Table 2

Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)
1	373	11	189,1	21	95,8	31	46,6
2	345,7	12	177,8	22	83,2	32	51,6
3	318,4	13	167,8	23	73	33	46,4
4	302	14	150,5	24	73,9	34	34,2
5	280,3	15	136,4	25	66,1	35	32,3
6	262,2	16	123,7	26	64,4	36	29,1
7	241,6	17	120,3	27	65,8	37	30
8	220,5	18	123,9	28	60,6	38	29,6
9	209,8	19	113,9	29	58,4	39	28,5
10	196,7	20	99,6	30	45,5	40	24,1

a) Is the exponential decay model appropriate for designing a control chart procedure? If not, what is the appropriate model for fitting data in Table 2?
b) Based on model fitted in point a), design a suitable control chart with $ARL_0 = 250$.
c) Using the control chart designed at point b), determine if the cooling process of a different component of the same material (Table 3) is in-control or not.

Table 3

Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)
1	361,9	11	123,3	21	50,5	31	16
2	322,7	12	114,7	22	46,7	32	18,8
3	298,1	13	104,8	23	38,5	33	10,5
4	265,6	14	104,3	24	29	34	3,4
5	240,7	15	89,8	25	35	35	8,8
6	213,7	16	80	26	39	36	10,9
7	194,5	17	66,4	27	32	37	10,2
8	164,8	18	60,9	28	22,9	38	11,1
9	150,8	19	51,9	29	26,1	39	13,5
10	135,5	20	50,3	30	19,6	40	13,9

d) Design and implement a statistical test of hypothesis to check whether the first and second components (referring to Table 2 and Table 3, respectively) have a cooling history that is statistically different or not.

Exercise 1 (solution)

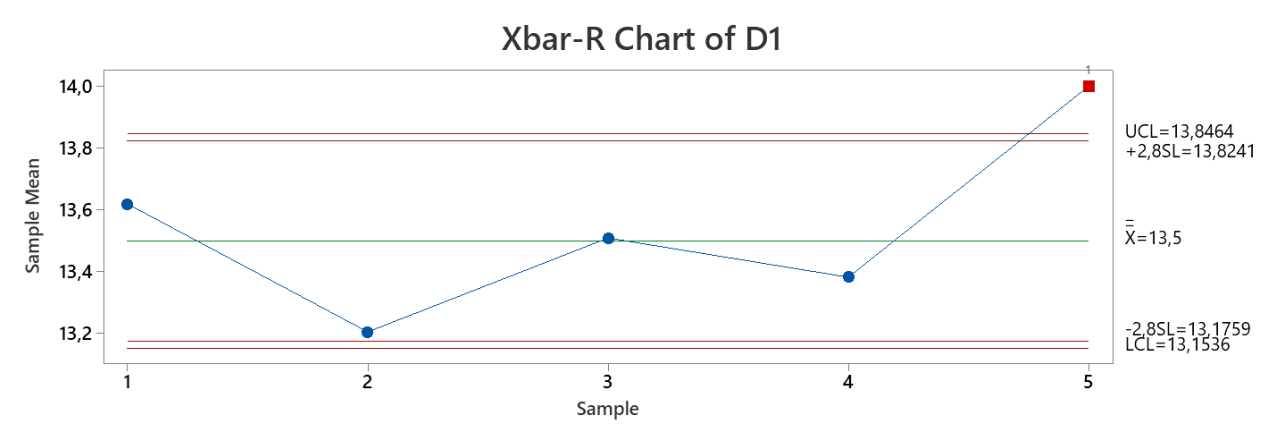
a) The two quality characteristics are independent. Therefore, the appropriate familywise correction is $\alpha^* = 1-(1-\alpha)^{1/2} = 0{,}005013$.

The control charts with $K = z_{\alpha^*/2} = 2.807$ have the following control limits:

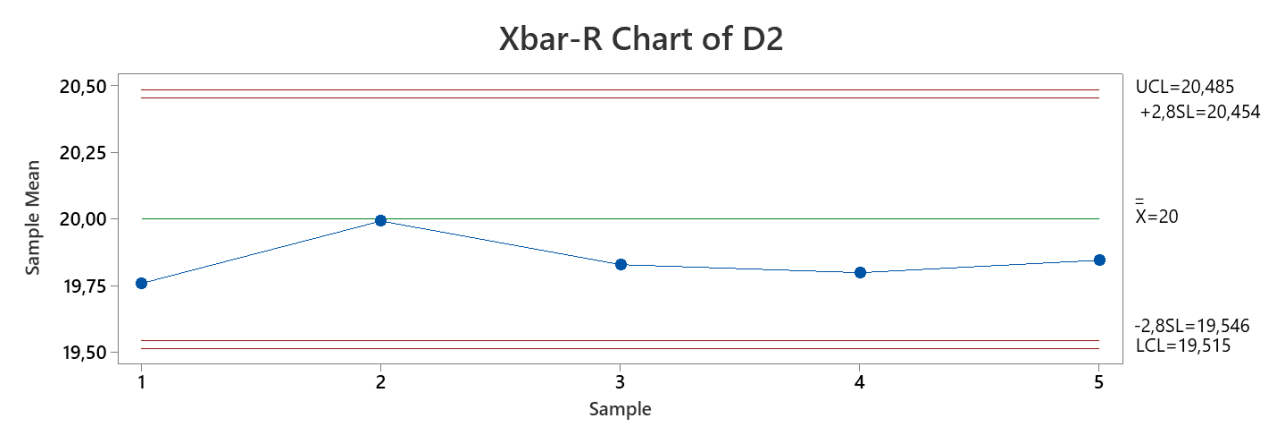
Diameter 1	Diameter 2
Xbar	Xbar
LCL =13,1759, UCL =13.8241	LCL =19,546, UCL =20,454

By applying these control charts to the provided data, sample 5 exhibits a violation of control limits in the chart for the mean of diameter 1. Moreover, all sample mean values of diameter 2 are below the center line, which may possibly indicate a small shift in the mean has occurred.

Diameter 1



Diameter 2



b) If no familywise correction was applied, $\alpha^* = 0.01$, and hence $\alpha = 1-(1-\alpha^*)^2 = 0{,}0199$. The corresponding Average Run Length is $ARL_0 = 50{,}25$. The expected one (with familywise correction) was $ARL_0 = 100$. Failing in using the proper correction would result is a much lower ARL.

c) The probability of detecting the shift is $P = 1 - \beta_{\bar{x},1} \cdot \beta_{\bar{x},2}$.

Let $\Delta\mu_1 = \Delta\mu_2 = 0.25\ mm$, then:

$$\beta_{\bar{x},1} = \phi\left(\frac{UCL_1 - \mu_1 - \Delta\mu_1}{\sigma_1/\sqrt{n}}\right) - \phi\left(\frac{LCL_1 - \mu_1 - \Delta\mu_1}{\sigma_1/\sqrt{n}}\right) =$$

$$\phi\left(\frac{13.8241 - 13.5 - 0.25}{0.2/\sqrt{3}}\right) - \phi\left(\frac{13.1759 - 13.5 - 0.25}{\frac{0.2}{\sqrt{3}}}\right) = 0.7395$$

$$\beta_{\bar{x},2} = \phi\left(\frac{UCL_2 - \mu_2 - \Delta\mu_2}{\sigma_2/\sqrt{n}}\right) - \phi\left(\frac{LCL_2 - \mu_2 - \Delta\mu_2}{\sigma_2/\sqrt{n}}\right) =$$

$$\phi\left(\frac{20.454 - 20.0 - 0.25}{0.28/\sqrt{3}}\right) - \phi\left(\frac{19.546 - 20.0 - 0.25}{\frac{0.28}{\sqrt{3}}}\right) = 0.8965$$

The resulting power is $P = 0.337$.

d) It is possible to express the power $P(n) = 1 - \beta_{\bar{x},1}(n) \cdot \beta_{\bar{x},2}(n)$ as a function of the sample size, keeping in mind that also control limits are functions of the sample size n.

By increasing the sample size, we get:

n	UCL1	LCL1	UCL2	LCL2	beta1	beta2	P
3	13,82412444	13,17588	20,45377	19,54623	0,739542	0,896253474	0,337183
4	13,7807	13,2193	20,39298	19,60702	0,620578	0,846438266	0,474719
5	13,75106571	13,24893	20,35149	19,64851	0,504753	0,791175871	0,600651
6	13,72919059	13,27081	20,32087	19,67913	0,399415	0,732356191	0,707486
7	13,71218926	13,28781	20,29706	19,70294	0,308471	0,671739564	0,792788
8	13,69848487	13,30152	20,27788	19,72212	0,233143	0,610881978	0,857577
9	13,68713333	13,31287	20,26199	19,73801	0,17284	0,551095069	0,904748
10	13,67753027	13,32247	20,24854	19,75146	0,125929	0,493432821	0,937862
11	13,66926847	13,33073	20,23698	19,76302	0,090321	0,43869758	0,960376
12	13,66206222	13,33794	20,22689	19,77311	0,063863	0,387459022	0,975256

The minimum sample size to have a power P>90% is $n = 9$.

Exercise 2 (solution)

Given a stationary AR(1) model $x_t = \xi + \phi_1 x_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, it is known that:

$$\mu = \frac{\xi}{1 - \phi_1}$$

$$\sigma^2 = \frac{\sigma_\varepsilon^2}{1 - {\phi_1}^2}$$

Therefore:

$$1 - \phi_1 = \frac{\xi}{\mu}$$

$$1 - {\phi_1}^2 = \frac{\sigma_\varepsilon^2}{\sigma^2}$$

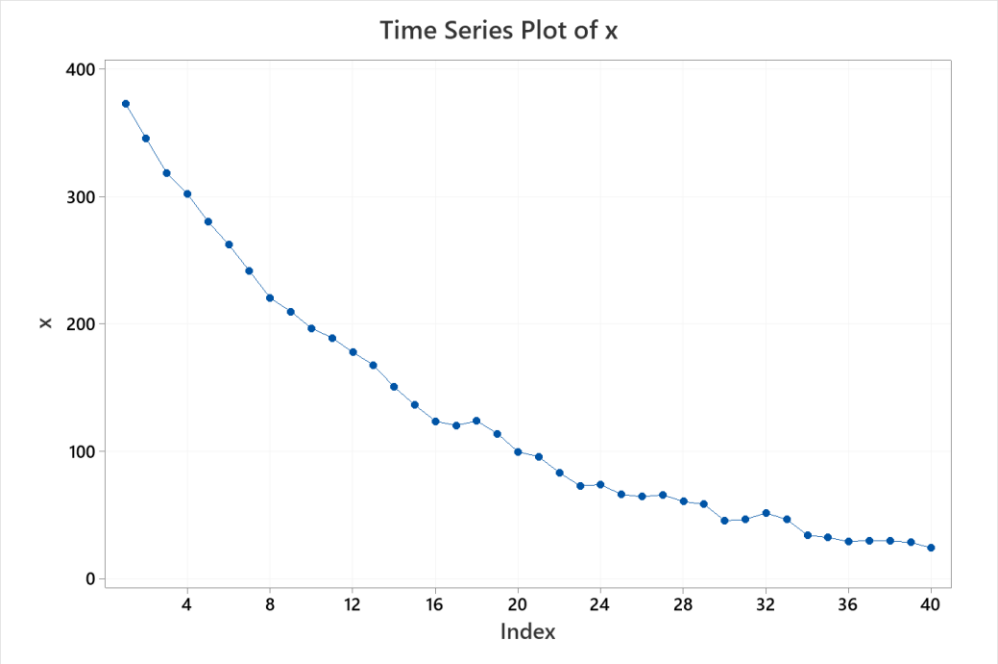
By solving the two equations with two unknowns:

$$\phi_1 = \sqrt{1 - \frac{\sigma_\varepsilon^2}{\sigma^2}}$$

$$\xi = \mu\left(1 - \sqrt{1 - \frac{\sigma_\varepsilon^2}{\sigma^2}}\right)$$

Exercise 3 (solution)

a) The cooling process for data in Table 2 is:



By fitting a model in the form $temp_t = \beta_1 \cdot e^{-0.07t} + \varepsilon_t$, we get:

EXE1

Regression Analysis: x versus exp2

Regression Equation

x = 398,79 exp2

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	398,79	1,74	229,79	0,000	1,00

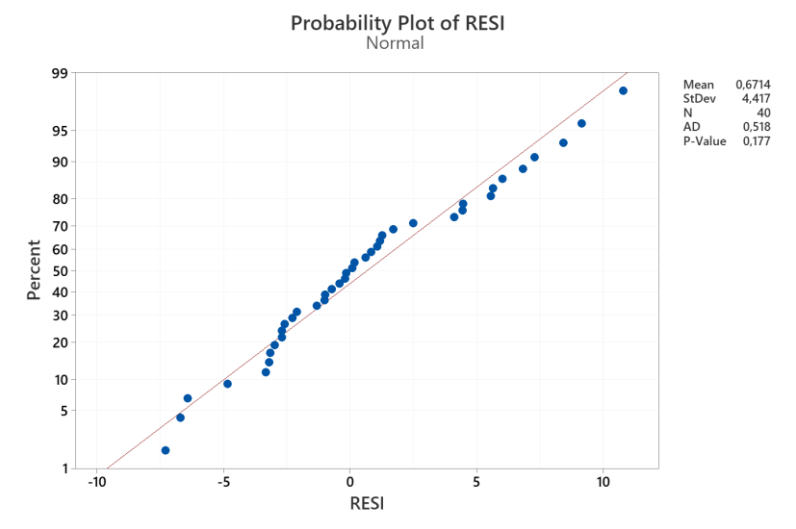
Model Summary

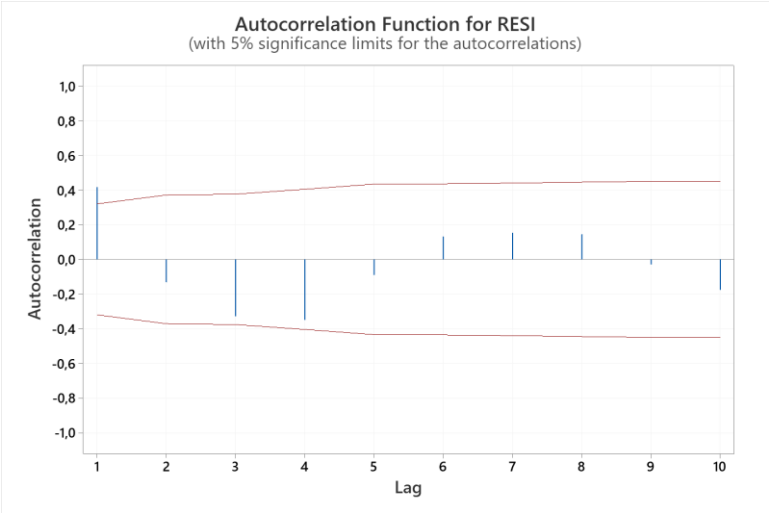
S	R-sq	R-sq(adj)	R-sq(pred)
4,46861	99,93%	99,92%	99,92%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1054364	1054364	52801,46	0,000
exp2	1	1054364	1054364	52801,46	0,000
Error	39	779	20		
Total	40	1055142			

The residuals are normal but not independent:





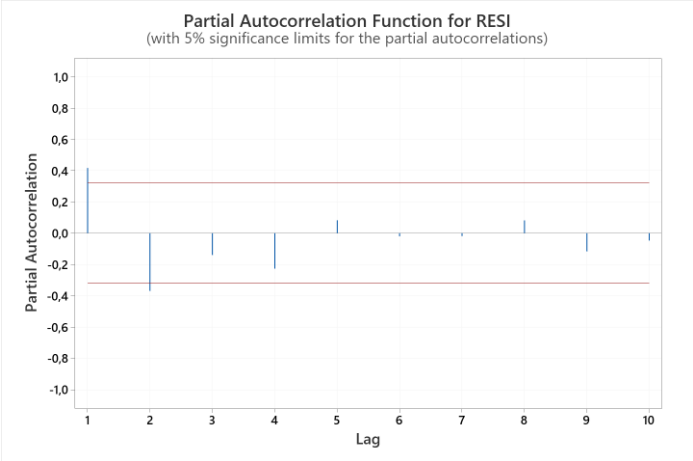
Bartlett's test at lag 1 (95% confidence):

$$|r_k| = 0.416$$

$$\frac{z_{\alpha/2}}{\sqrt{n}} = 0.354$$

The autocorrelation at lag 1 is significant.

PACF:



A more appropriate model should include an AR(1) or AR(2) term. Following the parsimony principle, we can try with an AR(1) term: $temp_t = \beta_1 \cdot e^{-0.07t} + \beta_2 temp_{t-1} + \varepsilon_t$.

By fitting this model, we get:

EXE1

Regression Analysis: x versus exp2; AR1

Method

Rows unused 1

Regression Equation

x = 214,8 exp2 + 0,430 AR1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	214,8	63,4	3,39	0,002	1354,18
AR1	0,430	0,148	2,90	0,006	1354,18

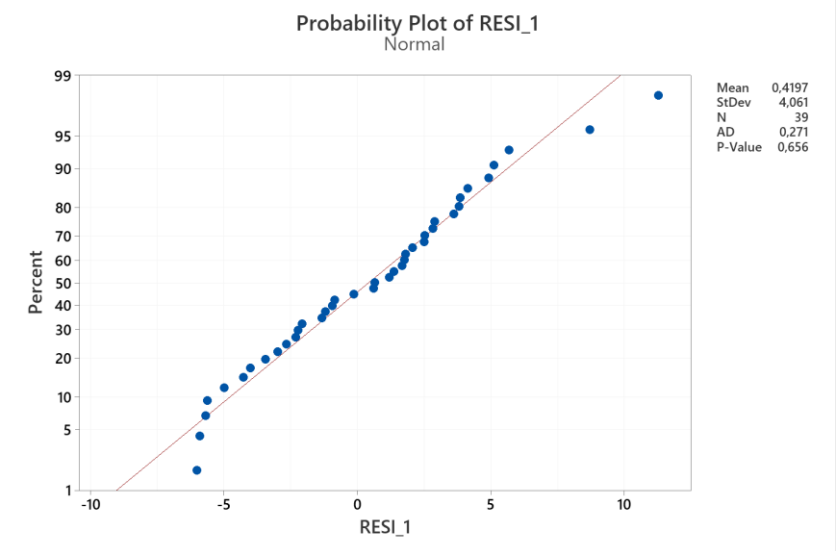
Model Summary

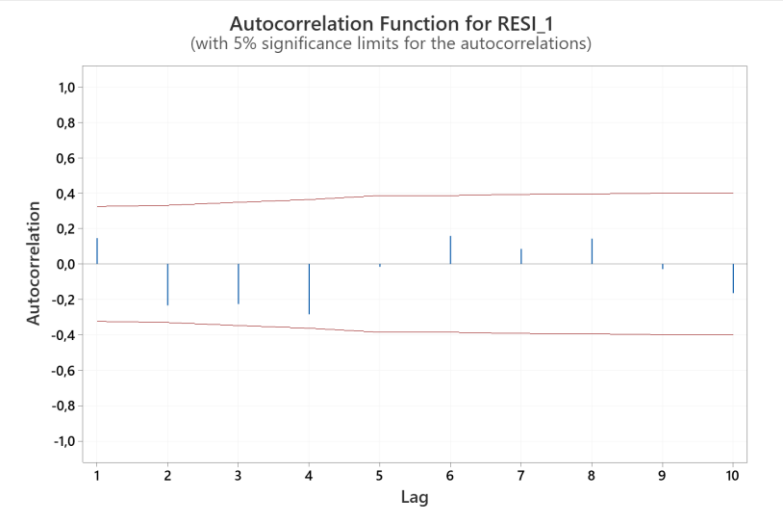
S	R-sq	R-sq(adj)	R-sq(pred)
4,13755	99,93%	99,93%	99,92%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	915380	457690	26735,23	0,000
exp2	1	196	196	11,47	0,002
AR1	1	144	144	8,40	0,006
Error	37	633	17		
Total	39	916013			

The residuals of this model are normal and independent, thus the model is adequate.





Test

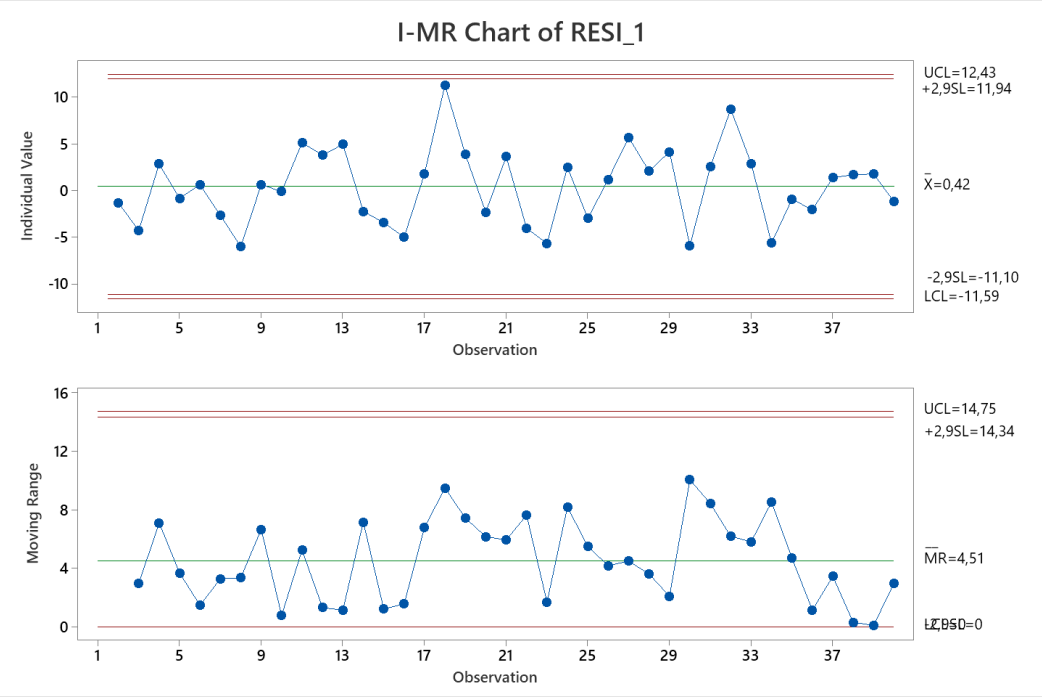
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs

Observed	Expected	P-Value
21	20,38	0,841

b) Given $ARL_0 = 250$, the Type I error is $\alpha = 0.004$, thus $K = z_{\alpha/2} = 2.878$.

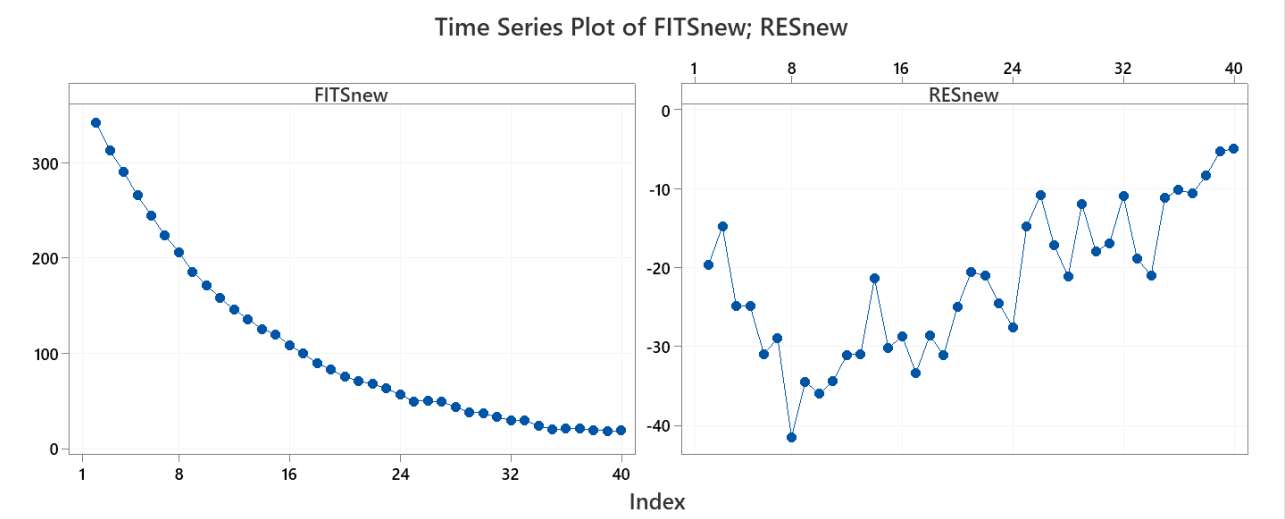
The special cause control chart for the process is the following:



The process is in-control.

c) The same model shall be applied to the new data.

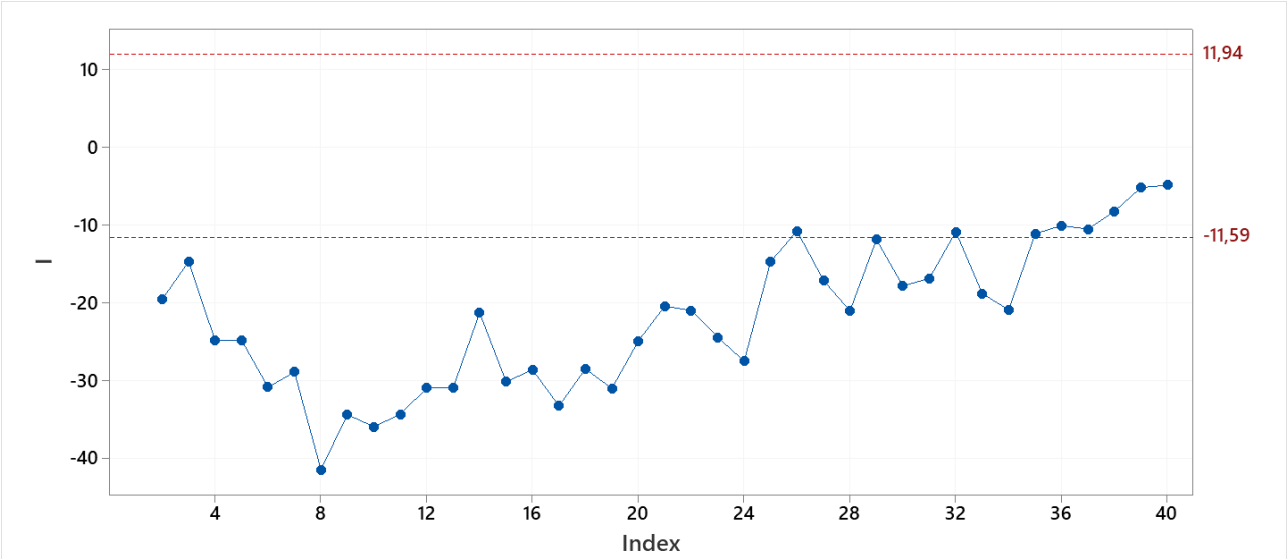
The resulting fits and residuals are the following:

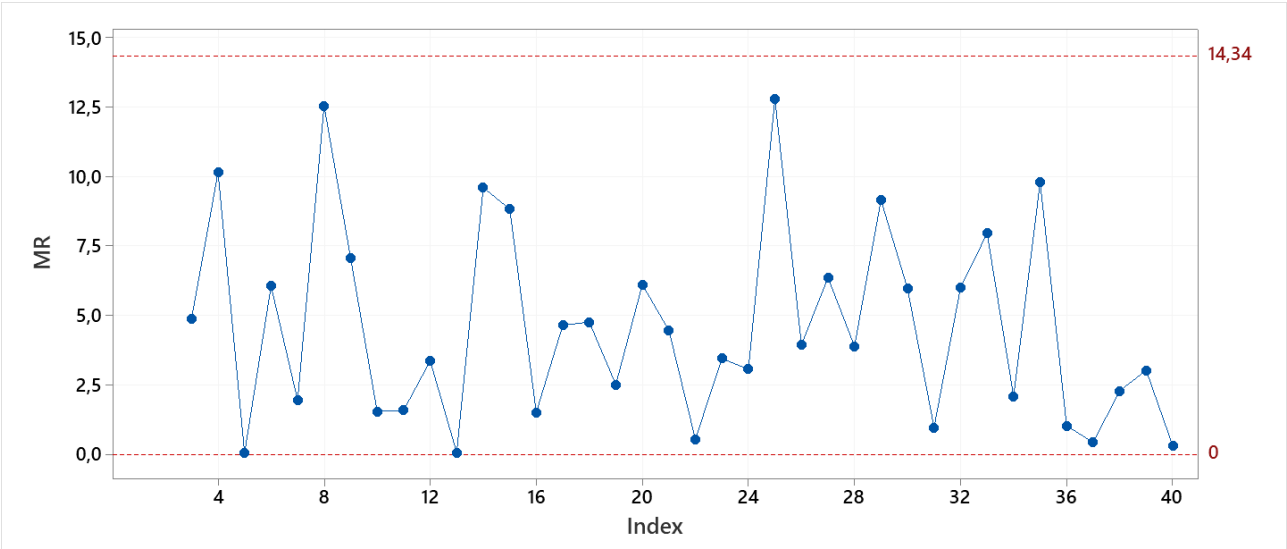


FITSnew	RESnew
342,3551	-19,6551
312,8745	-14,7745
290,5253	-24,9253
265,575	-24,875
244,6347	-30,9347
223,4831	-28,9831
206,3307	-41,5307
185,2647	-34,4647
171,5105	-36,0105
157,7202	-34,4202
145,7504	-31,0504
135,7832	-30,9832
125,6808	-21,3808
120,0156	-30,2156
108,6989	-28,6989
99,74673	-33,3467
89,48088	-28,5809
82,99672	-31,0967
75,28603	-24,986
71,01699	-20,517
67,76406	-21,0641
63,01686	-24,5169
56,58813	-27,5881
49,79664	-14,7966
49,85313	-10,8531

49,22022	-17,2202
44,01639	-21,1164
38,05787	-11,9579
37,52664	-17,9266
32,95335	-16,9534
29,74729	-10,9473
29,40532	-18,9053
24,39486	-20,9949
19,99786	-11,1979
21,06672	-10,1667
20,8013	-10,6013
19,41088	-8,31088
18,7821	-5,2821
18,867	-4,967

Using the previously designed control chart, the new cooling pattern results to be out-of-control.





Indeed, as shown in the following figure, the new cooling process is characterized by a faster decay than the previous (in-control) one.

d) A suitable way to check whether the first and second components (referring to Table 2 and Table 3, respectively) have a cooling time series that is statistically different is to fit the same model to the two time series and check if process parameters are statistically different.

Time series 1

Regression Equation

x = 214,8 exp2 + 0,430 AR1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	214,8	63,4	3,39	0,002	1354,18
AR1	0,430	0,148	2,90	0,006	1354,18

Time series 2

Regression Equation

xnew = 9,9 exp2 + 0,8726 AR1new

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	9,9	10,7	0,92	0,363	29,35
AR1new	0,8726	0,0308	28,34	0,000	29,35

Both models have normal and independent residuals. The results of two 2-sample t tests on the model coefficients (with different variances) with a familywise confidence of 95% are:

Test on β_1:

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Sample 1	40	214,8	63,4	10
Sample 2	40	9,9	10,7	1,7

Estimation for Difference

Difference	97,5% CI for Difference
204,9	(181,2; 228,6)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
20,16	41	0,000

Test on β_2:

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Sample 1	40	0,430	0,148	0,023
Sample 2	40	0,8726	0,0308	0,0049

Estimation for Difference

Difference	97,5% CI for Difference
-0,4426	(-0,4982; -0,3870)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-18,52	42	0,000

The two cooling histories are statistically significant.

QUALITY DATA ANALYSIS

08/07/2022

General recommendations:

a) write the solutions in CLEAR and READABLE way on paper and show (qualitatively) all the relevant plots;
b) avoid (if not required) theoretical introductions or explanations covered during the course;
c) always state the assumptions and report all relevant steps/discussion/formulas/expression to present and motivate your solution;
d) when using hypothesis tests provide the numerical value of the test statistic and the test conclusion in terms of p-value.
e) 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.
f) Exam duration: 2h 10min
g) <u>Multichance students should skip: point b) of exercise 2, point d) of exercise 3</u>

<u>Exercise 1 (3 points)</u>
A quality characteristic X_t follows a stationary AR(1) model $X_t = \xi + \phi_1 X_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ with positive autocorrelation coefficient and known σ_ε^2. Let $E(X_t) = \mu$ and $V(X_t) = \sigma^2$. Compute the expressions of ξ and ϕ_1 as functions of μ, σ^2 and σ_ε^2.

<u>Exercise 2 (15 points)</u>
In a finishing process for the production of an oil & gas component, two critical hole diameters are measured and monitored by randomly sampling $n = 3$ parts every hour. Under in-control conditions, it is known that the two diameters are independent and normally distributed with mean and standard deviation:

$$\mu_1 = 20.5\ mm,\ \mu_2 = 25.5\ mm,\ \sigma_1 = 0.2\ mm,\ \sigma_2 = 0.28\ mm$$

a) Design two univariate control charts for the mean with a familywise Type I error $\alpha = 0.01$ and determine if the data shown in Table 1 are in-control or not.

Table 1

Sample	Diameter hole 1 (mm)			Diameter hole 2 (mm)		
1	20,78	20,49	20,59	24,88	25,34	25,56
2	20,43	20,30	19,89	25,76	25,53	25,19
3	20,63	20,44	20,46	25,59	25,17	25,23
4	20,58	20,22	20,35	25,32	25,16	25,42
5	20,73	20,62	20,24	25,25	25,42	25,37

b) Determine the ARL_0 value if no familywise correction on the Type I error is applied and compare it with the ARL_0 of the chart designed at point a). Discuss the result.
c) In case both the diameters exhibit a shift of the mean $\Delta\mu_1 = \Delta\mu_2 = 0.3\ mm$, determine the probability of detecting it at the first sample after the shift using the control chart designed at point a).
d) What is the minimum sample size to be used to detect a simultaneous shift of the means $\Delta\mu_1 = \Delta\mu_2 = 0.3\ mm$ with a probability $P > 90\%$?

Exercise 3 (15 points)

In a thermal process, the cooling profile has a direct effect on the final quality and performance of the material. For Ti6Al4V components, it is known that the temperature of the material during the cooling process follows an exponential decay in the form $temp_t = \beta_1 \cdot e^{-0.07t} + \varepsilon_t$, where time t is expressed in seconds. Table 2 shows the material temperature during the cooling phase for one thermally treated part.

Table 2

Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)
1	373	11	189,1	21	95,8	31	46,6
2	345,7	12	177,8	22	83,2	32	51,6
3	318,4	13	167,8	23	73	33	46,4
4	302	14	150,5	24	73,9	34	34,2
5	280,3	15	136,4	25	66,1	35	32,3
6	262,2	16	123,7	26	64,4	36	29,1
7	241,6	17	120,3	27	65,8	37	30
8	220,5	18	123,9	28	60,6	38	29,6
9	209,8	19	113,9	29	58,4	39	28,5
10	196,7	20	99,6	30	45,5	40	24,1

a) Is the exponential decay model appropriate for designing a control chart procedure? If not, what is the appropriate model for fitting data in Table 2?
b) Based on model fitted in point a), design a suitable control chart with $ARL_0 = 250$.
c) Using the control chart designed at point b), determine if the cooling process of a different component of the same material (Table 3) is in-control or not.

Table 3

Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)	Time (s)	Temperature (°C)
1	361,9	11	123,3	21	50,5	31	16
2	322,7	12	114,7	22	46,7	32	18,8
3	298,1	13	104,8	23	38,5	33	10,5
4	265,6	14	104,3	24	29	34	3,4
5	240,7	15	89,8	25	35	35	8,8
6	213,7	16	80	26	39	36	10,9
7	194,5	17	66,4	27	32	37	10,2
8	164,8	18	60,9	28	22,9	38	11,1
9	150,8	19	51,9	29	26,1	39	13,5
10	135,5	20	50,3	30	19,6	40	13,9

d) Design and implement a statistical test of hypothesis to check whether the first and second components (referring to Table 2 and Table 3, respectively) have a cooling history that is statistically different or not.

Exercise 1 (solution)

Given a stationary AR(1) model $x_t = \xi + \phi_1 x_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, it is known that:

$$\mu = \frac{\xi}{1-\phi_1}$$
$$\sigma^2 = \frac{\sigma_\varepsilon^2}{1-{\phi_1}^2}$$

Therefore:

$$1-\phi_1 = \frac{\xi}{\mu}$$
$$1-{\phi_1}^2 = \frac{\sigma_\varepsilon^2}{\sigma^2}$$

By solving the two equations with two unknowns:

$$\phi_1 = \sqrt{1-\frac{\sigma_\varepsilon^2}{\sigma^2}}$$

$$\xi = \mu\left(1-\sqrt{1-\frac{\sigma_\varepsilon^2}{\sigma^2}}\right)$$

Exercise 2 (solution)

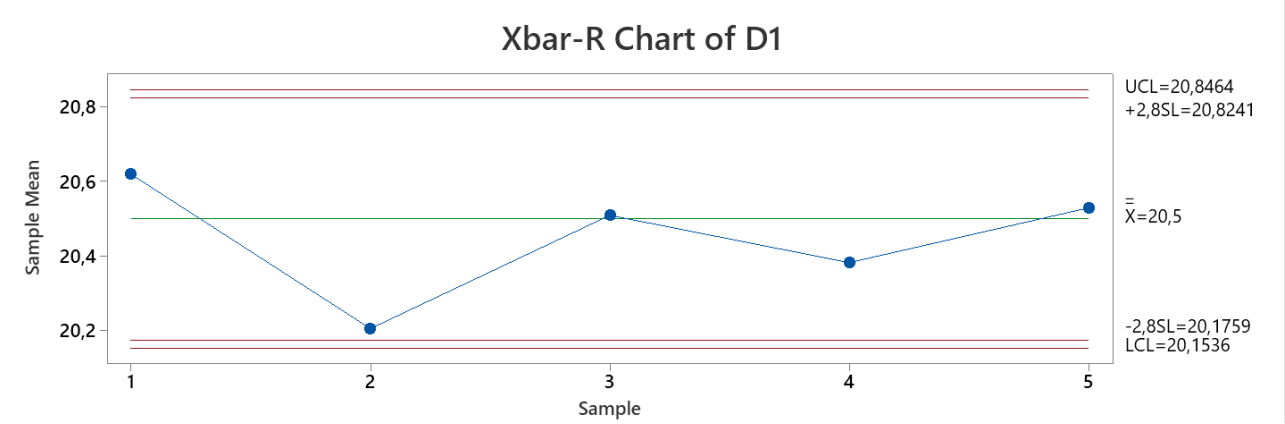
a) The two quality characteristics are independent. Therefore, the appropriate familywise correction is $\alpha^* = 1-(1-\alpha)^{1/2} = 0{,}005013$.

The control charts with $K = z_{\alpha^*/2} = 2.807$ have the following control limits:

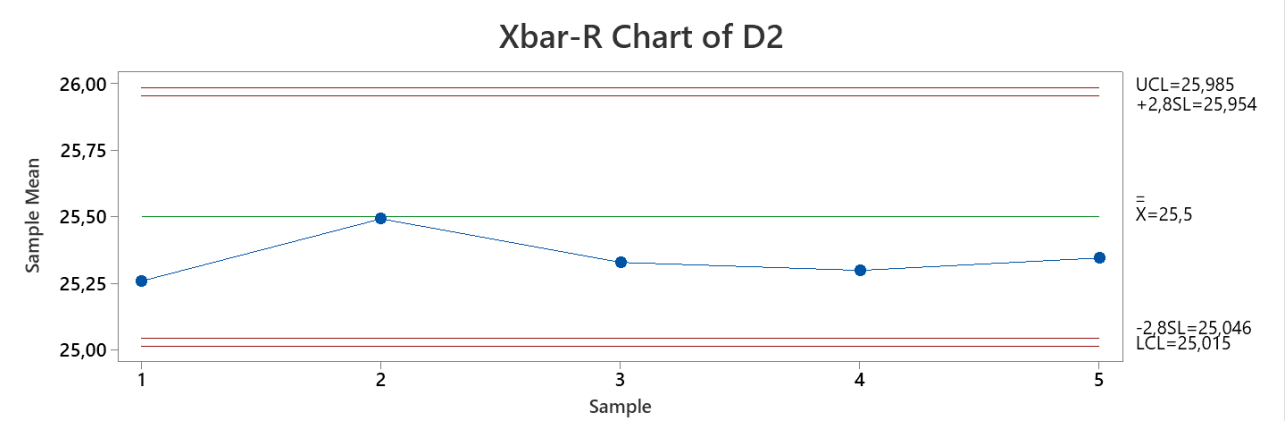
Diameter 1	Diameter 2
Xbar	Xbar
LCL =20.1536, UCL =20.8241	LCL =25.015, UCL =25.954

By applying these control charts to the provided data, no alarm is signalled, but all sample mean values of diameter 2 are below the center line, which may possibly indicate a small shift in the mean has occurred.

Diameter 1



Diameter 2



b) If no familywise correction was applied, $\alpha^* = 0.01$, and hence $\alpha = 1-(1-\alpha^*)^2 = 0{,}0199$. The corresponding Average Run Length is $ARL_0 = 50{,}25$. The expected one (with familywise correction) was $ARL_0 = 100$. Failing in using the proper correction would result is a much lower ARL.

c) The probability of detecting the shift is $P = 1 - \beta_{\bar{x},1} \cdot \beta_{\bar{x},2}$.

Let $\Delta\mu_1 = \Delta\mu_2 = 0.3\ mm$, then:

$$\beta_{\bar{x},1} = \phi\left(\frac{UCL_1 - \mu_1 - \Delta\mu_1}{\sigma_1/\sqrt{n}}\right) - \phi\left(\frac{LCL_1 - \mu_1 - \Delta\mu_1}{\sigma_1/\sqrt{n}}\right) =$$
$$\phi\left(\frac{20.8241 - 20.5 - 0.3}{0.2/\sqrt{3}}\right) - \phi\left(\frac{20.1536 - 20.5 - 0.3}{0.2/\sqrt{3}}\right) = 0.5266$$

$$\beta_{\bar{x},2} = \phi\left(\frac{UCL_2 - \mu_2 - \Delta\mu_2}{\sigma_2/\sqrt{n}}\right) - \phi\left(\frac{LCL_2 - \mu_2 - \Delta\mu_2}{\sigma_2/\sqrt{n}}\right) =$$
$$\phi\left(\frac{25.954 - 25.5 - 0.3}{0.28/\sqrt{3}}\right) - \phi\left(\frac{25.015 - 25.5 - 0.3}{0.28/\sqrt{3}}\right) = 0.8296$$

The resulting power is P =0.5166.

d) It is possible to express the power $P(n) = 1 - \beta_{\bar{x},1}(n) \cdot \beta_{\bar{x},2}(n)$ as a function of the sample size, keeping in mind that also control limits are functions of the sample size n.

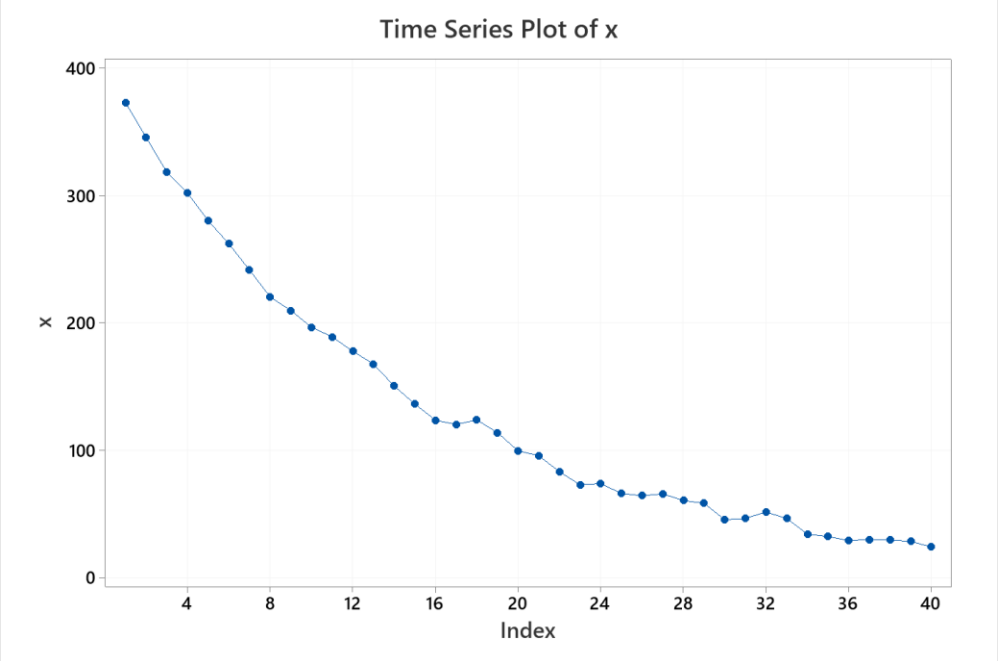
By increasing the sample size, we get:

n	UCL1	LCL1	UCL2	LCL2	beta1	beta2	P
3	20,82412444	20,17588	25,95377	25,04623	0,582746	0,829254948	0,516755
4	20,7807	20,2193	25,89298	25,10702	0,423479	0,746700187	0,683788
5	20,75106571	20,24893	25,85149	25,14851	0,292154	0,65954168	0,807312
6	20,72919059	20,27081	25,82087	25,17913	0,192907	0,572423119	0,889576
7	20,71218926	20,28781	25,79706	25,20294	0,122694	0,488937325	0,940011
8	20,69848487	20,30152	25,77788	25,22212	0,075552	0,411589674	0,968903
9	20,68713333	20,31287	25,76199	25,23801	0,045228	0,341899076	0,984537
10	20,67753027	20,32247	25,74854	25,25146	0,026408	0,280568119	0,992591

The minimum sample size to have a power P>90% is $n = 7$.

Exercise 3 (solution)

a) The cooling process for data in Table 2 is:



By fitting a model in the form $temp_t = \beta_1 \cdot e^{-0.07t} + \varepsilon_t$, we get:

EXE1

Regression Analysis: x versus exp2

Regression Equation

x = 398,79 exp2

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	398,79	1,74	229,79	0,000	1,00

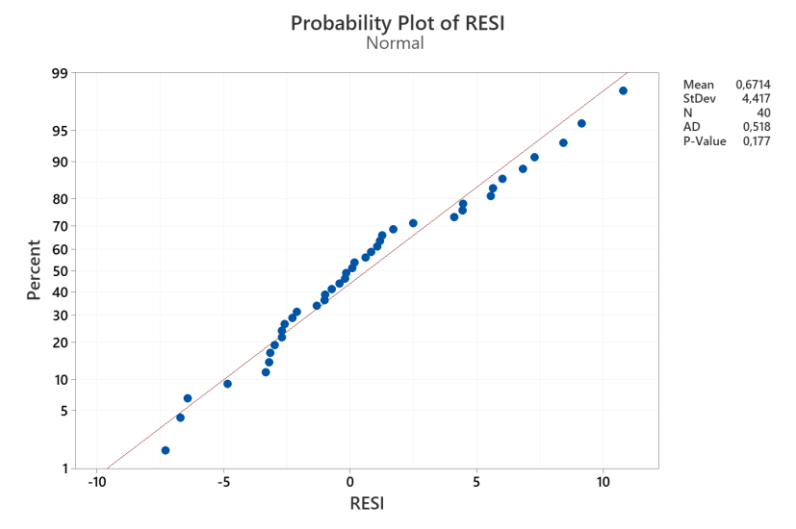
Model Summary

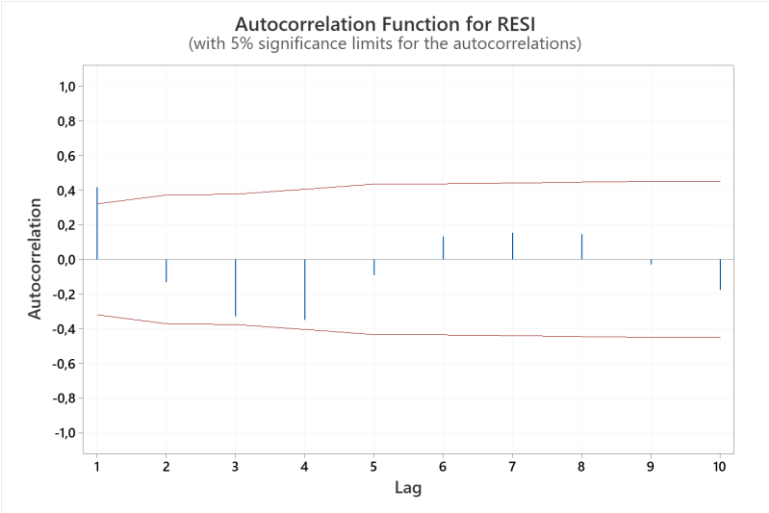
S	R-sq	R-sq(adj)	R-sq(pred)
4,46861	99,93%	99,92%	99,92%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1054364	1054364	52801,46	0,000
exp2	1	1054364	1054364	52801,46	0,000
Error	39	779	20		
Total	40	1055142			

The residuals are normal but not independent:





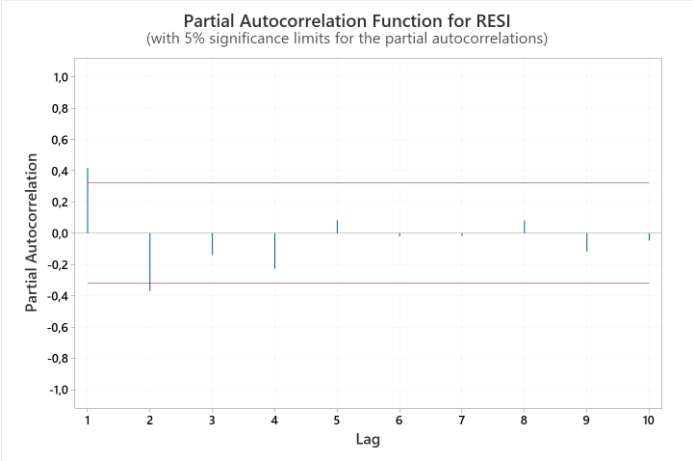
Bartlett's test at lag 1 (95% confidence):

$$|r_k| = 0.416$$

$$\frac{z_{\alpha/2}}{\sqrt{n}} = 0.354$$

The autocorrelation at lag 1 is significant.

PACF:



A more appropriate model should include an AR(1) or AR(2) term. Following the parsimony principle, we can try with an AR(1) term: $temp_t = \beta_1 \cdot e^{-0.07t} + \beta_2 temp_{t-1} + \varepsilon_t$.

By fitting this model, we get:

EXE1

Regression Analysis: x versus exp2; AR1

Method

Rows unused 1

Regression Equation

x = 214,8 exp2 + 0,430 AR1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	214,8	63,4	3,39	0,002	1354,18
AR1	0,430	0,148	2,90	0,006	1354,18

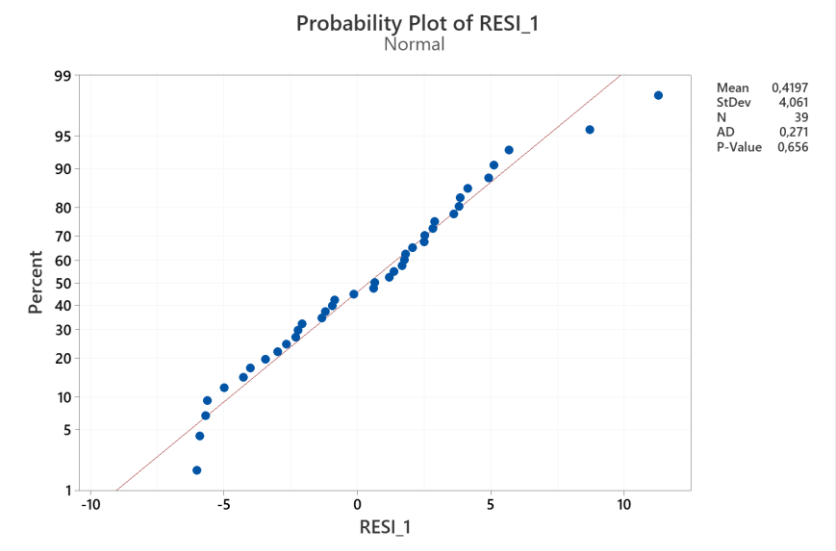
Model Summary

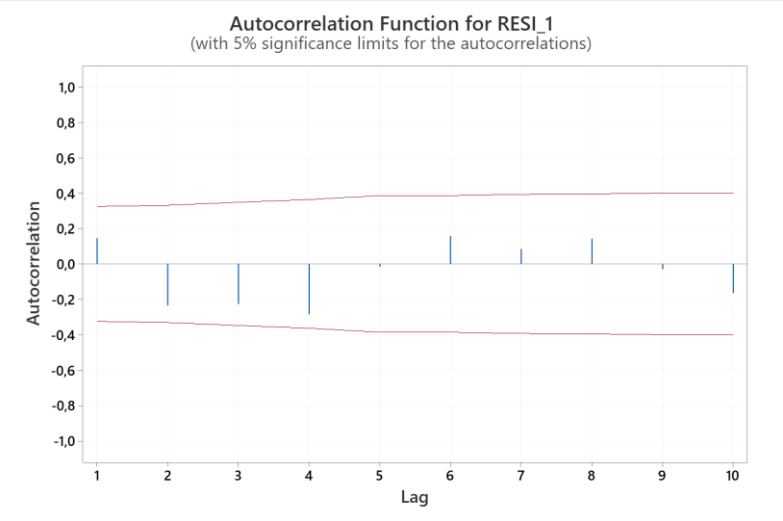
S	R-sq	R-sq(adj)	R-sq(pred)
4,13755	99,93%	99,93%	99,92%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	915380	457690	26735,23	0,000
exp2	1	196	196	11,47	0,002
AR1	1	144	144	8,40	0,006
Error	37	633	17		
Total	39	916013			

The residuals of this model are normal and independent, thus the model is adequate.





Test

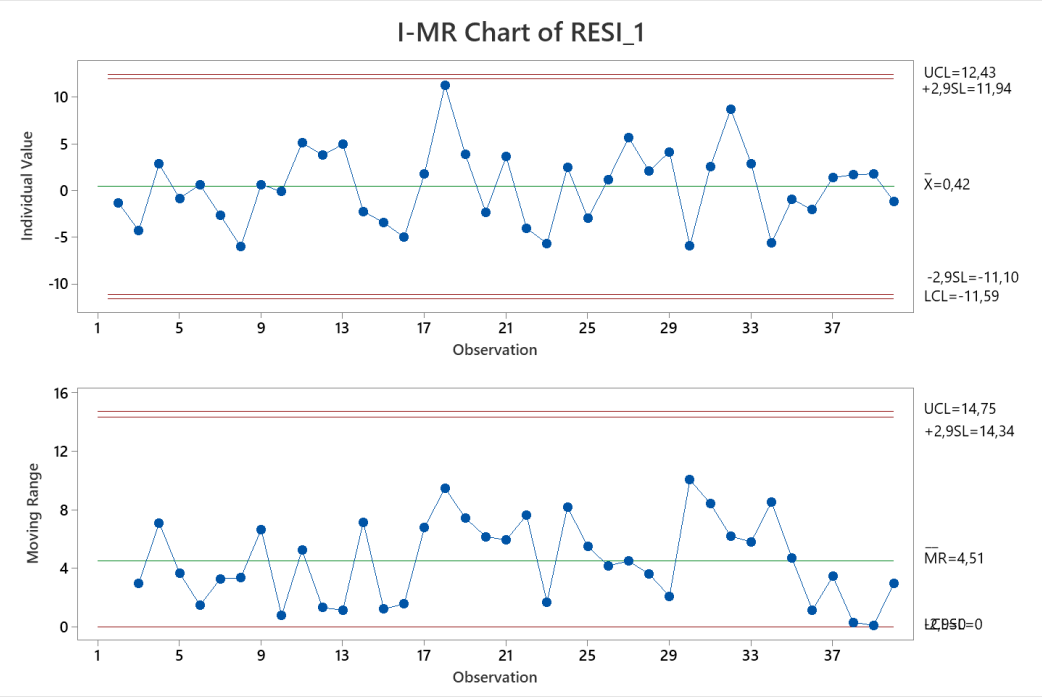
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs

Observed	Expected	P-Value
21	20,38	0,841

b) Given $ARL_0 = 250$, the Type I error is $\alpha = 0.004$, thus $K = z_{\alpha/2} = 2.878$.

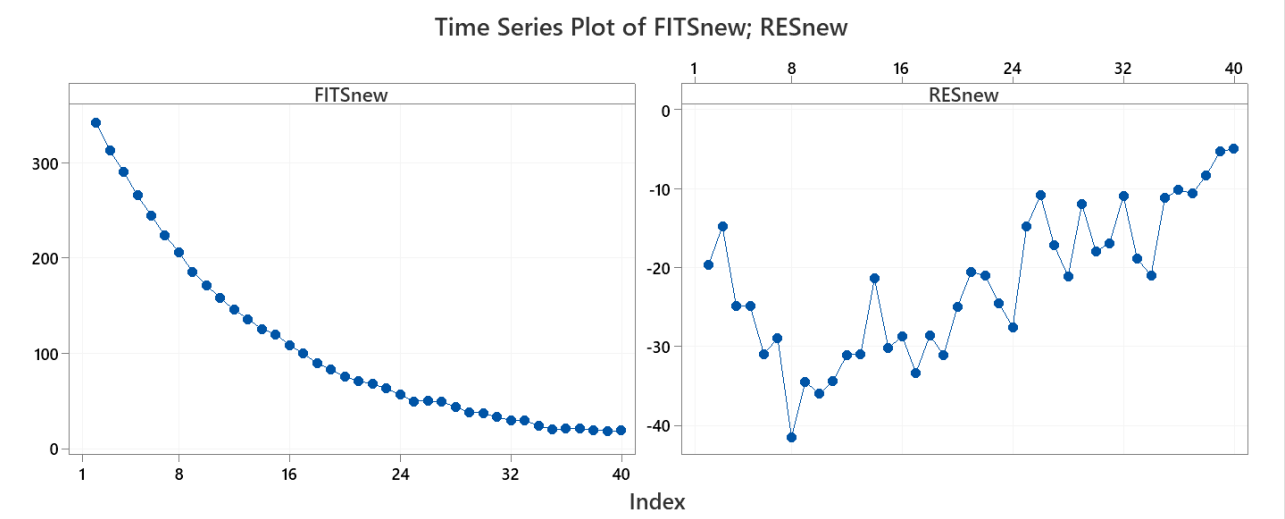
The special cause control chart for the process is the following:



The process is in-control.

c) The same model shall be applied to the new data.

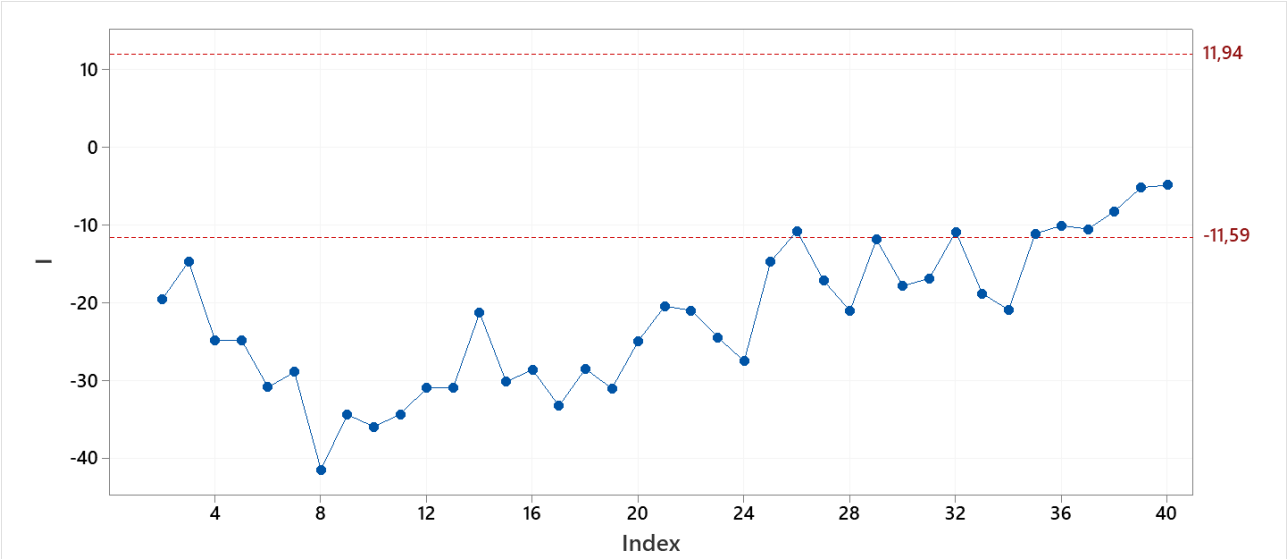
The resulting fits and residuals are the following:

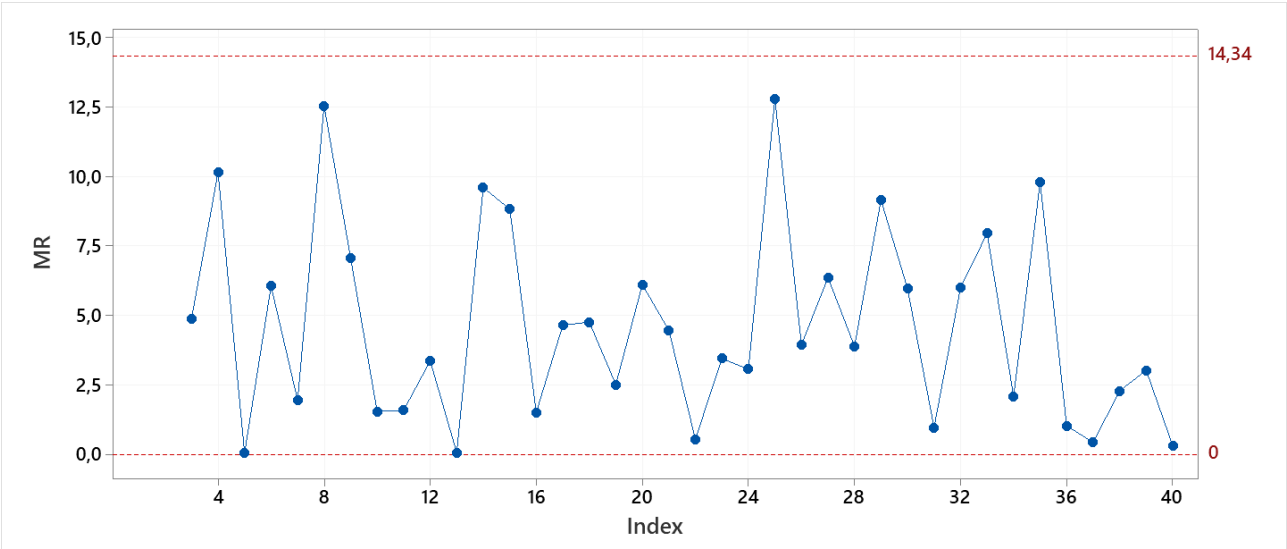


FITSnew	RESnew
342,3551	-19,6551
312,8745	-14,7745
290,5253	-24,9253
265,575	-24,875
244,6347	-30,9347
223,4831	-28,9831
206,3307	-41,5307
185,2647	-34,4647
171,5105	-36,0105
157,7202	-34,4202
145,7504	-31,0504
135,7832	-30,9832
125,6808	-21,3808
120,0156	-30,2156
108,6989	-28,6989
99,74673	-33,3467
89,48088	-28,5809
82,99672	-31,0967
75,28603	-24,986
71,01699	-20,517
67,76406	-21,0641
63,01686	-24,5169
56,58813	-27,5881
49,79664	-14,7966
49,85313	-10,8531

49,22022	-17,2202
44,01639	-21,1164
38,05787	-11,9579
37,52664	-17,9266
32,95335	-16,9534
29,74729	-10,9473
29,40532	-18,9053
24,39486	-20,9949
19,99786	-11,1979
21,06672	-10,1667
20,8013	-10,6013
19,41088	-8,31088
18,7821	-5,2821
18,867	-4,967

Using the previously designed control chart, the new cooling pattern results to be out-of-control.





Indeed, as shown in the following figure, the new cooling process is characterized by a faster decay than the previous (in-control) one.

~~d)~~ A suitable way to check whether the first and second components (referring to Table 2 and Table 3, respectively) have a cooling time series that is statistically different is to fit the same model to the two time series and check if process parameters are statistically different.

Time series 1

Regression Equation

x = 214,8 exp2 + 0,430 AR1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	214,8	63,4	3,39	0,002	1354,18
AR1	0,430	0,148	2,90	0,006	1354,18

Time series 2

Regression Equation

xnew = 9,9 exp2 + 0,8726 AR1new

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
exp2	9,9	10,7	0,92	0,363	29,35
AR1new	0,8726	0,0308	28,34	0,000	29,35

Both models have normal and independent residuals. The results of two 2-sample t tests on the model coefficients (with different variances) with a familywise confidence of 95% are:

Test on β_1:

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Sample 1	40	214,8	63,4	10
Sample 2	40	9,9	10,7	1,7

Estimation for Difference

Difference	97,5% CI for Difference
204,9	(181,2; 228,6)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
20,16	41	0,000

Test on β_2:

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Sample 1	40	0,430	0,148	0,023
Sample 2	40	0,8726	0,0308	0,0049

Estimation for Difference

Difference	97,5% CI for Difference
-0,4426	(-0,4982; -0,3870)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-18,52	42	0,000

The two cooling histories are statistically significant.

QUALITY DATA ANALYSIS

17/06/2022

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.
- 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.
- Exam duration: 2h 10min
- **Multichance students should skip: point b) in Exercise 1 and point c) in Exercise 2**

Exercise 1 (15 points)

In a metal coating process, the thickness of the coating is measured by means of a quartz microbalance. It is also known that the thickness slowly reduces over time as the cathode wears out. Table 1 shows consecutive measurements acquired every hour using the same cathode.

Table 1

Time (h)	Thickness (μm)	Time (h)	Thickness (μm)	Time (h)	Thickness (μm)	Time (h)	Thickness (μm)
1	3,58	11	4,9	21	4,4	31	1,74
2	3,68	12	5,95	22	2,2	32	5,35
3	3,32	13	4,03	23	3,58	33	2,73
4	11,56	14	3,82	24	3,97	34	2,94
5	3,86	15	3,3	25	3,48	35	1,35
6	5,02	16	6,36	26	4,44	36	4,2
7	2,67	17	2,53	27	1,9	37	3,18
8	4,09	18	2,4	28	1,79	38	3,78
9	5,94	19	3	29	6,18	39	2,45
10	4,23	20	2,48	30	1,78	40	1,06

a) Design a trend control chart for the data in Table 1 with an average run length under in-control conditions equal to $ARL_0 = 300$.
b) Using the control chart designed at point a), determine if the new observations in Table 2 are in-control or not.

Table 2

Time (h)	Thickness (μm)
41	3,28
42	3,01
43	2,25
44	1,11
45	0,86

c) Knowing that parts with a metal coating thickness lower than 1.5 μm are not conforming, use the model fitted at point a) to determine the time (in hours) after which the probability of producing non-conforming parts is at least 10%.

Exercise 2 (15 points)

During a milling process, three vibration signals are acquired by means of accelerometers mounted in three different places of the machine. For monitoring purposes, the root mean square (RMS) of each signal is computed and analyzed. Based on previous tests, it is known that under in-control milling conditions the three RMS signals follow a multivariate normal distribution with the following parameters:

$$\boldsymbol{\mu} = [11.3\ 14.61\ 12.12]'$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4.4 & 3.6 & 0.7 \\ 3.6 & 4.6 & 1.5 \\ 0.7 & 1.5 & 0.8 \end{bmatrix}$$

Table 3 shows the RMS data collected during the ten most recent milling operations.

Table 3

Signal 1 RMS	Signal 2 RMS	Signal 3 RMS
11,12	12,25	11,57
12,63	17,98	11,83
7,88	12,73	11,25
11,5	13,89	13,47
10,87	14,41	12,16
8,98	12,32	11,67
10	12,98	11,73
10,9	15,28	11,55
14,66	17,31	13,75
13,72	16,79	12,05

a) How many principal components are needed to explain at least 95% of the overall data variability? Report the eigenvalues and eigenvectors of the retained principal components (PCs).
b) Design univariate control charts on the PCs retained at point a) with a familywise type I error $\alpha = 0.01$ and determine if data in Table 3 are in-control or not.
c) Design a T^2 control chart on the PCs retained at point a) with a type I error $\alpha = 0.01$ and determine if data in Table 3 are in-control or not.
d) The head of the quality department is interested in analyzing the signal data reconstructed by applying the PCA and using the first k retained PCs (i.e., data obtained by back-transforming from the PC space to the original variable space). The aim is to evaluate to what extent the salient information enclosed in the signals is preserved. Determine the mean and variance of the reconstructed RMS of signal 1 using, respectively, $k = 1$ and $k = 3$ PCs. Discuss the result.

Exercise 3 (3 points)

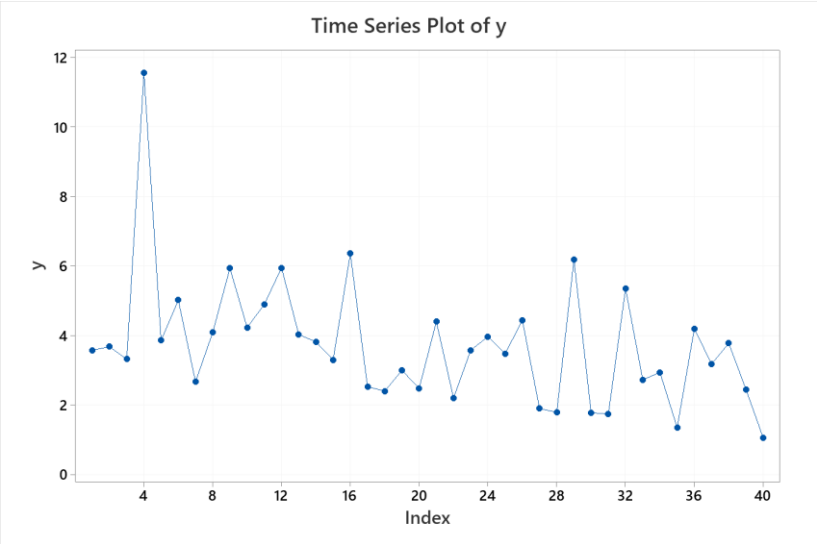
A chemical process for the production of jelly is monitored by means of a $\bar{X} - S$ control chart with known parameters. Based on historical evidence, it is known that an out-of-control increase of the mean always occurs with a simultaneous increase of the standard deviation of the process.

Determine the power of the $\bar{X} - S$ control chart in detecting a simultaneous increase of the process mean, $\mu_1 = \mu_0 + \Delta$, and of the standard deviation, $\sigma_1 = \lambda\sigma_0$, being known that: $\mu_0 = 100$, $\sigma_0 = 9.5$, $\lambda = 0.5\ \Delta$, $\Delta = 10$, $K = 3$, $n = 5$ (sample size).

Exercise 1 solution

a)

The time series plot highlights a slight decreasing trend of the coating thickness:



By fitting a trend model to these data we get:

WORKSHEET 2

Regression Analysis: y versus t

Regression Equation

y = 5,085 - 0,0661 t

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5,085	0,545	9,33	0,000	
t	-0,0661	0,0232	-2,85	0,007	1,00

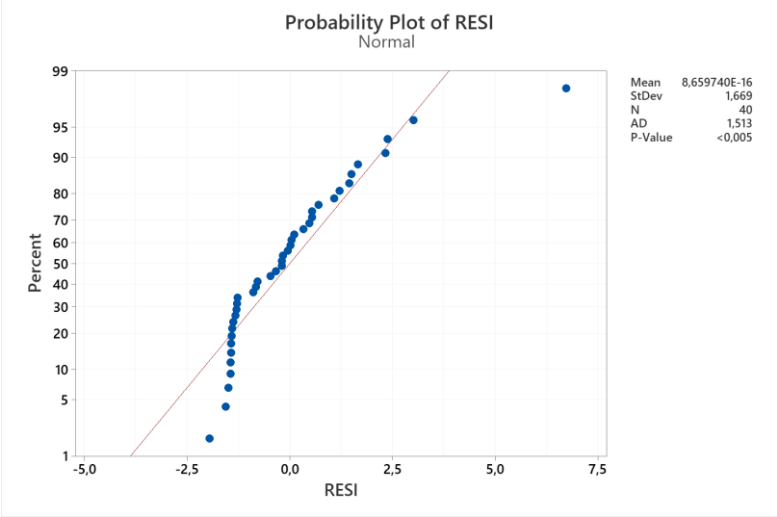
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1,69063	17,65%	15,48%	6,65%

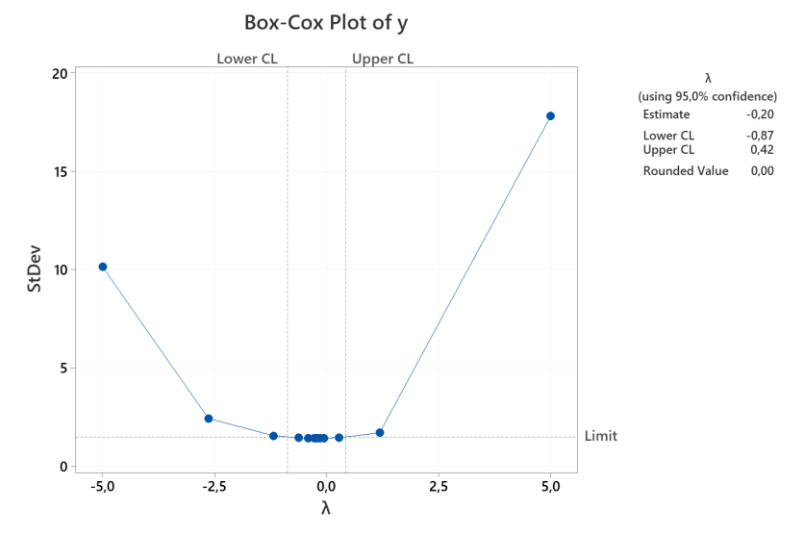
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	23,28	23,280	8,14	0,007
t	1	23,28	23,280	8,14	0,007
Error	38	108,61	2,858		
Total	39	131,89			

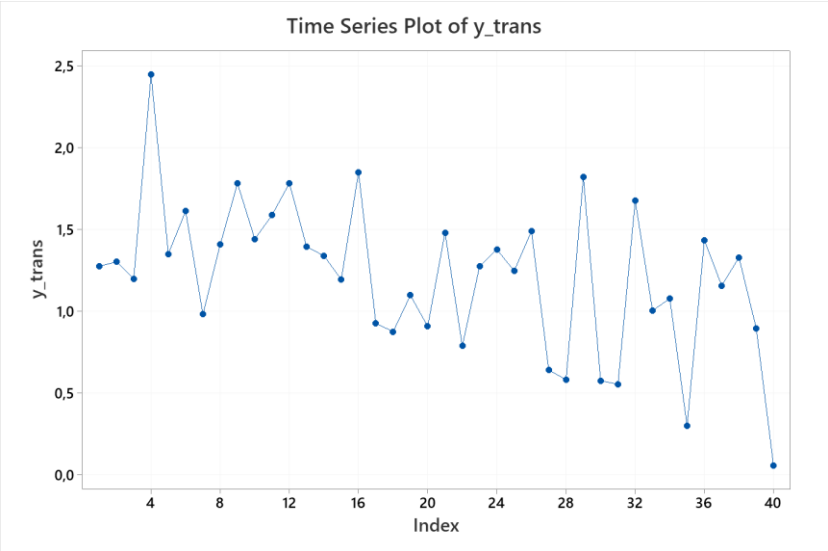
But there is a violation of the normality assumption affects the model residuals:



Such violation is caused by a skewed distribution of the measurements. It is possible to transform the data with the Box-Cox approach and then fit the trend model to the transformed data, as follows:



The data transformed with a natural logarithm transformation have the following time series pattern:



The trend model fitted on the transformed data is the following:

WORKSHEET 2

Regression Analysis: y_trans versus t

Regression Equation

y_trans = 1,602 - 0,01893 t

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1,602	0,132	12,11	0,000	
t	-0,01893	0,00562	-3,37	0,002	1,00

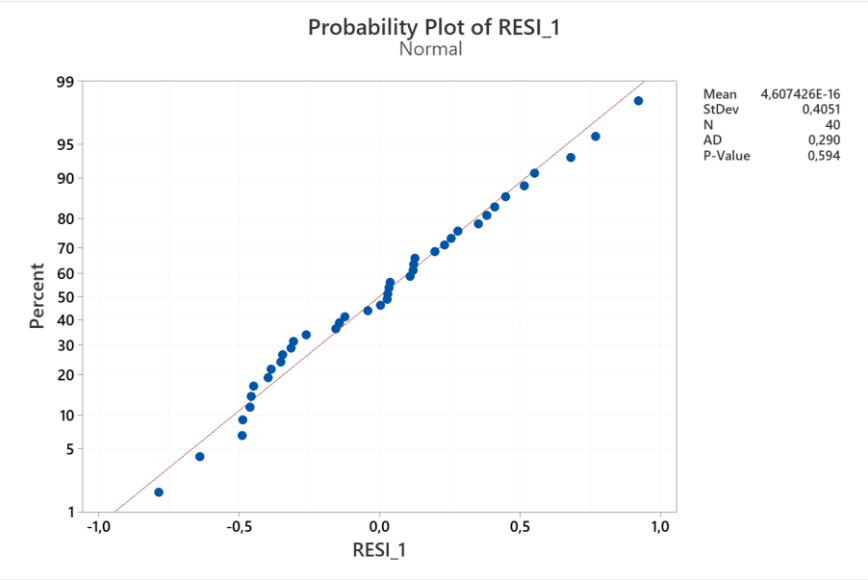
Model Summary

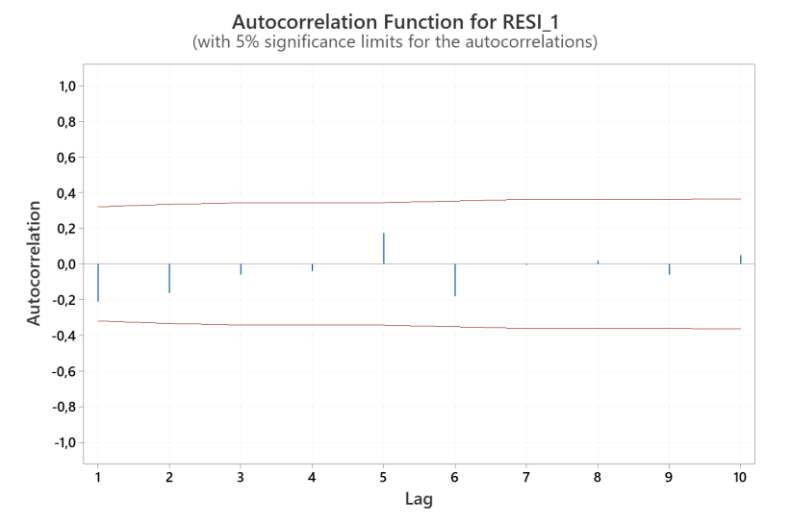
S	R-sq	R-sq(adj)	R-sq(pred)
0,410438	22,98%	20,96%	13,40%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1,910	1,9105	11,34	0,002
t	1	1,910	1,9105	11,34	0,002
Error	38	6,401	0,1685		
Total	39	8,312			

The model is significant and now the residuals meet the assumptions:





Test

Null hypothesis	H₀: The order of the data is random
Alternative hypothesis	H₁: The order of the data is not random

Number of Runs

Observed	Expected	P-Value
19	20,80	0,560

Given $ARL_0 = 300$, the type I error for the trend control chart is $\alpha = 0.0033$. The resulting control chart for the transformed data is the following:

$$UCL = b_0 + b_1 t + z_{\alpha/2} \frac{\overline{MR}}{d_2(2)}$$

$$UCL = b_0 + b_1 t$$

$$LCL = b_0 + b_1 t - z_{\alpha/2}\frac{\overline{MR}}{d_2(2)}$$

Where:

$\overline{MR} = 0{,}4764$

$d_2(2) = 1{,}128$

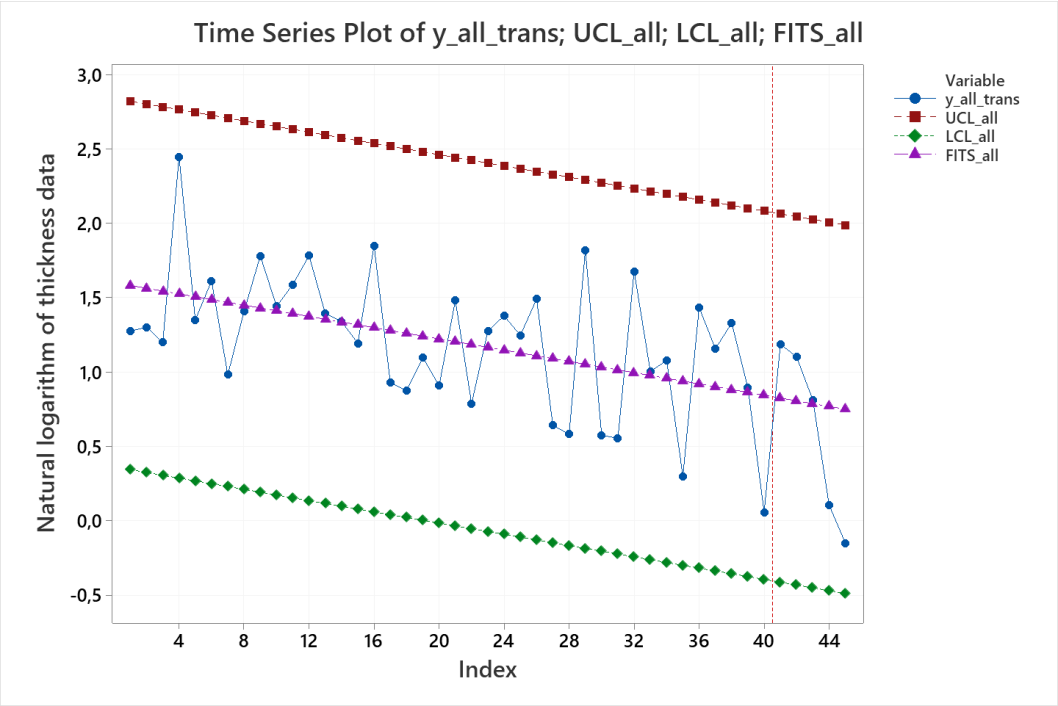
$z_{\alpha/2} = 2{,}938$

b)

Before plotting the new data onto the control chart designed in point a), we shall transform them with the natural logarithm transformation:

y_new	y_new_trans
3,28	1,18784
3,01	1,10194
2,25	0,81093
1,11	0,10436
0,86	-0,15082

The new data are in-control:



c)

Let $LSL = 1{,}5$ µm and let $\gamma \geq 10\%$ be the probability of producing a non-conforming part, then:

$$\gamma = P(y_t{}^* \leq LSL^*) = \Phi\left(\frac{LSL^* - \mu_t}{\sigma_\varepsilon}\right) \geq 0{,}1$$

Where $y_t{}^*$ is the natural logarithm of the coating thickness and LSL^* is the natural logarithm of the lower specification limit, as it results from the Box-Cox transformation.

Thus:

$$\gamma = \Phi\left(\frac{LSL^* - \mu_t}{\sigma_\varepsilon}\right) = \Phi\left(\frac{LSL^* - (b_0 + b_1 t)}{\sigma_\varepsilon}\right)$$

Where:

$LSL^* = 0{,}405$

$\sigma_\varepsilon = 0{,}41$

$b_0 = 1{,}602$

$b_1 = -0{,}01893$

We get the estimate of γ as a function of time t as shown in the table below:

t	gamma
1	0,002038
2	0,002356
3	0,002719
4	0,003131
5	0,003599
6	0,004129
7	0,004727

8	0,005401
9	0,00616
10	0,007012
11	0,007965
12	0,009031
13	0,01022
14	0,011544
15	0,013013
16	0,014642
17	0,016443
18	0,01843
19	0,020619
20	0,023023
21	0,02566
22	0,028545
23	0,031695
24	0,035126
25	0,038857
26	0,042904
27	0,047285
28	0,052018
29	0,057119
30	0,062606
31	0,068496
32	0,074804
33	0,081547
34	0,088737
35	0,096389
36	0,104516
37	0,113128
38	0,122234
39	0,131843
40	0,141961
41	0,152592
42	0,163739
43	0,175401
44	0,187576
45	0,200259
46	0,213445
47	0,227124
48	0,241283
49	0,255908
50	0,270984

Based on available model, the probability of producing at least 10% of non-conforming parts is achieved after 36 hours of coating process.

Exercise 2 (Solution)

a)

By applying the PCA on the known variance-covariance matrix, the eigenvalues (i.e., the variances of the PCs) are the following:

$\lambda_1 = 8.42364$

$\lambda_2 = 1.26052$

$\lambda_3 = 0.11584$

The first PC explains about 86% of the overall data variability. The first two PCs explains 98.8% of the overall variability. Thus, retaining the first 2 PCs is needed. Their loadings are:

u1	u2
-0,672330	-0,679682
-0,712197	0,485889
-0,201862	0,549495

b)

Being known that the scores along the first two PCs are normally distributed with:

$$\mu_{PC1} = 0, \mu_{PC2} = 0$$

$$\sigma^2_{PC1} = \lambda_1 = 8.42364,$$

$$\sigma^2_{PC2} = \lambda_2 = 1.26052$$

It is possible to design two univariate control charts for the mean of the first two PCs as follows (n=1 since we have individual observations):

PC1

$$UCL = \mu_{PC1} + K\sigma_{PC1}$$
$$CL = \mu_{PC1}$$
$$LCL = \mu_{PC1} - K\sigma_{PC1}$$

PC2

$$UCL = \mu_{PC2} + K\sigma_{PC2}$$
$$CL = \mu_{PC2}$$
$$LCL = \mu_{PC2} - K\sigma_{PC2}$$

The familywise Type I error is $\alpha = 0.01$.

The Type I error to be used in each control chart (since scores are independent by construction) is $\alpha^* = 1-(1-\alpha)^{1/2} = 0{,}005013$.

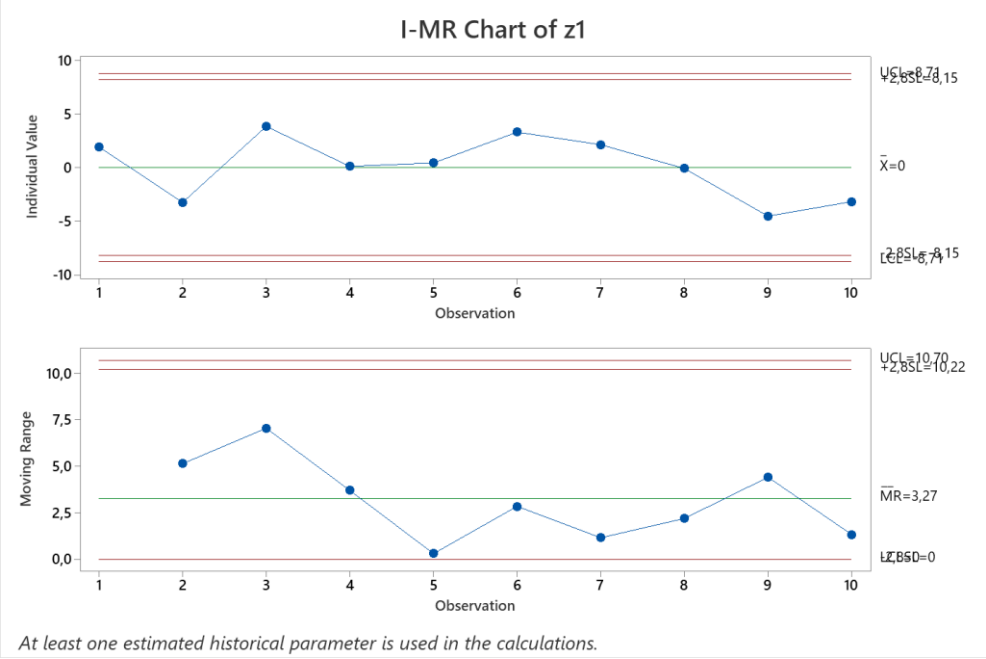
The control charts with $K = z_{\alpha^*/2} = 2.807$ have the following control limits:

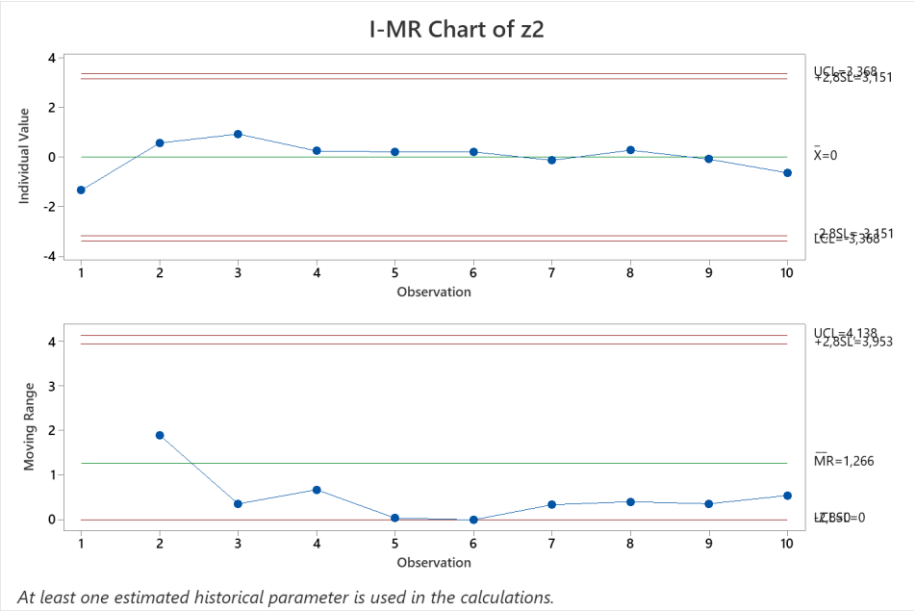
PC1		PC2	
I	MR		
LCL = -8.15, UCL = 8.15	LCL = 0, UCL = 10.22	LCL = -3.151, UCL = 3.151	LCL = 0, UCL = 3.953

The new data can be projected onto the space spanned by the first 2 PCs. The following scores are computed:

z1	z2
1,91283	-1,32658
-3,23576	0,57412
3,81392	0,93298
0,10580	0,25604
0,42347	0,21707
3,28157	0,21690
2,11364	-0,12272
-0,09318	0,28421
-4,51100	-0,07615
-3,16550	-0,62406

The control charts applied to the ten new observations are the following (ignore the additional control limits that Minitab shows, by default, at K=3).

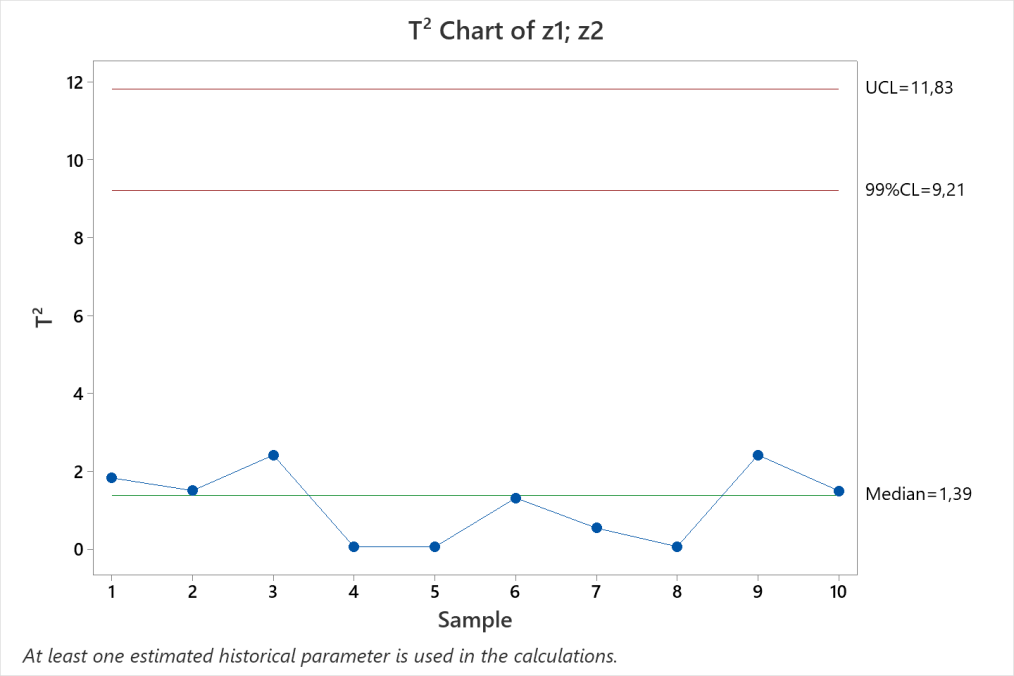




There is no violation of the control limits, although hugging is present along the second PC (which can be a symptom of a change in the process).

c)

The T^2 control chart on the scores of the first 2 PCs with known mean and variance and $\alpha = 0.01$ is:



This control chart indicates that the process is in-control according to the last ten observations.

d)

Let k=1 (only the first PC is retained). Then, the reconstructed data can be estimated as:

$$\hat{\boldsymbol{x}}_j(k) = \boldsymbol{\mu} + z_{j1}\boldsymbol{u}_1$$

For signal 1:

$$\hat{x}_{1j}(k) = \mu_1 + z_{j1}u_{11}$$

Being $\mu_{PC1} = 0, \sigma^2_{PC1} = \lambda_1 = 8.42364$, its mean and variance are:

$$E\left(\hat{x}_{1j}(k)\right) = E(\mu_1 + z_{j1}u_{11}) = \mu_1 = 11.3$$

$$V\left(\hat{x}_{1j}(k)\right) = V(\mu_1 + z_{j1}u_{11}) = \lambda_1 {u_{11}}^2 = 3.80$$

Let k=3 (no data reduction). Then, the reconstructed data can be estimated as:

$$\hat{\boldsymbol{x}}_j(k) = \boldsymbol{\mu} + z_{j1}\boldsymbol{u}_1 + z_{j2}\boldsymbol{u}_2 + z_{j3}\boldsymbol{u}_3$$

For signal 1:

$$\hat{x}_{1j}(k) = \mu_1 + z_{j1}u_{11} + z_{j2}u_{21} + z_{j3}u_{31}$$

Its mean and variance are:

$$E\left(\hat{x}_{1j}(k)\right) = E(\mu_1 + z_{j1}u_{11}) = \mu_1 = 11.3$$

$$V\left(\hat{x}_{1j}(k)\right) = V(\mu_1 + z_{j1}u_{11} + z_{j2}u_{21} + z_{j3}u_{31}) = \lambda_1 {u_{11}}^2 + \lambda_2 {u_{21}}^2 + \lambda_3 {u_{31}}^2 = 4.4 = V(x_{1j})$$

The mean of the reconstructed data is equal to the mean of the original data regardless of the number k of retained PCs.

The variance of the reconstructed data, instead, depends on the number k of retained PCs. When k=p (in this case, k=3), the reconstructed data coincide with the original data, as no dimensionality reduction is applied.

Exercise 3 (solution)

The power of the $\bar{X} - S$ control chart is:

$$P = 1 - \beta_{\bar{X}} * \beta_S$$

Where $\beta_{\bar{X}}$ is the type II error of the $\bar{X}$ control chart, whereas β_S is the type II error of the S control chart.

Let: $\mu_1 = \mu_0 + \Delta$ and $\sigma_1 = \lambda\sigma_0$, with:

- $\mu_0 = 100, \sigma_0 = 9.5$
- $\lambda = 0.5\ \Delta$
- $\Delta = 10$
- $K = 3$
- $n = 5$ (sample size)

Then:

$$\beta_{\bar{X}} = \Phi\left(\frac{\mu_0 + \frac{K\sigma_0}{\sqrt{n}} - (\mu_0 + \Delta)}{\lambda\sigma_0/\sqrt{n}}\right) - \Phi\left(\frac{\mu_0 - \frac{K\sigma_0}{\sqrt{n}} - (\mu_0 + \Delta)}{\frac{\lambda\sigma_0}{\sqrt{n}}}\right) =$$

$$\beta_{\bar{X}} = \Phi\left(\frac{\frac{K\sigma_0}{\sqrt{n}} - \Delta}{\lambda\sigma_0/\sqrt{n}}\right) - \Phi\left(\frac{-\frac{K\sigma_0}{\sqrt{n}} - \Delta}{\lambda\sigma_0/\sqrt{n}}\right) =$$

$$\beta_{\bar{X}} = \Phi\left(\frac{K}{\lambda} - \frac{\Delta\sqrt{n}}{\lambda\sigma_0}\right) - \Phi\left(-\frac{K}{\lambda} - \frac{\Delta\sqrt{n}}{\lambda\sigma_0}\right) =$$

$$\beta_{\bar{X}} = \Phi\left(\frac{3}{5} - \frac{2\sqrt{5}}{9{,}5}\right) - \Phi\left(-\frac{3}{5} - \frac{2\sqrt{5}}{9{,}5}\right) = 0{,}409$$

While:

$$\beta_S = P\left(X^2_{n-1} \leq \frac{X^2_{\alpha/2,n-1}}{\lambda^2}\right) - P\left(X^2_{n-1} \leq \frac{X^2_{1-\frac{\alpha}{2},n-1}}{\lambda^2}\right) =$$

$$\beta_S = P\left(X^2_{n-1} \leq \frac{17{,}8}{5^2}\right) - P\left(X^2_{n-1} \leq \frac{0{,}1058}{5^2}\right) = 0{,}05$$

Thus, the power of the control chart in the presence of the simultaneous shift of the mean and the standard deviation is:

$$P = 1 - 0{,}409 * 0{,}05 = 0{,}98$$

QUALITY DATA ANALYSIS

17/06/2022

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.
- 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.
- Exam duration: 2h 10min
- **Multichance students should skip: point b) in Exercise 2 and point c) in Exercise 3**

Exercise 1 (3 points)

A chemical process for the production of jelly is monitored by means of a $\bar{X} - S$ control chart with known parameters. Based on historical evidence, it is known that an out-of-control increase of the mean always occurs with a simultaneous increase of the standard deviation of the process.
Determine the power of the $\bar{X} - S$ control chart in detecting a simultaneous increase of the process mean, $\mu_1 = \mu_0 + \Delta$, and of the standard deviation, $\sigma_1 = \lambda\sigma_0$, being known that: $\mu_0 = 100$, $\sigma_0 = 9.5$, $\lambda = 0.5$ Δ, $\Delta = 5$, $K = 3$, $n = 7$ (sample size).

Exercise 2 (15 points)

In a metal coating process, the thickness of the coating is measured by means of a quartz microbalance. It is also known that the thickness slowly reduces over time as the cathode wears out. Table 1 shows consecutive measurements acquired every hour using the same cathode.

Table 1

Time (h)	Thickness (μm)	Time (h)	Thickness (μm)	Time (h)	Thickness (μm)	Time (h)	Thickness (μm)
1	8,95	11	12,25	21	11	31	4,35
2	9,2	12	14,875	22	5,5	32	13,375
3	8,3	13	10,075	23	8,95	33	6,825
4	28,9	14	9,55	24	9,925	34	7,35
5	9,65	15	8,25	25	8,7	35	3,375
6	12,55	16	15,9	26	11,1	36	10,5
7	6,675	17	6,325	27	4,75	37	7,95
8	10,225	18	6	28	4,475	38	9,45
9	14,85	19	7,5	29	15,45	39	6,125
10	10,575	20	6,2	30	4,45	40	2,65

a) Design a trend control chart for the data in Table 1 with an average run length under in-control conditions equal to $ARL_0 = 300$.
b) Using the control chart designed at point a), determine if the new observations in Table 2 are in-control or not.

Table 2

Time (h)	Thickness (μm)
41	37,7
42	27,75
43	44,55
44	61,7
45	78,9

c) Knowing that parts with a metal coating thickness lower than 3 μm are not conforming, use the model fitted at point a) to determine the time (in hours) after which the probability of producing non-conforming parts is at least 10%.

Exercise 3 (15 points)

During a milling process, three vibration signals are acquired by means of accelerometers mounted in three different places of the machine. For monitoring purposes, the root mean square (RMS) of each signal is computed and analyzed. Based on previous tests, it is known that under in-control milling conditions the three RMS signals follow a multivariate normal distribution with the following parameters:

$$\boldsymbol{\mu} = [8.3\ 11.61\ 9.12]'$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4.4 & 3.6 & 0.7 \\ 3.6 & 4.6 & 1.5 \\ 0.7 & 1.5 & 0.8 \end{bmatrix}$$

Table 3 shows the RMS data collected during the ten most recent milling operations.

Table 3

Signal 1 RMS	Signal 2 RMS	Signal 3 RMS
8,12	9,25	8,57
9,63	14,98	8,83
4,88	9,73	8,25
8,5	10,89	10,47
7,87	11,41	9,16
5,98	9,32	8,67
7	9,98	8,73
7,9	12,28	8,55
11,66	14,31	10,75
10,72	13,79	9,05

a) How many principal components are needed to explain at least 95% of the overall data variability? Report the eigenvalues and eigenvectors of the retained principal components (PCs).
b) Design univariate control charts on the PCs retained at point a) with a familywise type I error $\alpha = 0.005$ and determine if data in Table 3 are in-control or not.
c) Design a T^2 control chart on the PCs retained at point a) with a type I error $\alpha = 0.005$ and determine if data in Table 3 are in-control or not.
d) The head of the quality department is interested in analyzing the signal data reconstructed by applying the PCA and using the first k retained PCs (i.e., data obtained by back-transforming from the PC space to the original variable space). The aim is to evaluate to what extent the salient information enclosed in the signals is preserved. Determine the mean and variance of the reconstructed RMS of signal 1 using, respectively, $k = 1$ and $k = 3$ PCs. Discuss the result.

Exercise 1 (solution)

The power of the $\bar{X} - S$ control chart is:

$$P = 1 - \beta_{\bar{X}} * \beta_S$$

Where $\beta_{\bar{X}}$ is the type II error of the $\bar{X}$ control chart, whereas β_S is the type II error of the S control chart.

Let: $\mu_1 = \mu_0 + \Delta$ and $\sigma_1 = \lambda\sigma_0$, with:

- $\mu_0 = 100$, $\sigma_0 = 9.5$
- $\lambda = 0.5\ \Delta$
- $\Delta = 5$
- $K = 3$
- $n = 7$ (sample size)

Then:

$$\beta_{\bar{X}} = \Phi\left(\frac{\mu_0 + \frac{K\sigma_0}{\sqrt{n}} - (\mu_0 + \Delta)}{\lambda\sigma_0/\sqrt{n}}\right) - \Phi\left(\frac{\mu_0 - \frac{K\sigma_0}{\sqrt{n}} - (\mu_0 + \Delta)}{\frac{\lambda\sigma_0}{\sqrt{n}}}\right) =$$

$$\beta_{\bar{X}} = \Phi\left(\frac{\frac{K\sigma_0}{\sqrt{n}} - \Delta}{\lambda\sigma_0/\sqrt{n}}\right) - \Phi\left(\frac{-\frac{K\sigma_0}{\sqrt{n}} - \Delta}{\lambda\sigma_0/\sqrt{n}}\right) =$$

$$\beta_{\bar{X}} = \Phi\left(\frac{K}{\lambda} - \frac{\Delta\sqrt{n}}{\lambda\sigma_0}\right) - \Phi\left(-\frac{K}{\lambda} - \frac{\Delta\sqrt{n}}{\lambda\sigma_0}\right) =$$

$$\beta_{\bar{X}} = \Phi\left(\frac{3}{2.5} - \frac{2\sqrt{7}}{9.5}\right) - \Phi\left(-\frac{3}{2.5} - \frac{2\sqrt{7}}{9.5}\right) = 0{,}7$$

While:

$$\beta_S = P\left(X^2_{n-1} \leq \frac{X^2_{\alpha/2,n-1}}{\lambda^2}\right) - P\left(X^2_{n-1} \leq \frac{X^2_{1-\frac{\alpha}{2},n-1}}{\lambda^2}\right) =$$

$$\beta_S = P\left(X^2_{n-1} \leq \frac{21.74}{2.5^2}\right) - P\left(X^2_{n-1} \leq \frac{0.4234}{2.5^2}\right) = 0{,}253$$

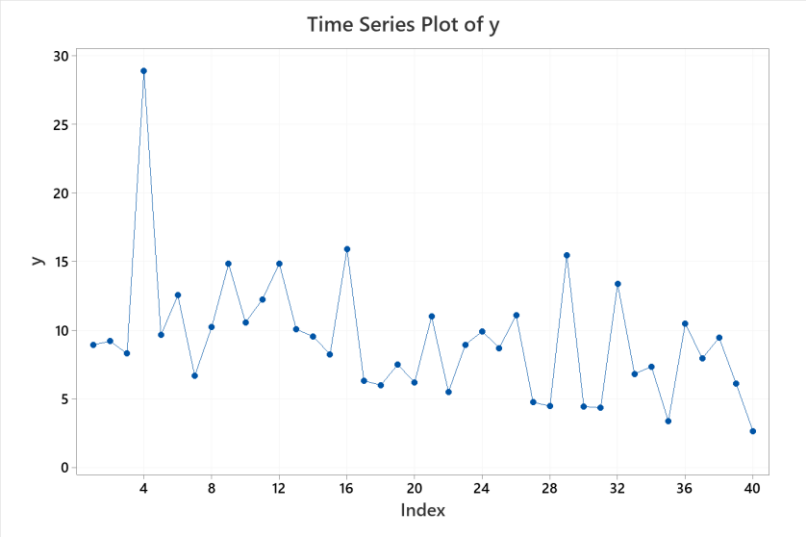
Thus, the power of the control chart in the presence of the simultaneous shift of the mean and the standard deviation is:

$$P = 1 - 0{,}7 * 0{,}253 = 0{,}82$$

Exercise 2 solution

a)

The time series plot highlights a slight decreasing trend of the coating thickness:



By fitting a trend model to these data we get:

EXE1

Regression Analysis: y versus t

Regression Equation

y = 12,71 - 0,1652 t

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	12,71	1,36	9,33	0,000	
t	-0,1652	0,0579	-2,85	0,007	1,00

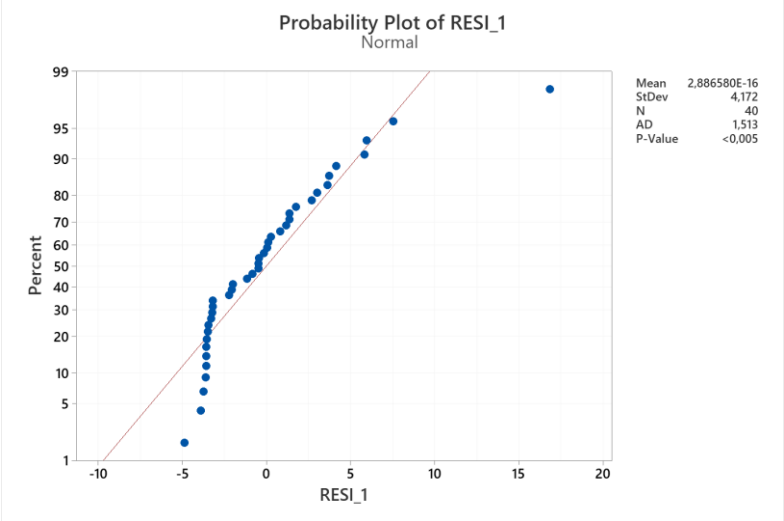
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4,22657	17,65%	15,48%	6,65%

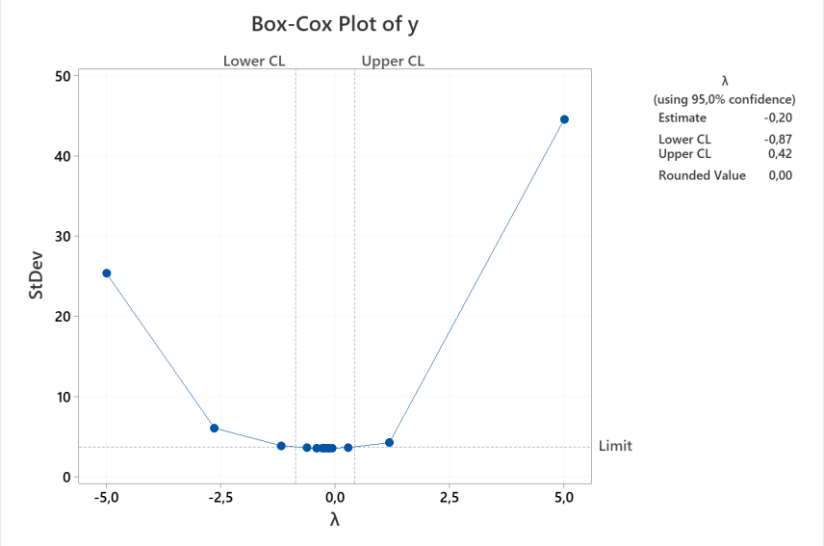
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	145,5	145,50	8,14	0,007
t	1	145,5	145,50	8,14	0,007
Error	38	678,8	17,86		
Total	39	824,3			

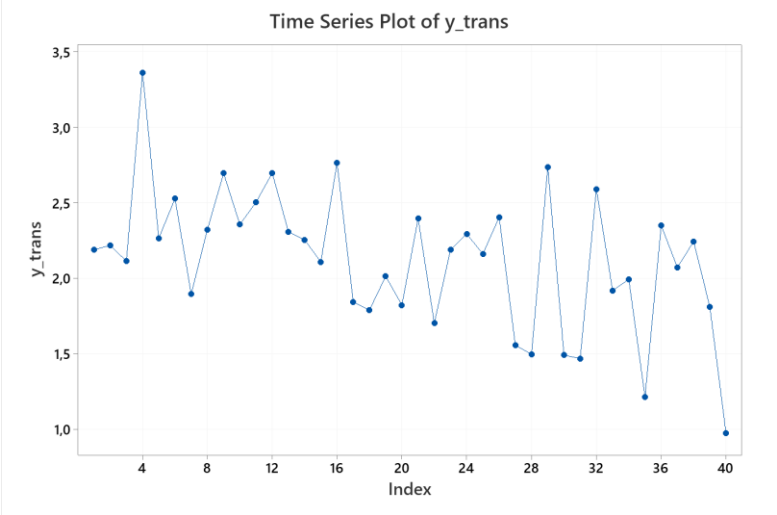
But a violation of the normality assumption affects the model residuals:



Such violation is caused by a skewed distribution of the measurements. It is possible to transform the data with the Box-Cox approach and then fit the trend model to the transformed data, as follows:



The data transformed with a natural logarithm transformation have the following time series pattern:



The trend model fitted on the transformed data is the following:

EXE1

Regression Analysis: y_trans versus t

Regression Equation

y_trans = 2,518 - 0,01893 t

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2,518	0,132	19,04	0,000	
t	-0,01893	0,00562	-3,37	0,002	1,00

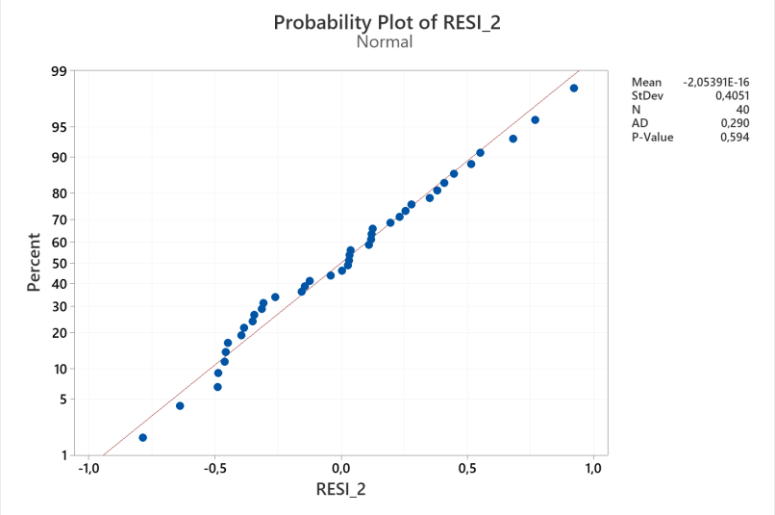
Model Summary

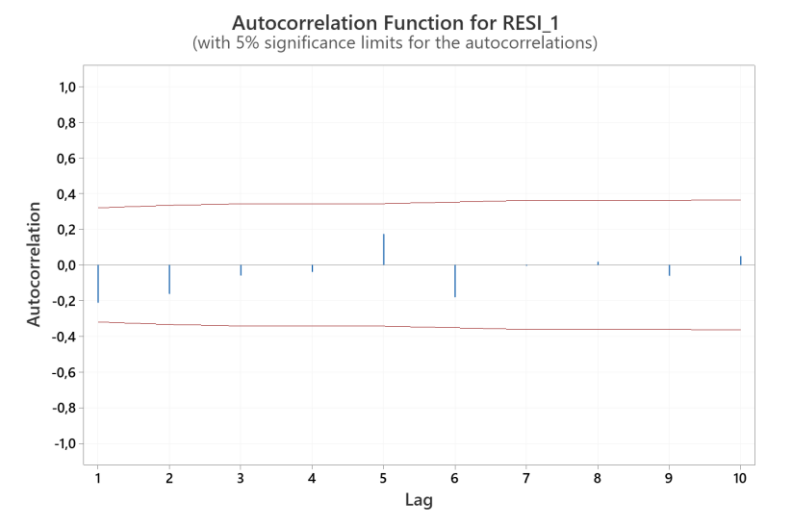
S	R-sq	R-sq(adj)	R-sq(pred)
0,410438	22,98%	20,96%	13,40%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1,910	1,9105	11,34	0,002
t	1	1,910	1,9105	11,34	0,002
Error	38	6,401	0,1685		
Total	39	8,312			

The model is significant and now the residuals meet the assumptions:





Test

Null hypothesis	H₀: The order of the data is random
Alternative hypothesis	H₁: The order of the data is not random

Number of Runs

Observed	Expected	P-Value
19	20,80	0,560

Given $ARL_0 = 300$, the type I error for the trend control chart is $\alpha = 0.0033$. The resulting control chart for the transformed data is the following:

$$UCL = b_0 + b_1 t + z_{\alpha/2} \frac{\overline{MR}}{d_2(2)}$$

$$UCL = b_0 + b_1 t$$

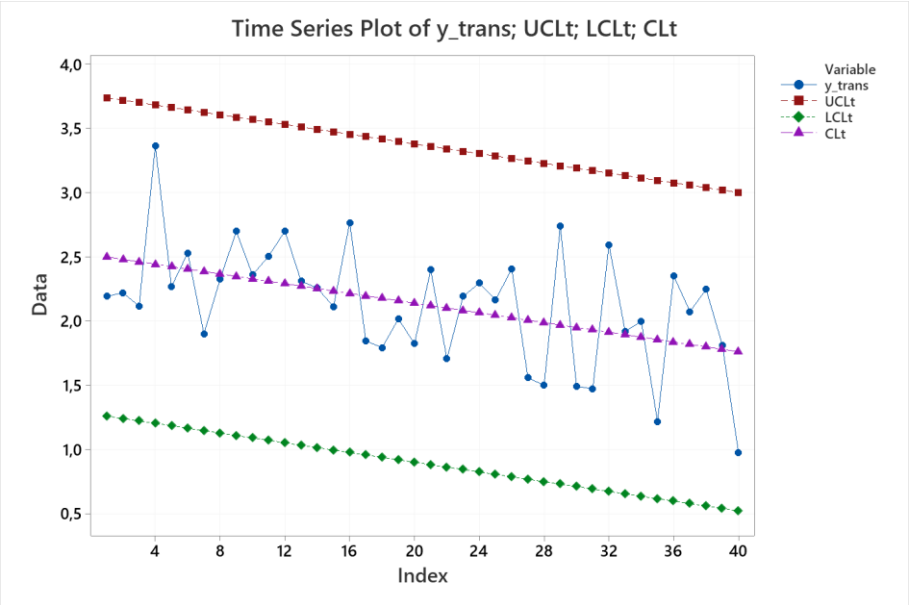
$$LCL = b_0 + b_1 t - z_{\alpha/2} \frac{\overline{MR}}{d_2(2)}$$

Where:

$\overline{MR} = 0{,}4764$

$d_2(2) = 1{,}128$

$z_{\alpha/2} = 2{,}935$

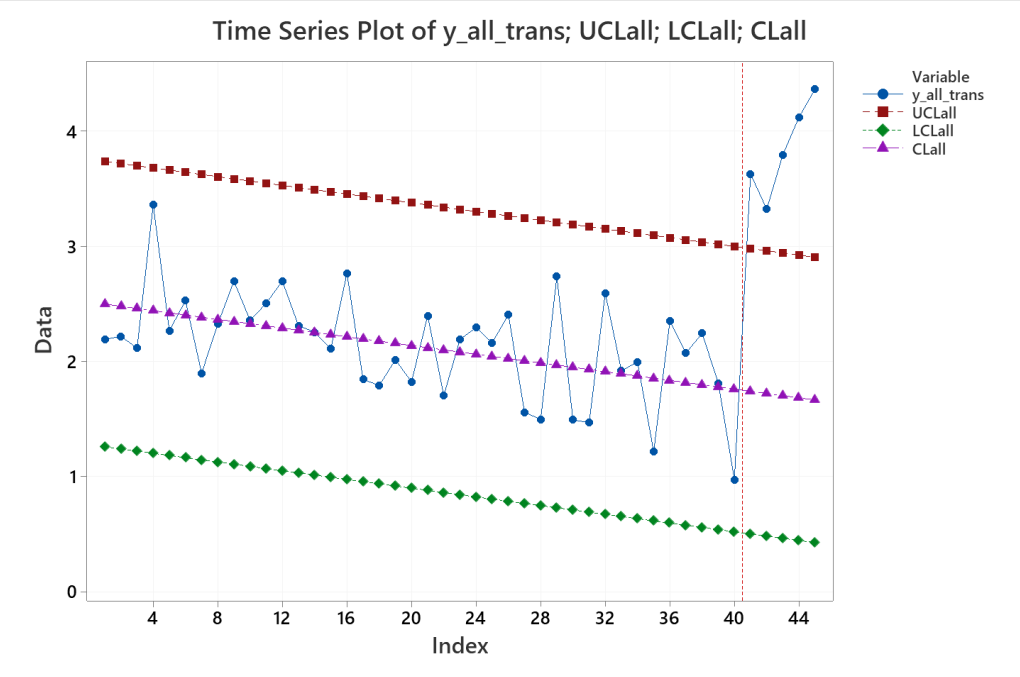


b)

Before plotting the new data onto the control chart designed in point a), we shall transform them with the natural logarithm transformation:

ynew	ynew trans
37,7	3,62966
27,75	3,323236
44,55	3,796612
61,7	4,122284
78,9	4,368181

The new data are out-of-control:



c)

Let $LSL = 3$ µm and let $\gamma \geq 10\%$ be the probability of producing a non-conforming part, then:

$$\gamma = P({y_t}^* \leq LSL^*) = \Phi\left(\frac{LSL^* - \mu_t}{\sigma_\varepsilon}\right) \geq 0{,}1$$

Where ${y_t}^*$is the natural logarithm of the coating thickness and LSL^* is the natural logarithm of the lower specification limit, as it results from the Box-Cox transformation.

Thus:

$$\gamma = \Phi\left(\frac{LSL^* - \mu_t}{\sigma_\varepsilon}\right) = \Phi\left(\frac{LSL^* - (b_0 + b_1 t)}{\sigma_\varepsilon}\right)$$

Where:

$LSL^* = 1.0986$

$\sigma_\varepsilon = 0{,}41$

$b_0 = 2{,}518$

$b_1 = -0{,}01893$

We get the estimate of γ as a function of time t as shown in the table below:

t	gamma
1	0,000318032
2	0,000376414
3	0,000444622
4	0,000524139
5	0,000616644
6	0,000724028
7	0,00084842
8	0,000992206
9	0,001158057
10	0,00134895
11	0,0015682
12	0,001819484
13	0,002106866
14	0,002434835
15	0,002808325
16	0,003232748
17	0,003714023
18	0,004258605
19	0,004873508
20	0,005566335
21	0,006345298
22	0,007219242
23	0,008197661
24	0,009290711
25	0,010509223
26	0,011864706
27	0,013369346
28	0,015036
29	0,016878182

30	0,018910044
31	0,021146344
32	0,023602412
33	0,026294103
34	0,029237739
35	0,03245005
36	0,035948094
37	0,039749178
38	0,043870759
39	0,048330348
40	0,053145389
41	0,058333145
42	0,063910567
43	0,069894157
44	0,076299826
45	0,083142747
46	0,090437203
47	0,098196433
48	0,106432479
49	0,11515603
50	0,124376267

Based on available model, the probability of producing at least 10% of non-conforming parts is achieved after 48 hours of coating process.

Exercise 3 (Solution)

By applying the PCA on the known variance-covariance matrix, the eigenvalues (i.e., the variances of the PCs) are the following:

$\lambda_1 = 8.42364$

$\lambda_2 = 1.26052$

$\lambda_3 = 0.11584$

The first PC explains about 86% of the overall data variability. The first two PCs explains 98.8% of the overall variability. Thus, retaining the first 2 PCs is needed. Their loadings are:

u1	u2
-0,672330	-0,679682
-0,712197	0,485889
-0,201862	0,549495

b)

Being known that the scores along the first two PCs are normally distributed with:

$$\mu_{PC1} = 0, \mu_{PC2} = 0$$

$$\sigma^2_{PC1} = \lambda_1 = 8.42364,$$

$$\sigma^2_{PC2} = \lambda_2 = 1.26052$$

It is possible to design two univariate control charts for the mean of the first two PCs as follows (n=1 since we have individual observations):

PC1

$$UCL = \mu_{PC1} + K\sigma_{PC1}$$
$$CL = \mu_{PC1}$$
$$LCL = \mu_{PC1} - K\sigma_{PC1}$$

PC2

$$UCL = \mu_{PC2} + K\sigma_{PC2}$$
$$CL = \mu_{PC2}$$
$$LCL = \mu_{PC2} - K\sigma_{PC2}$$

The familywise Type I error is $\alpha = 0.005$.

The Type I error to be used in each control chart (since scores are independent by construction) is $\alpha^* = 1-(1-\alpha)^{1/2} = 0{,}002503$.

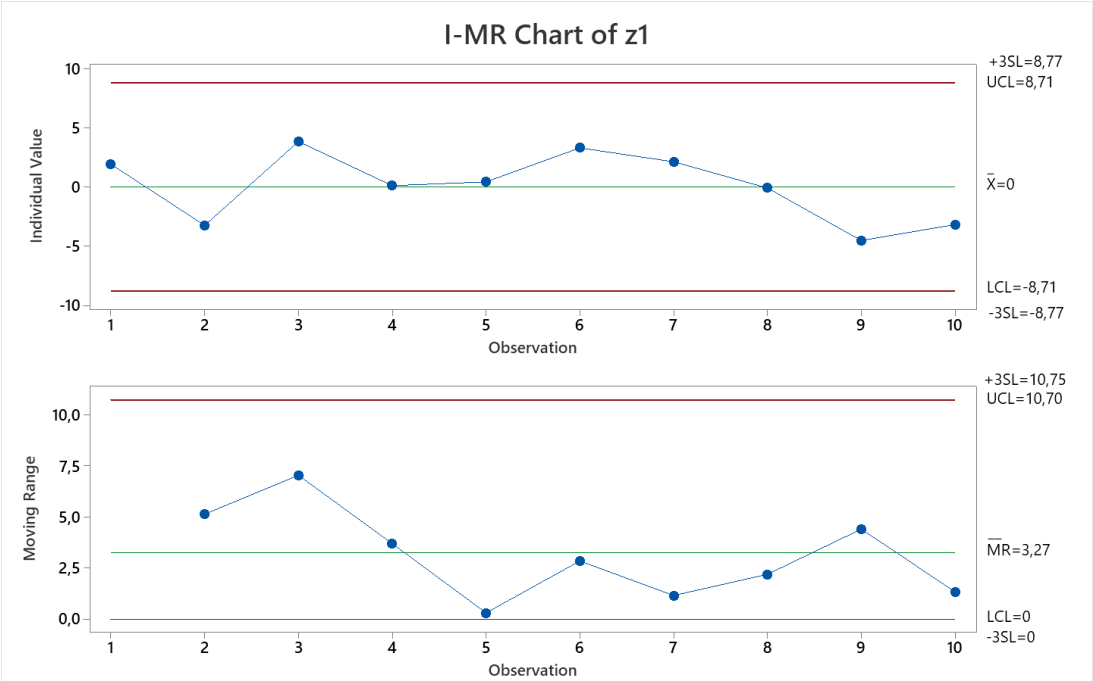
The control charts with $K = z_{\alpha^*/2} = 3.023$ have the following control limits:

PC1		PC2	
I	MR		
LCL = -8.77, UCL = 8.77	LCL = 0, UCL = 10.75	LCL = -3.394, UCL = 3.394	LCL = 0, UCL = 4.160

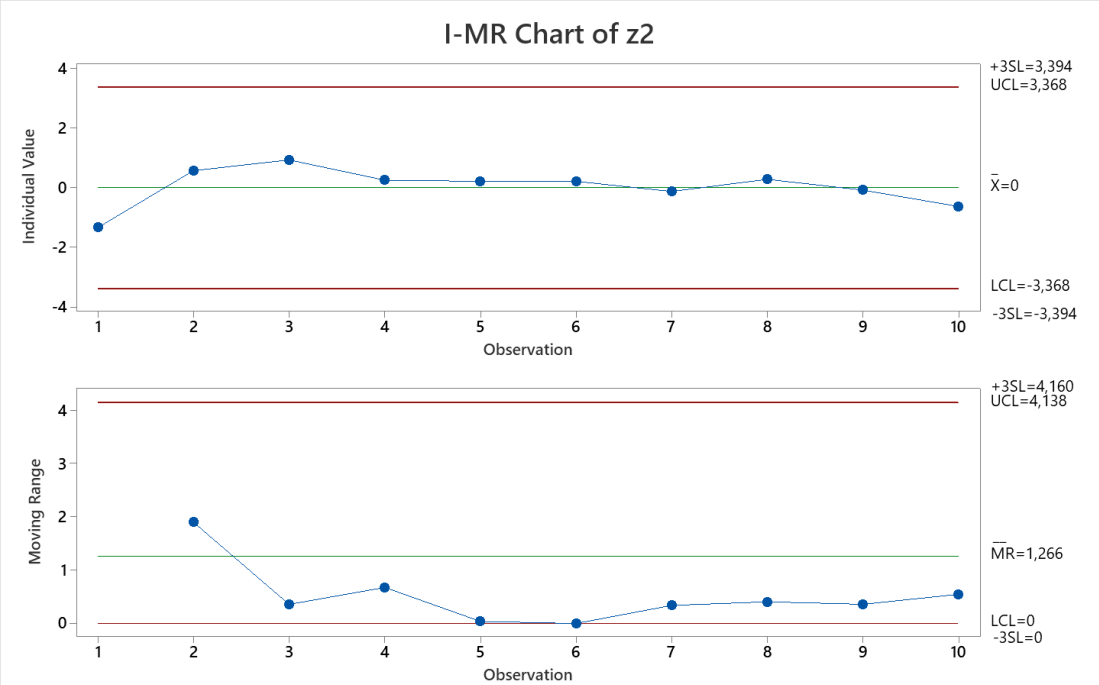
The new data can be projected onto the space spanned by the first 2 PCs. The following scores are computed:

z1	z2
1,91283	-1,32658
-3,23576	0,57412
3,81392	0,93298
0,10580	0,25604
0,42347	0,21707
3,28157	0,21690
2,11364	-0,12272
-0,09318	0,28421
-4,51100	-0,07615
-3,16550	-0,62406

The control charts applied to the ten new observations are the following:



At least one estimated historical parameter is used in the calculations.

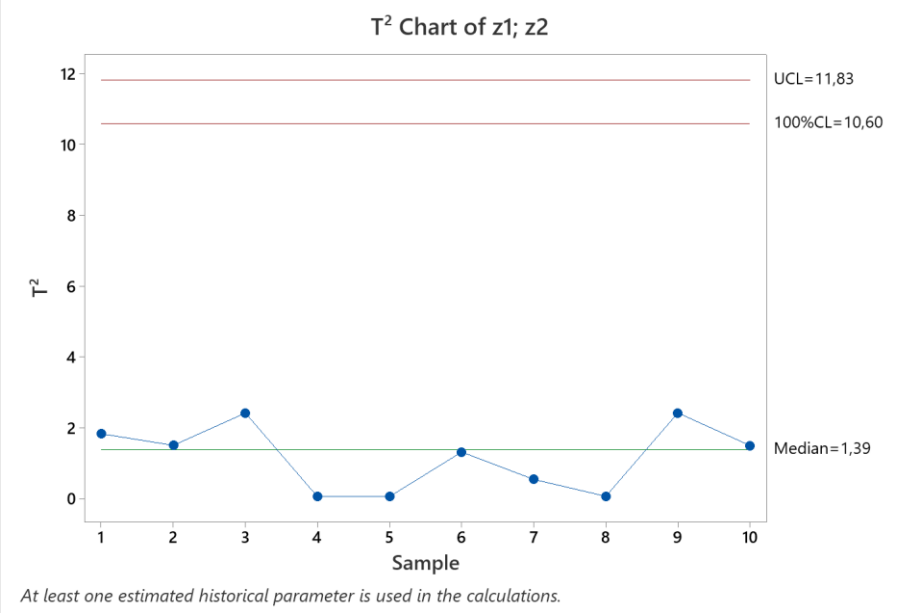


At least one estimated historical parameter is used in the calculations.

There is no violation of the control limits, although hugging is present along the second PC (which can be a symptom of a change in the process).

c)

The T^2 control chart on the scores of the first 2 PCs with known mean and variance and $\alpha = 0.005$ is:



This control chart indicates that the process is in-control according to the last ten observations.

d)

Let k=1 (only the first PC is retained). Then, the reconstructed data can be estimated as:

$$\hat{\boldsymbol{x}}_j(k) = \boldsymbol{\mu} + z_{j1}\boldsymbol{u}_1$$

For signal 1:

$$\hat{x}_{1j}(k) = \mu_1 + z_{j1}u_{11}$$

Being $\mu_{PC1} = 0, \sigma^2_{PC1} = \lambda_1 = 8.42364$, its mean and variance are:

$$E\left(\hat{x}_{1j}(k)\right) = E(\mu_1 + z_{j1}u_{11}) = \mu_1 = 8.3$$

$$V\left(\hat{x}_{1j}(k)\right) = V(\mu_1 + z_{j1}u_{11}) = \lambda_1 {u_{11}}^2 = 3.80$$

Let k=3 (no data reduction). Then, the reconstructed data can be estimated as:

$$\hat{\boldsymbol{x}}_j(k) = \boldsymbol{\mu} + z_{j1}\boldsymbol{u}_1 + z_{j2}\boldsymbol{u}_2 + z_{j3}\boldsymbol{u}_3$$

For signal 1:

$$\hat{x}_{1j}(k) = \mu_1 + z_{j1}u_{11} + z_{j2}u_{21} + z_{j3}u_{31}$$

Its mean and variance are:

$$E\left(\hat{x}_{1j}(k)\right) = E(\mu_1 + z_{j1}u_{11}) = \mu_1 = 8.3$$

$$V\left(\hat{x}_{1j}(k)\right) = V(\mu_1 + z_{j1}u_{11} + z_{j2}u_{21} + z_{j3}u_{31}) = \lambda_1 {u_{11}}^2 + \lambda_2 {u_{21}}^2 + \lambda_3 {u_{31}}^2 = 4.4 = V(x_{1j})$$

The mean of the reconstructed data is equal to the mean of the original data regardless of the number k of retained PCs.

The variance of the reconstructed data, instead, depends on the number k of retained PCs. When k=p (in this case, k=3), the reconstructed data coincide with the original data, as no dimensionality reduction is applied.

QUALITY DATA ANALYSIS

17/06/2022

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.
- 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.
- Exam duration: 2h 10min
- **Multichance students should skip: point c) in Exercise 2 and point b) in Exercise 3**

Exercise 1 (3 points)

A chemical process for the production of jelly is monitored by means of a $\bar{X} - S$ control chart with known parameters. Based on historical evidence, it is known that an out-of-control increase of the mean always occurs with a simultaneous increase of the standard deviation of the process.

Determine the power of the $\bar{X} - S$ control chart in detecting a simultaneous increase of the process mean, $\mu_1 = \mu_0 + \Delta$, and of the standard deviation, $\sigma_1 = \lambda\sigma_0$, being known that: $\mu_0 = 80$, $\sigma_0 = 5$, $\lambda = 0.5\Delta$, $\Delta = 3$, $K = 3$, $n = 7$ (sample size).

Exercise 2 (15 points)

During a milling process, three vibration signals are acquired by means of accelerometers mounted in three different places of the machine. For monitoring purposes, the root mean square (RMS) of each signal is computed and analyzed. Based on previous tests, it is known that under in-control milling conditions the three RMS signals follow a multivariate normal distribution with the following parameters:

$$\boldsymbol{\mu} = [11.3\ 14.61\ \ 12.12]'$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4.4 & 3.6 & 1.1 \\ 3.6 & 4.6 & 1.5 \\ 1.1 & 1.5 & 0.8 \end{bmatrix}$$

Table 1 shows the RMS data collected during the ten most recent milling operations.

Table 1

Signal 1 RMS	Signal 2 RMS	Signal 3 RMS
11,12	12,25	11,57
12,63	17,98	11,83
7,88	12,73	11,25
11,5	13,89	13,47
10,87	14,41	12,16
8,98	12,32	11,67
10	12,98	11,73
10,9	15,28	11,55
14,66	17,31	13,75
13,72	16,79	12,05

a) How many principal components are needed to explain at least 95% of the overall data variability? Report the eigenvalues and eigenvectors of the retained principal components (PCs).
b) Design univariate control charts on the PCs retained at point a) with a familywise type I error $\alpha = 0.005$ and determine if data in Table 1 are in-control or not.
c) Design a T^2 control chart on the PCs retained at point a) with a type I error $\alpha = 0.005$ and determine if data in Table 1 are in-control or not.
d) The head of the quality department is interested in analyzing the signal data reconstructed by applying the PCA and using the first k retained PCs (i.e., data obtained by back-transforming from the PC space to the original variable space). The aim is to evaluate to what extent the salient information enclosed in the signals is preserved. Determine the mean and variance of the reconstructed RMS of signal 1 using, respectively, $k = 1$ and $k = 3$ PCs. Discuss the result.

Exercise 3 (15 points)

In a metal coating process, the thickness of the coating is measured by means of a quartz microbalance. It is also known that the thickness slowly reduces over time as the cathode wears out. Table 2 shows consecutive measurements acquired every hour using the same cathode.

Table 2

Time (h)	Thickness (μm)	Time (h)	Thickness (μm)	Time (h)	Thickness (μm)	Time (h)	Thickness (μm)
1	8,95	11	12,25	21	11	31	4,35
2	9,2	12	14,875	22	5,5	32	13,375
3	8,3	13	10,075	23	8,95	33	6,825
4	28,9	14	9,55	24	9,925	34	7,35
5	9,65	15	8,25	25	8,7	35	3,375
6	12,55	16	15,9	26	11,1	36	10,5
7	6,675	17	6,325	27	4,75	37	7,95
8	10,225	18	6	28	4,475	38	9,45
9	14,85	19	7,5	29	15,45	39	6,125
10	10,575	20	6,2	30	4,45	40	2,65

a) Design a trend control chart for the data in Table 2 with an average run length under in-control conditions equal to $ARL_0 = 300$.
b) Using the control chart designed at point a), determine if the new observations in Table 3 are in-control or not.

Table 3

Time (h)	Thickness (μm)
41	25,12
42	23,65
43	28,11
44	20,98
45	27,70

c) Knowing that parts with a metal coating thickness lower than 3 μm are not conforming, use the model fitted at point a) to determine the time (in hours) after which the probability of producing non-conforming parts is at least 10%.

Exercise 1 (solution)

The power of the $\bar{X} - S$ control chart is:

$$P = 1 - \beta_{\bar{X}} * \beta_S$$

Where $\beta_{\bar{X}}$ is the type II error of the $\bar{X}$ control chart, whereas β_S is the type II error of the S control chart.

Let: $\mu_1 = \mu_0 + \Delta$ and $\sigma_1 = \lambda\sigma_0$, with:

- $\mu_0 = 80$, $\sigma_0 = 5$
- $\lambda = 0.5\ \Delta$
- $\Delta = 3$
- $K = 3$
- $n = 7$ (sample size)

Then:

$$\beta_{\bar{X}} = \Phi\left(\frac{\mu_0 + \frac{K\sigma_0}{\sqrt{n}} - (\mu_0 + \Delta)}{\lambda\sigma_0/\sqrt{n}}\right) - \Phi\left(\frac{\mu_0 - \frac{K\sigma_0}{\sqrt{n}} - (\mu_0 + \Delta)}{\frac{\lambda\sigma_0}{\sqrt{n}}}\right) =$$

$$\beta_{\bar{X}} = \Phi\left(\frac{\frac{K\sigma_0}{\sqrt{n}} - \Delta}{\lambda\sigma_0/\sqrt{n}}\right) - \Phi\left(\frac{-\frac{K\sigma_0}{\sqrt{n}} - \Delta}{\lambda\sigma_0/\sqrt{n}}\right) =$$

$$\beta_{\bar{X}} = \Phi\left(\frac{K}{\lambda} - \frac{\Delta\sqrt{n}}{\lambda\sigma_0}\right) - \Phi\left(-\frac{K}{\lambda} - \frac{\Delta\sqrt{n}}{\lambda\sigma_0}\right) =$$

$$\beta_{\bar{X}} = \Phi\left(\frac{3}{1.5} - \frac{2\sqrt{7}}{5}\right) - \Phi\left(-\frac{3}{1.5} - \frac{2\sqrt{7}}{5}\right) = 0{,}826$$

While:

$$\beta_S = P\left(X^2_{n-1} \leq \frac{X^2_{\alpha/2,n-1}}{\lambda^2}\right) - P\left(X^2_{n-1} \leq \frac{X^2_{1-\frac{\alpha}{2},n-1}}{\lambda^2}\right) =$$

$$\beta_S = P\left(X^2_{n-1} \leq \frac{21.74}{1.5^2}\right) - P\left(X^2_{n-1} \leq \frac{0.4234}{1.5^2}\right) = 0{,}86$$

Thus, the power of the control chart in the presence of the simultaneous shift of the mean and the standard deviation is:

$$P = 1 - 0{,}826 * 0{,}86 = 0{,}29$$

Exercise 2 (Solution)

By applying the PCA on the known variance-covariance matrix, the eigenvalues (i.e., the variances of the PCs) are the following:

$\lambda_1 = 8.54034$

$\lambda_2 = 0.99676$

$\lambda_3 = 0.26290$

The first PC explains about 87% of the overall data variability. The first two PCs explains 97.3% of the overall variability. Thus, retaining the first 2 PCs is needed. Their loadings are:

u1	u2
-0,67262	0,731098
-0,70276	-0,58269
-0,23178	-0,35491

b)

Being known that the scores along the first two PCs are normally distributed with:

$$\mu_{PC1} = 0, \mu_{PC2} = 0$$

$$\sigma^2_{PC1} = \lambda_1 = 8.54034,$$

$$\sigma^2_{PC2} = \lambda_2 = 0.99676$$

It is possible to design two univariate control charts for the mean of the first two PCs as follows (n=1 since we have individual observations):

PC1

$$UCL = \mu_{PC1} + K\sigma_{PC1}$$
$$CL = \mu_{PC1}$$
$$LCL = \mu_{PC1} - K\sigma_{PC1}$$

PC2

$$UCL = \mu_{PC2} + K\sigma_{PC2}$$
$$CL = \mu_{PC2}$$
$$LCL = \mu_{PC2} - K\sigma_{PC2}$$

The familywise Type I error is $\alpha = 0.005$.

The Type I error to be used in each control chart (since scores are independent by construction) is $\alpha^* = 1-(1-\alpha)^{1/2} = 0{,}002503$.

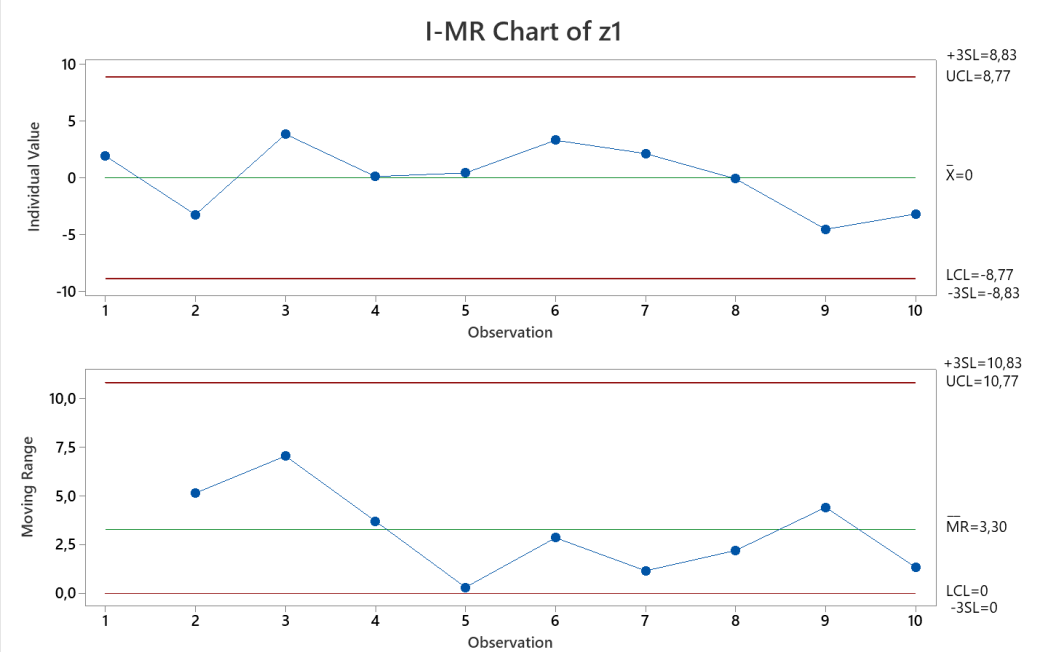
The control charts with $K = z_{\alpha^*/2} = 3.023$ have the following control limits:

PC1		PC2	
I	MR		
LCL = -8.83, UCL = 8.83	LCL = 0, UCL = 10.83	LCL = -3.018, UCL = 3.018	LCL = 0, UCL = 3.699

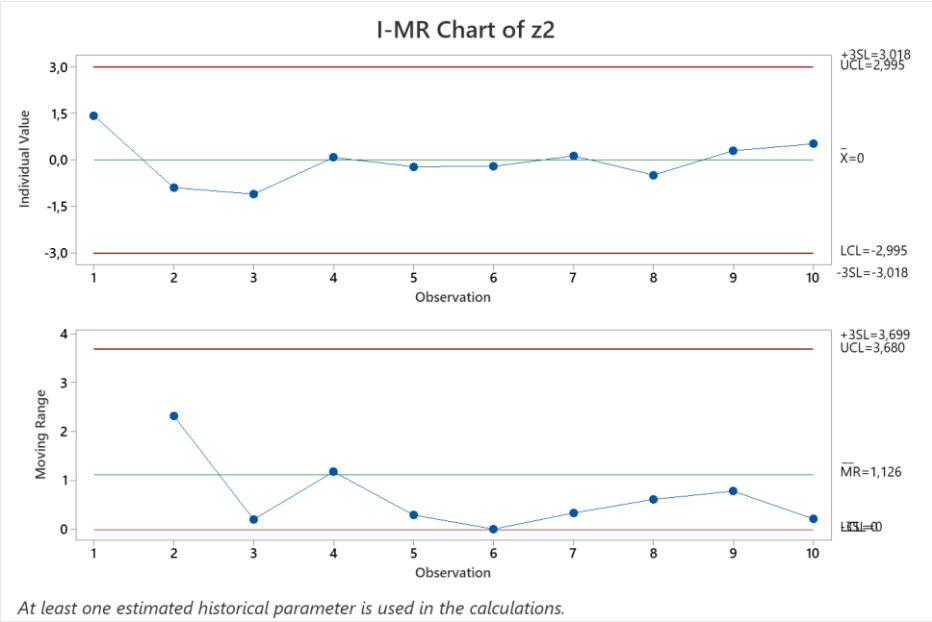
The new data can be projected onto the space spanned by the first 2 PCs. The following scores are computed:

z1	z2
1,907049	1,438762
-3,19565	-0,88839
3,823181	-1,09612
0,058564	0,086628
0,420506	-0,21203
3,274084	-0,20207
2,110288	0,137779
-0,06969	-0,48054
-4,53523	0,304709
-3,14352	0,523828

The control charts applied to the ten new observations are the following:



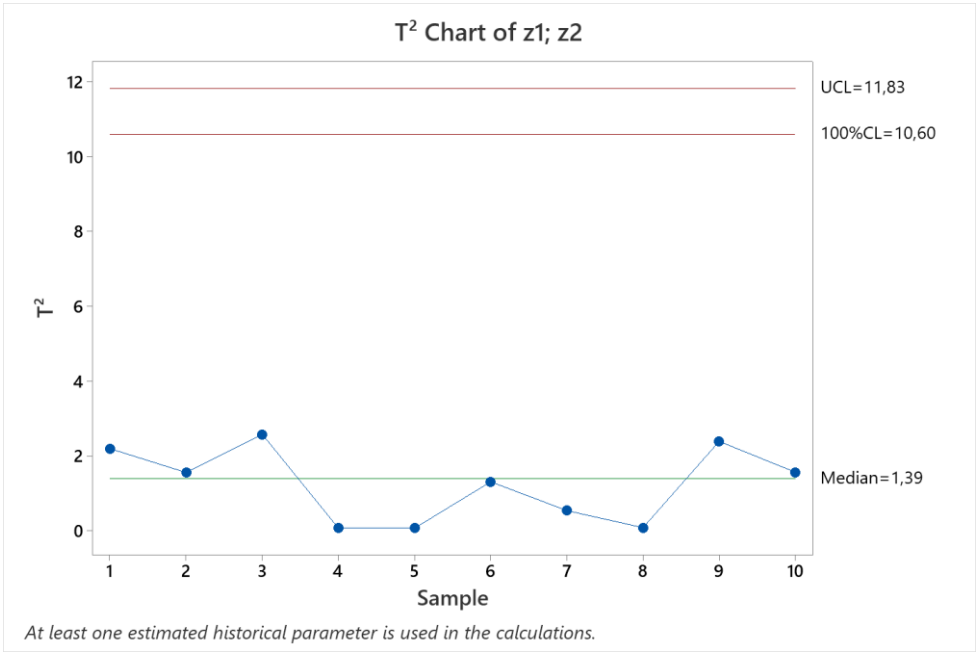
At least one estimated historical parameter is used in the calculations.



All the new observations are within the control limits, although hugging is present along the second PC (which can be a symptom of a change in the process).

c)

The T^2 control chart on the scores of the first 2 PCs with known mean and variance and $\alpha = 0.005$ is:



This control chart indicates that the process is in-control in the last ten observations.

d)

Let k=1 (only the first PC is retained). Then, the reconstructed data can be estimated as:

$$\hat{\boldsymbol{x}}_j(k) = \boldsymbol{\mu} + z_{j1}\boldsymbol{u}_1$$

For signal 1:

$$\hat{x}_{1j}(k) = \mu_1 + z_{j1}u_{11}$$

Being $\mu_{PC1} = 0, \sigma^2_{PC1} = \lambda_1 = 8.54034$, its mean and variance are:

$$E\left(\hat{x}_{1j}(k)\right) = E(\mu_1 + z_{j1}u_{11}) = \mu_1 = 11.3$$

$$V\left(\hat{x}_{1j}(k)\right) = V(\mu_1 + z_{j1}u_{11}) = \lambda_1 {u_{11}}^2 = 3.86$$

Let k=3 (no data reduction). Then, the reconstructed data can be estimated as:

$$\hat{\boldsymbol{x}}_j(k) = \boldsymbol{\mu} + z_{j1}\boldsymbol{u}_1 + z_{j2}\boldsymbol{u}_2 + z_{j3}\boldsymbol{u}_3$$

For signal 1:

$$\hat{x}_{1j}(k) = \mu_1 + z_{j1}u_{11} + z_{j2}u_{21} + z_{j3}u_{31}$$

Its mean and variance are:

$$E\left(\hat{x}_{1j}(k)\right) = E(\mu_1 + z_{j1}u_{11}) = \mu_1 = 11.3$$

$$V\left(\hat{x}_{1j}(k)\right) = V(\mu_1 + z_{j1}u_{11} + z_{j2}u_{21} + z_{j3}u_{31}) = \lambda_1 {u_{11}}^2 + \lambda_2 {u_{21}}^2 + \lambda_3 {u_{31}}^2 = 4.4 = V(x_{1j})$$

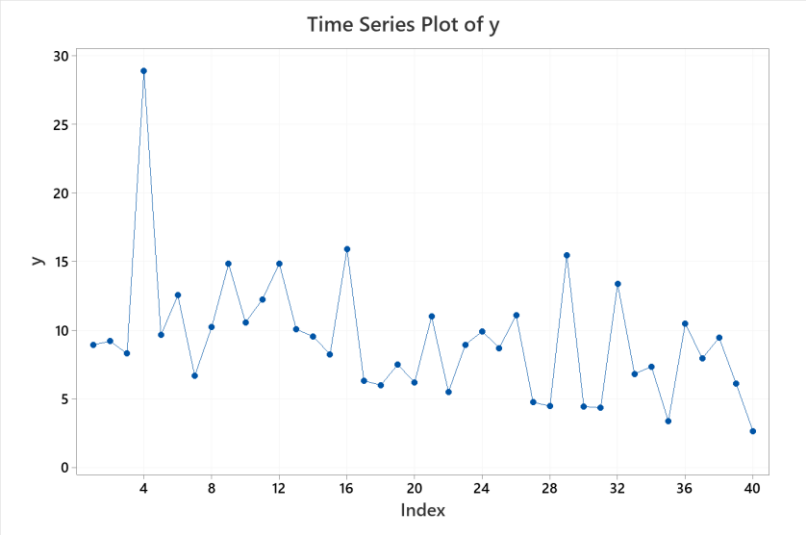
The mean of the reconstructed data is equal to the mean of the original data regardless of the number k of retained PCs.

The variance of the reconstructed data, instead, depends on the number k of retained PCs. When k=p (in this case, k=3), the reconstructed data coincide with the original data, as no dimensionality reduction is applied.

Exercise 3 solution

a)

The time series plot highlights a slight decreasing trend of the coating thickness:



By fitting a trend model to these data we get:

EXE1

Regression Analysis: y versus t

Regression Equation

y = 12,71 - 0,1652 t

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	12,71	1,36	9,33	0,000	
t	-0,1652	0,0579	-2,85	0,007	1,00

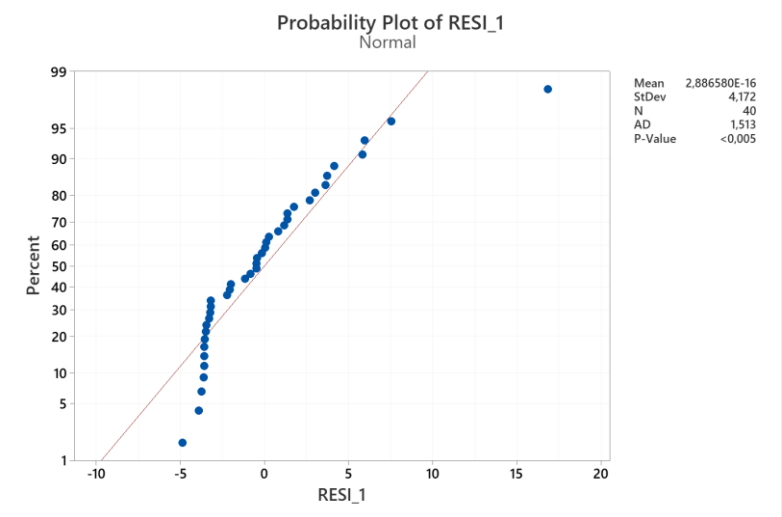
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4,22657	17,65%	15,48%	6,65%

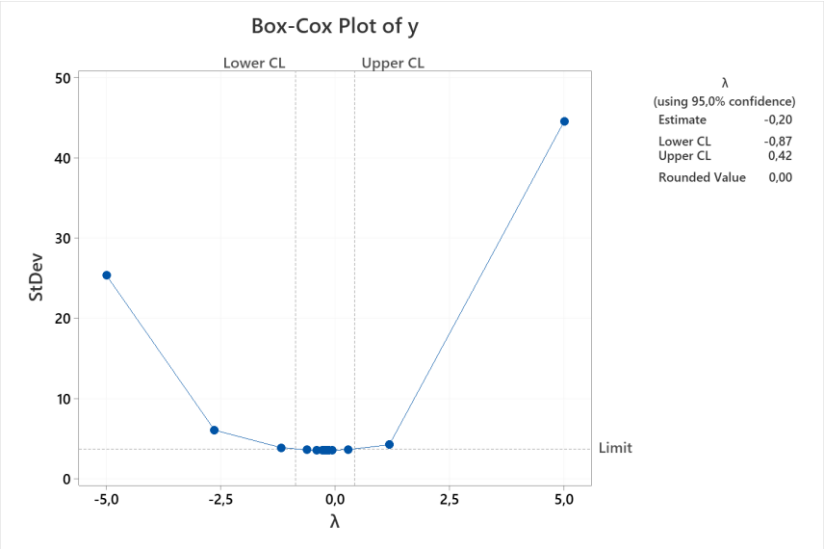
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	145,5	145,50	8,14	0,007
t	1	145,5	145,50	8,14	0,007
Error	38	678,8	17,86		
Total	39	824,3			

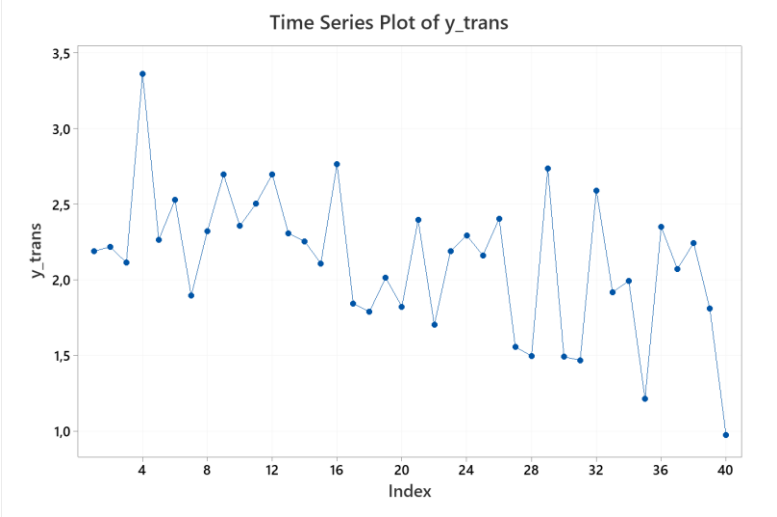
But a violation of the normality assumption affects the model residuals:



Such violation is caused by a skewed distribution of the measurements. It is possible to transform the data with the Box-Cox approach and then fit the trend model to the transformed data, as follows:



The data transformed with a natural logarithm transformation have the following time series pattern:



The trend model fitted on the transformed data is the following:

EXE1

Regression Analysis: y_trans versus t

Regression Equation

y_trans = 2,518 - 0,01893 t

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2,518	0,132	19,04	0,000	
t	-0,01893	0,00562	-3,37	0,002	1,00

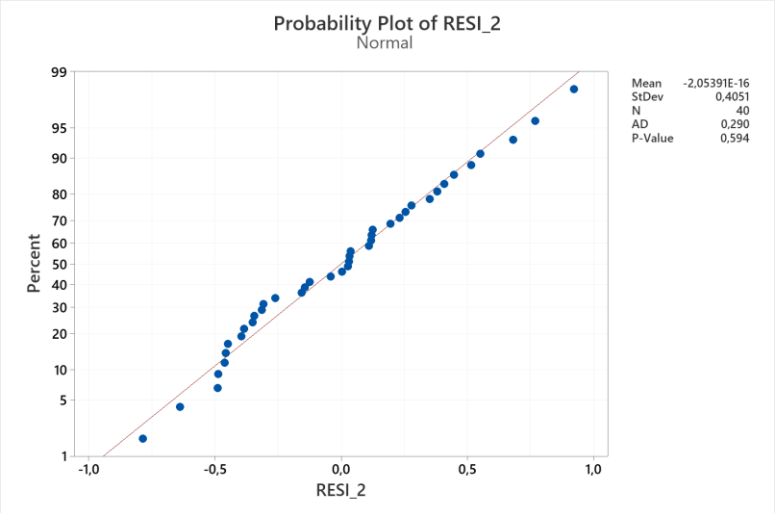
Model Summary

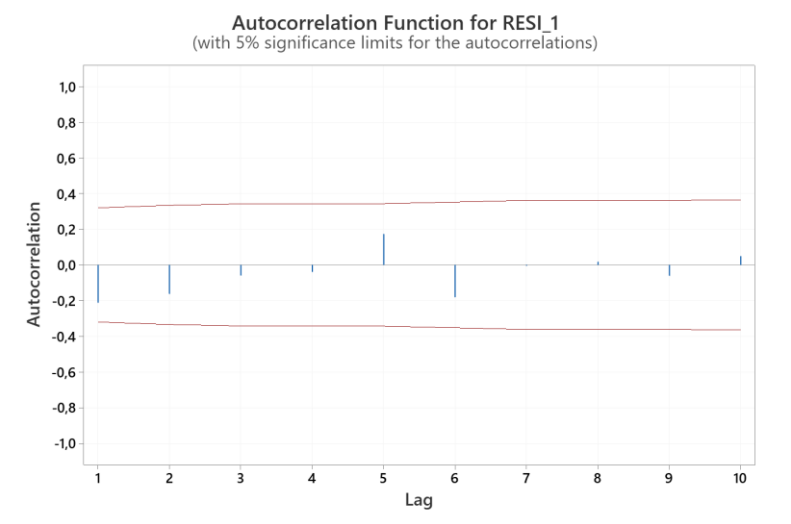
S	R-sq	R-sq(adj)	R-sq(pred)
0,410438	22,98%	20,96%	13,40%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1,910	1,9105	11,34	0,002
t	1	1,910	1,9105	11,34	0,002
Error	38	6,401	0,1685		
Total	39	8,312			

The model is significant and now the residuals meet the assumptions:





Test

Null hypothesis	H₀: The order of the data is random
Alternative hypothesis	H₁: The order of the data is not random

Number of Runs

Observed	Expected	P-Value
19	20,80	0,560

Given $ARL_0 = 300$, the type I error for the trend control chart is $\alpha = 0.0033$. The resulting control chart for the transformed data is the following:

$$UCL = b_0 + b_1 t + z_{\alpha/2} \frac{\overline{MR}}{d_2(2)}$$

$$UCL = b_0 + b_1 t$$

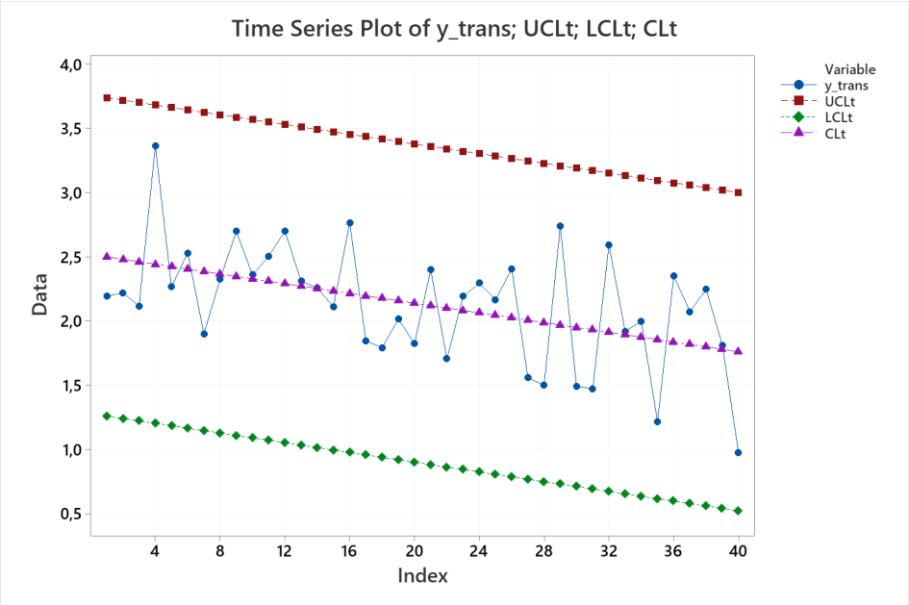
$$LCL = b_0 + b_1 t - z_{\alpha/2} \frac{\overline{MR}}{d_2(2)}$$

Where:

$\overline{MR} = 0{,}4764$

$d_2(2) = 1{,}128$

$z_{\alpha/2} = 2{,}938$

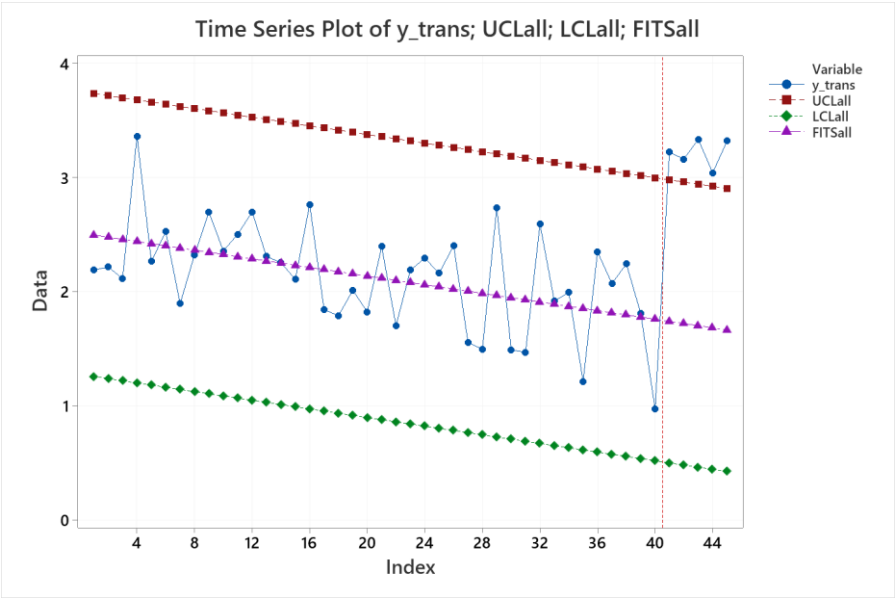


b)

Before plotting the new data onto the control chart designed in point a), we shall transform them with the natural logarithm transformation:

ynew	ynew trans
25,12	3,22366
23,65	3,16336
28,11	3,33613
20,98	3,04357
27,70	3,32143

The new data are out-of-control:



c)

Let $LSL = 3$ µm and let $\gamma \geq 10\%$ be the probability of producing a non-conforming part, then:

$$\gamma = P({y_t}^* \leq LSL^*) = \Phi\left(\frac{LSL^* - \mu_t}{\sigma_\varepsilon}\right) \geq 0{,}1$$

Where ${y_t}^*$is the natural logarithm of the coating thickness and LSL^* is the natural logarithm of the lower specification limit, as it results from the Box-Cox transformation.

Thus:

$$\gamma = \Phi\left(\frac{LSL^* - \mu_t}{\sigma_\varepsilon}\right) = \Phi\left(\frac{LSL^* - (b_0 + b_1 t)}{\sigma_\varepsilon}\right)$$

Where:

$LSL^* = 1.0986$

$\sigma_\varepsilon = 0{,}41$

$b_0 = 2{,}518$

$b_1 = -0{,}01893$

We get the estimate of γ as a function of time t as shown in the table below:

t	gamma
1	0,000318032
2	0,000376414
3	0,000444622
4	0,000524139
5	0,000616644
6	0,000724028
7	0,00084842
8	0,000992206
9	0,001158057
10	0,00134895
11	0,0015682
12	0,001819484
13	0,002106866
14	0,002434835
15	0,002808325
16	0,003232748
17	0,003714023
18	0,004258605
19	0,004873508
20	0,005566335
21	0,006345298
22	0,007219242
23	0,008197661
24	0,009290711
25	0,010509223
26	0,011864706
27	0,013369346
28	0,015036
29	0,016878182
30	0,018910044
31	0,021146344

32	0,023602412
33	0,026294103
34	0,029237739
35	0,03245005
36	0,035948094
37	0,039749178
38	0,043870759
39	0,048330348
40	0,053145389
41	0,058333145
42	0,063910567
43	0,069894157
44	0,076299826
45	0,083142747
46	0,090437203
47	0,098196433
48	0,106432479
49	0,11515603
50	0,124376267

Based on available model, the probability of producing at least 10% of non-conforming parts is achieved after 48 hours of coating process.

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)

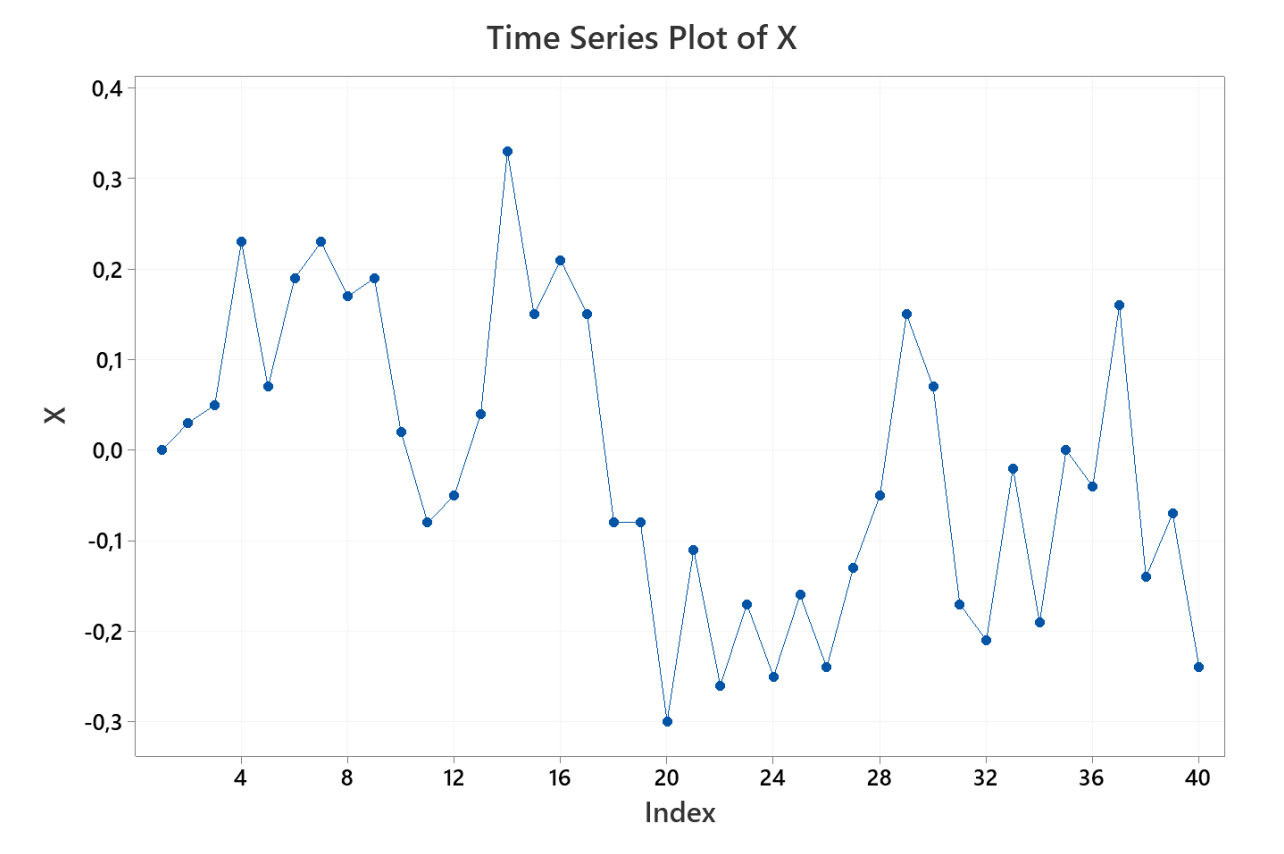
A manufacturing process is modelled by means of an AR(2) model. In case of a sudden shift of the process mean with entity $\delta\sigma_x$ (where σ_x is the standard deviation of the process data), what is the new average of the model residuals? Express the result as a function of σ_ε and of model parameters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

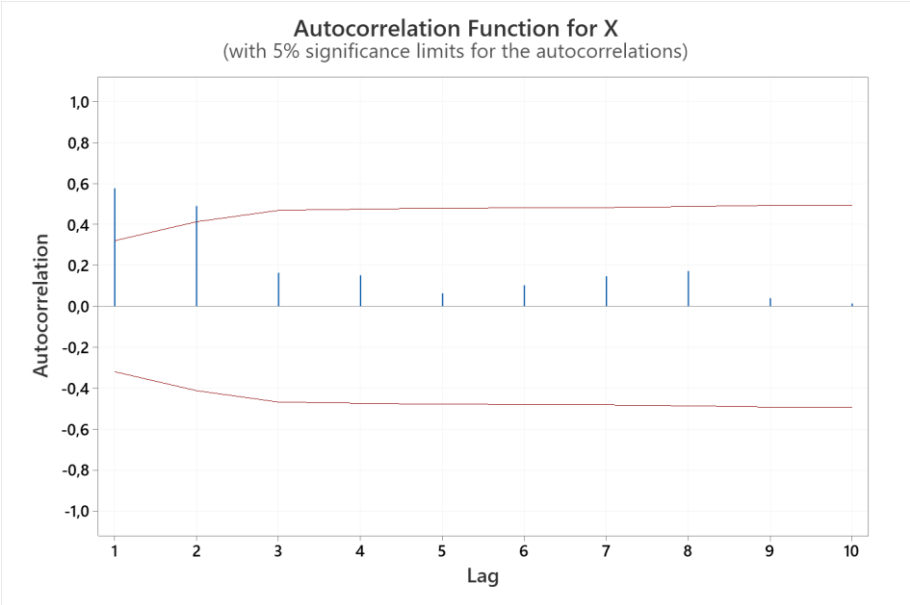
Test

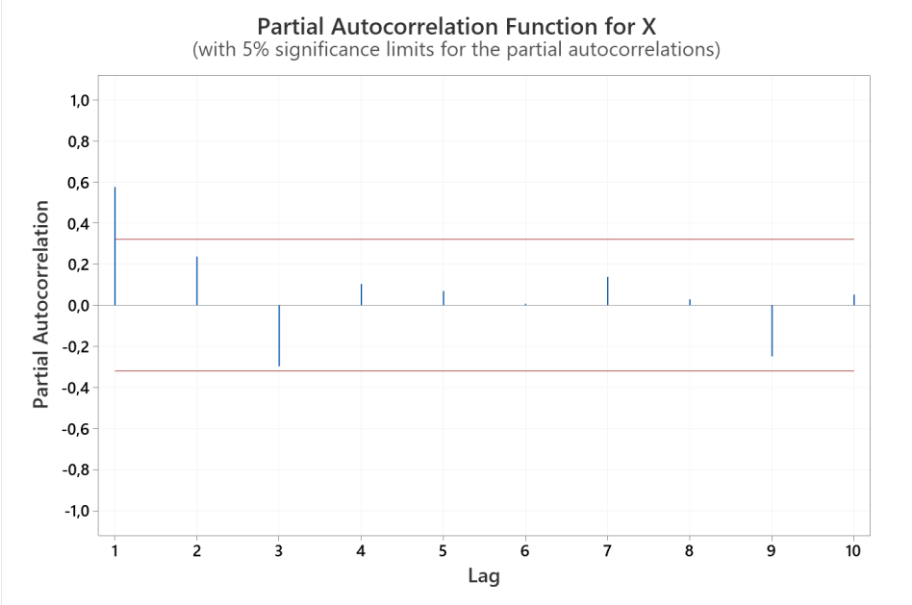
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

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

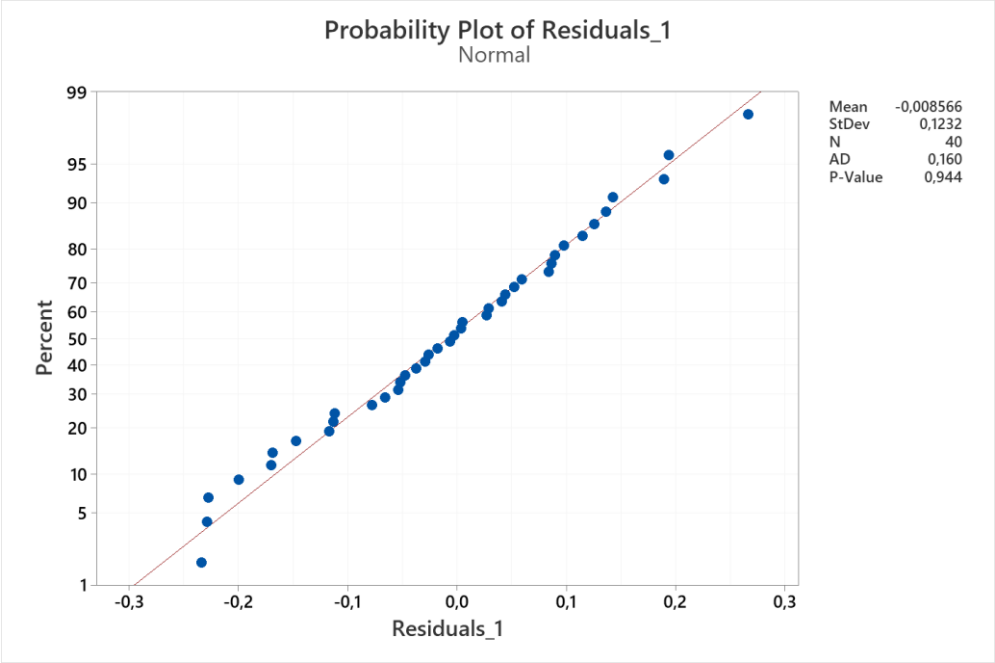
Number of observations: 40

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.

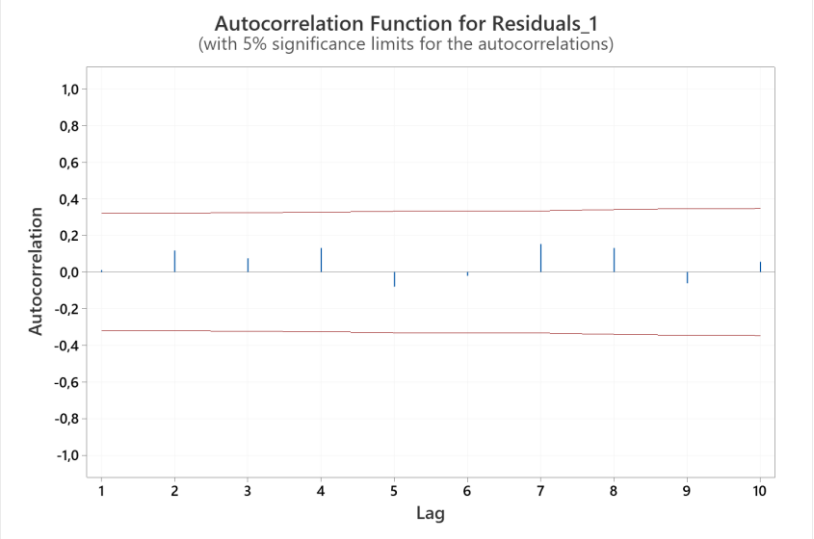


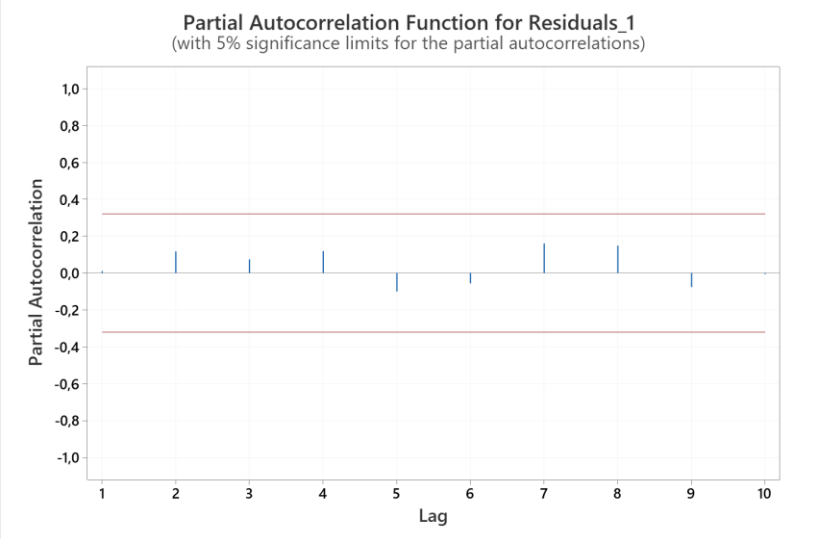
Test

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

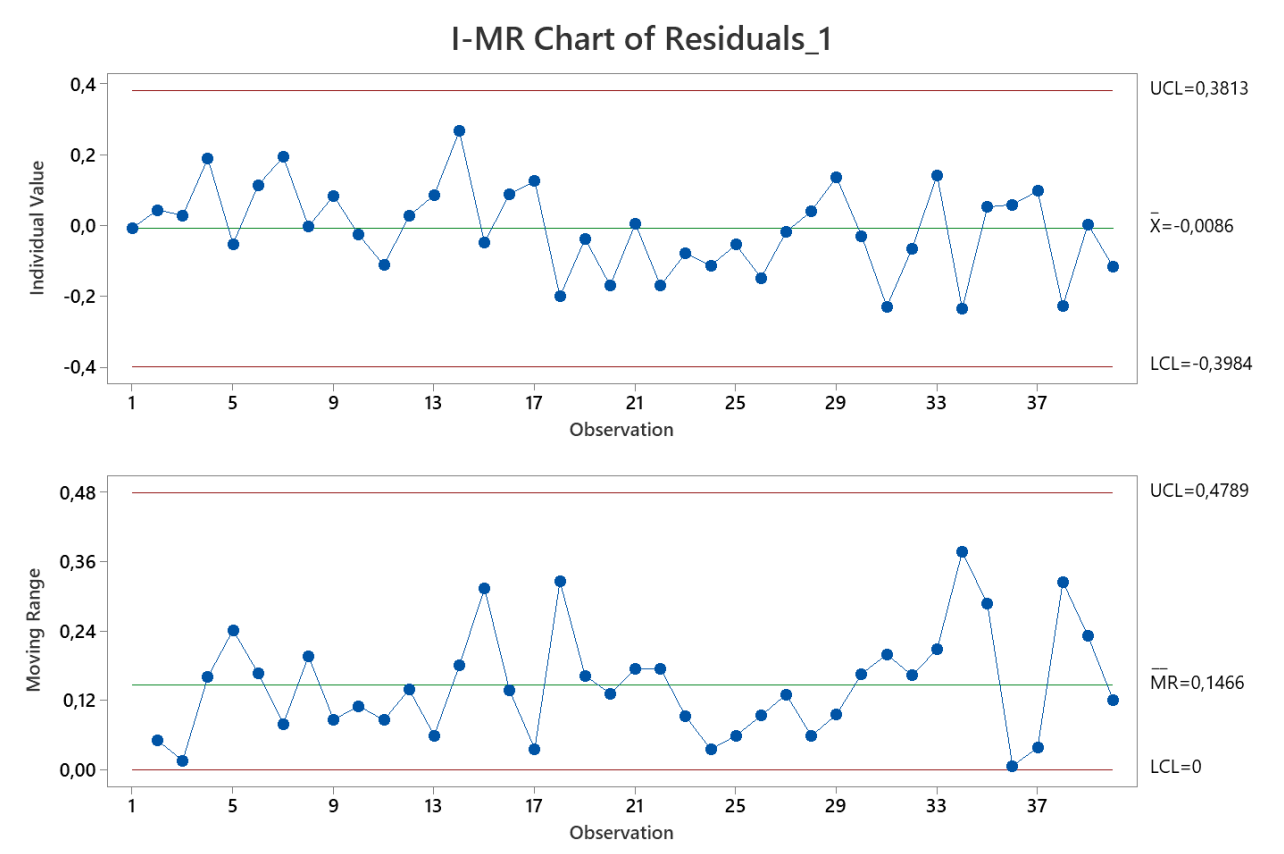
Observed	Expected	P-Value
18	20,95	0,343





b)

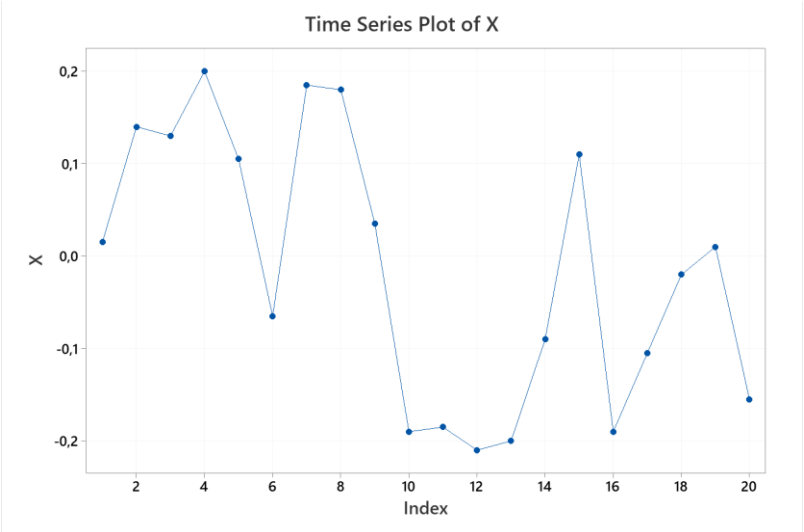
The control chart on the model residuals is the following:



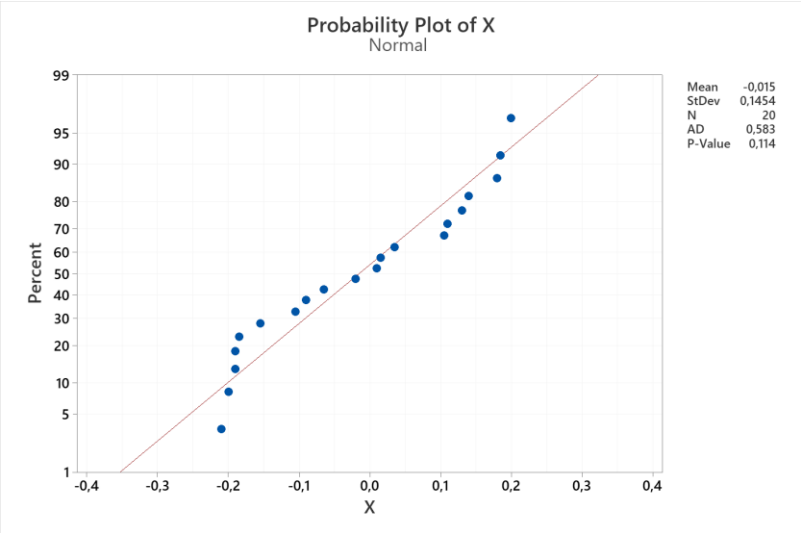
No violation of the limit is signalled, although a small shift may have occurred, since the residuals of the first half of samples are predominantly above the center line in the I chart, whereas the second half is predominantly below. However, this shift is too small to generate an alarm in Shewhart 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:



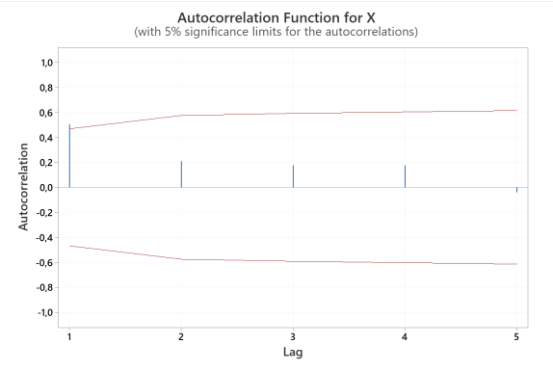
Test

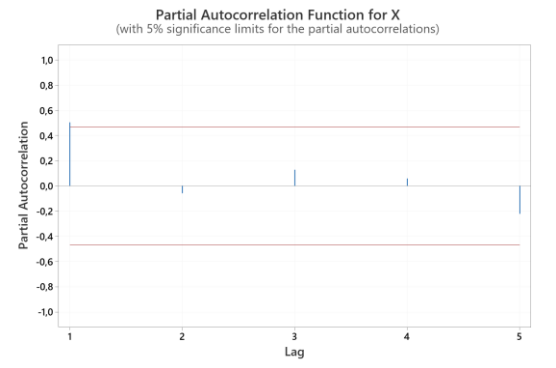
Null hypothesis H_0: The order of the data is random
Alternative hypothesis H_1: The order of the data is not random

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

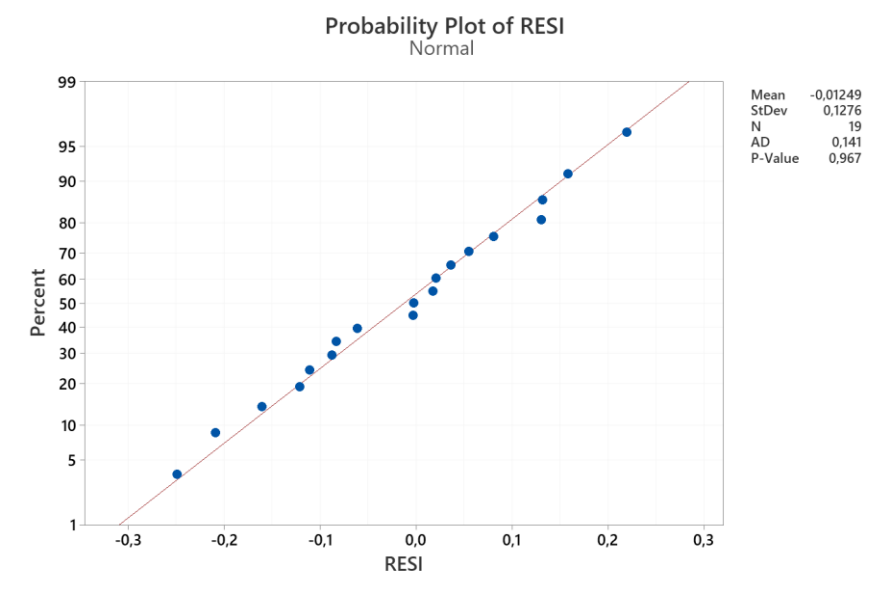
Model Summary

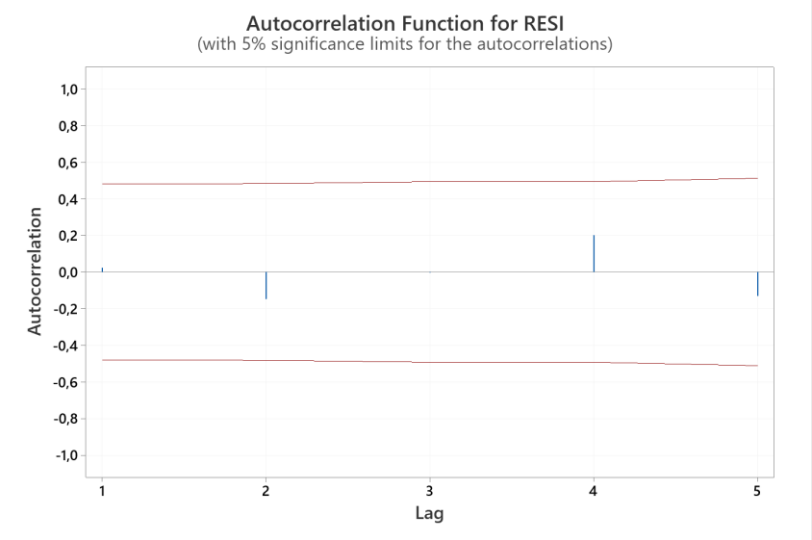
S	R-sq	R-sq(adj)	R-sq(pred)
0,128234	27,07%	23,02%	22,67%

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

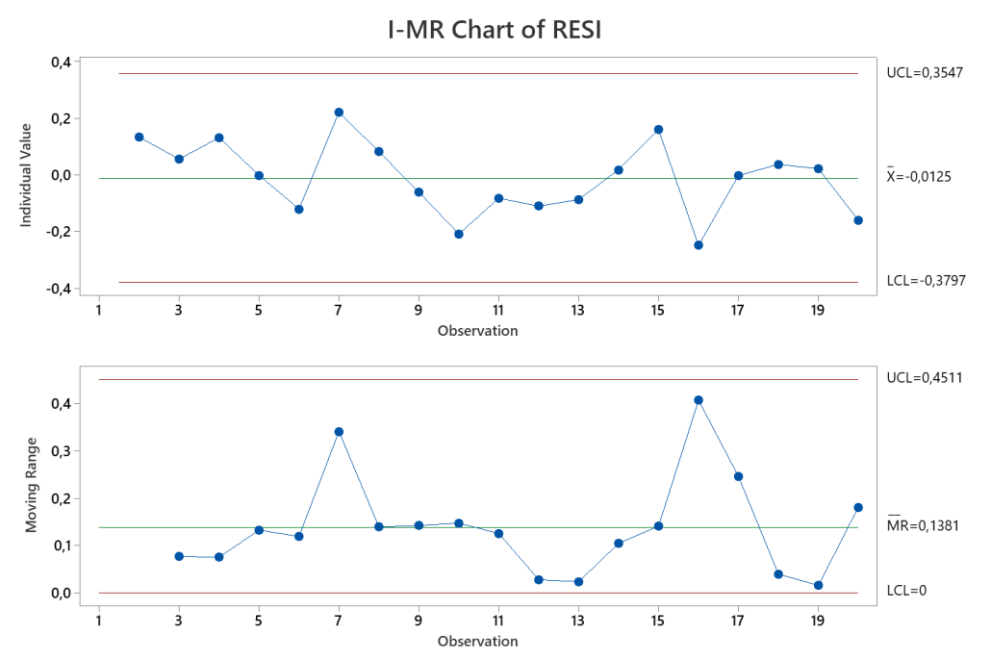
Null hypothesis H_0: The order of the data is random
Alternative hypothesis H_1: The order of the data is not random

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)

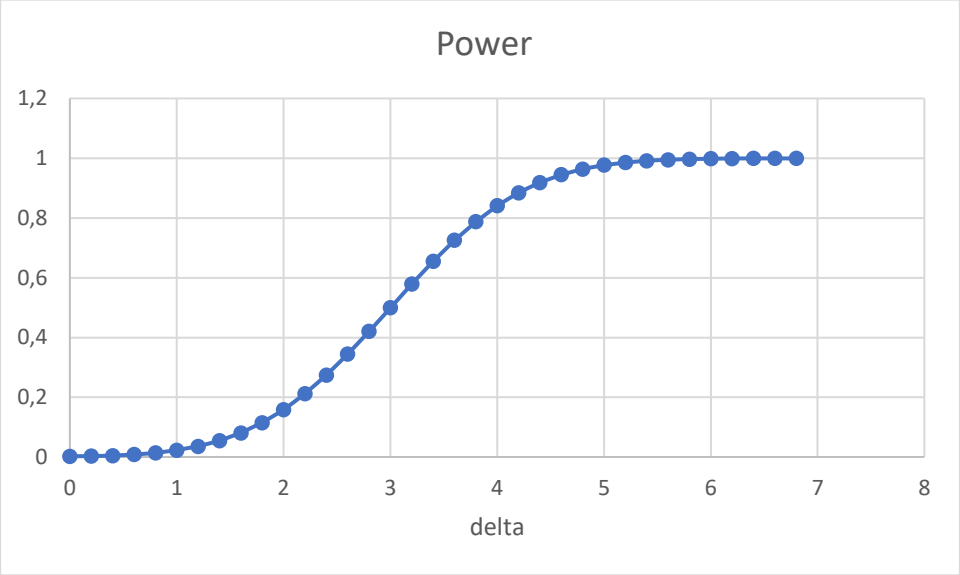
The type II error is:

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

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

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

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

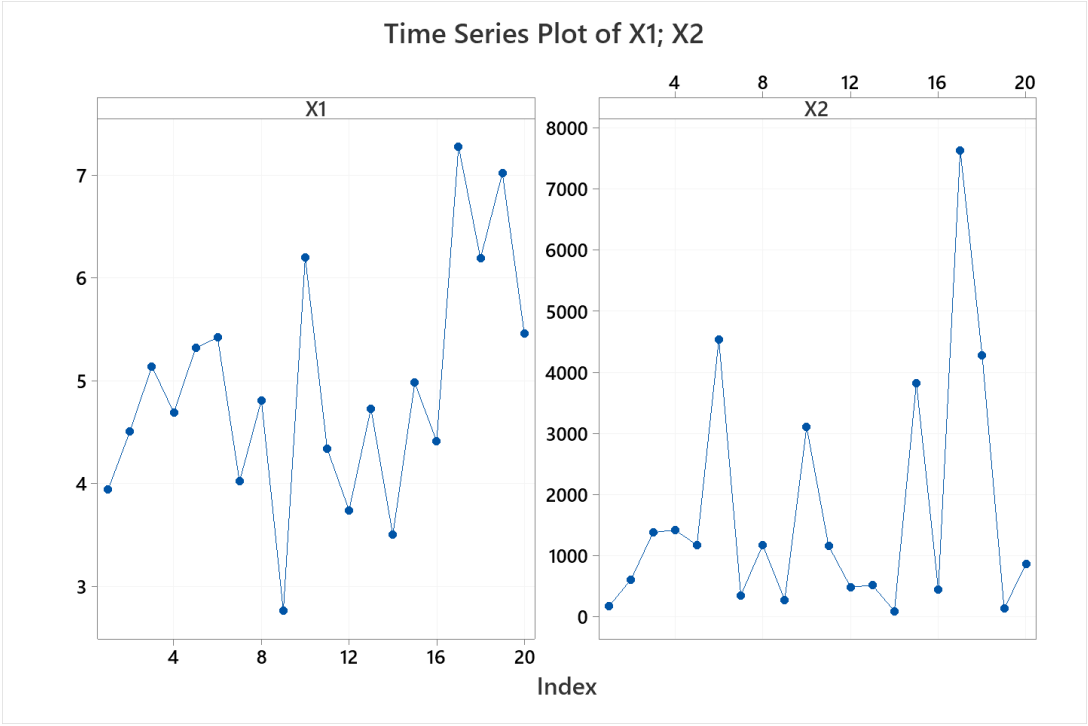


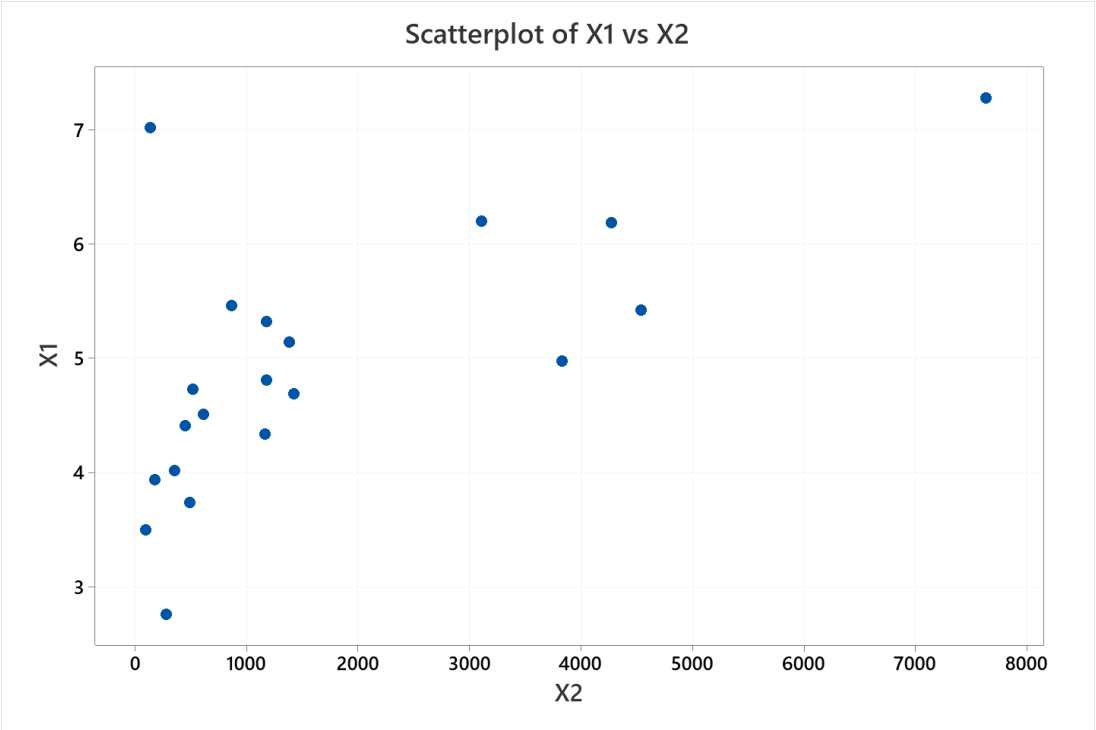
Therefore, the minimum value of δ that is detected with a power of 90% is $\delta = 4.4$.

Exercise 2

a)

Data snooping.





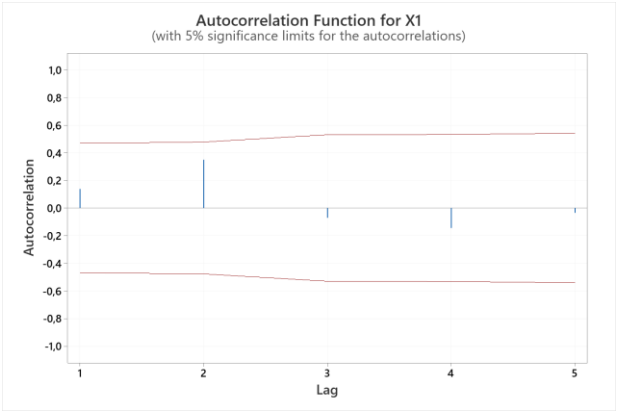
The second variable looks quite skewed. Check of assumptions:

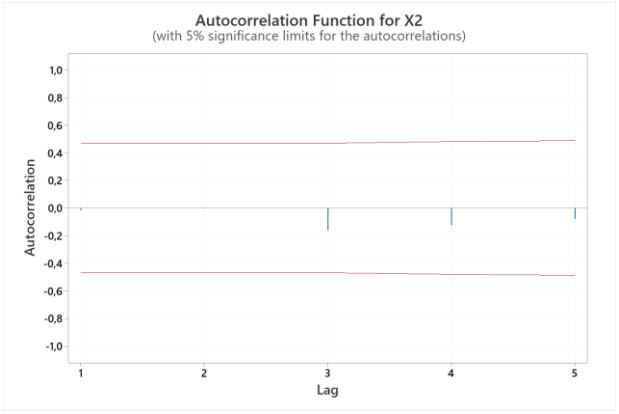
Test

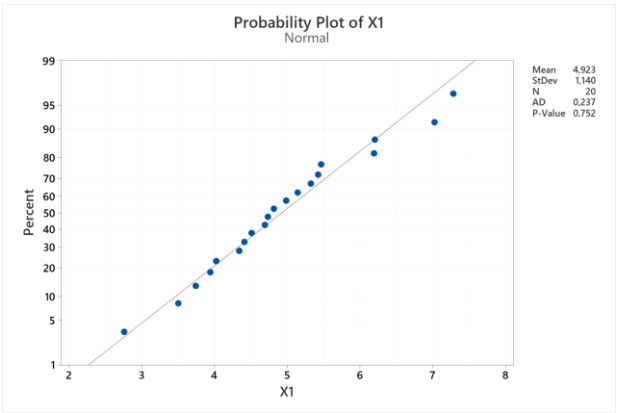
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

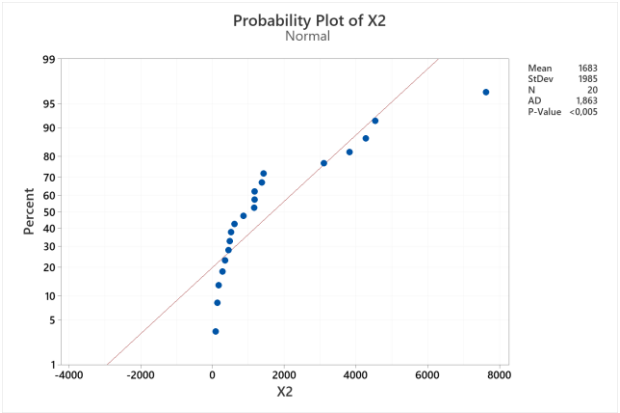
	Number of Runs		
Variable	Observed	Expected	P-Value
X1	10	10,90	0,676
X2	9	8,50	0,755

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

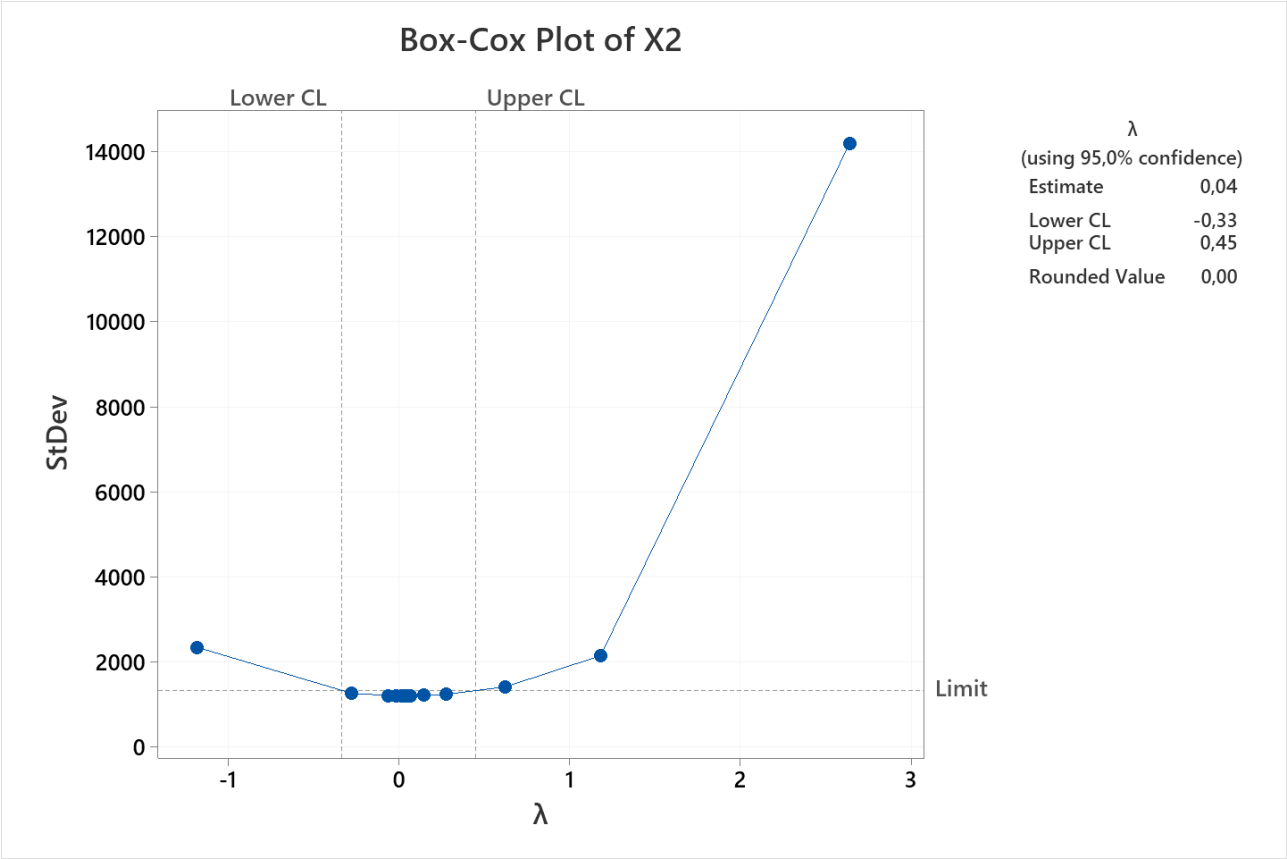




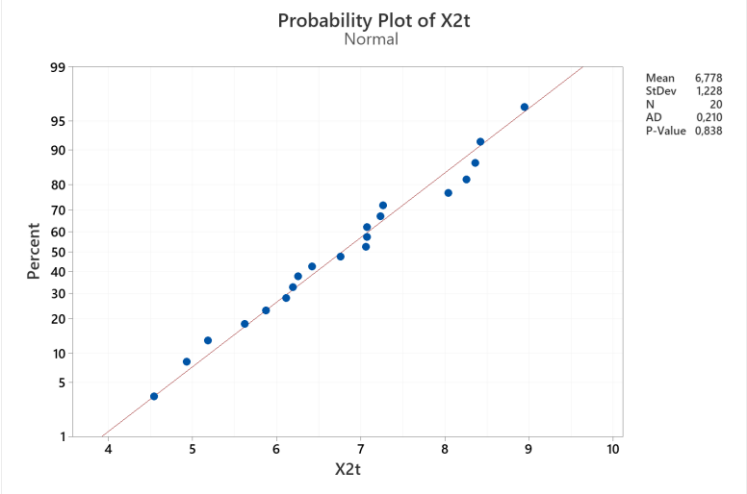




The first variable is normal and independent. The second variable meets the randomness assumption but it violates the normality assumption. We can try to apply the Box-Cox transformation.



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



Randomness is still met too.

Test

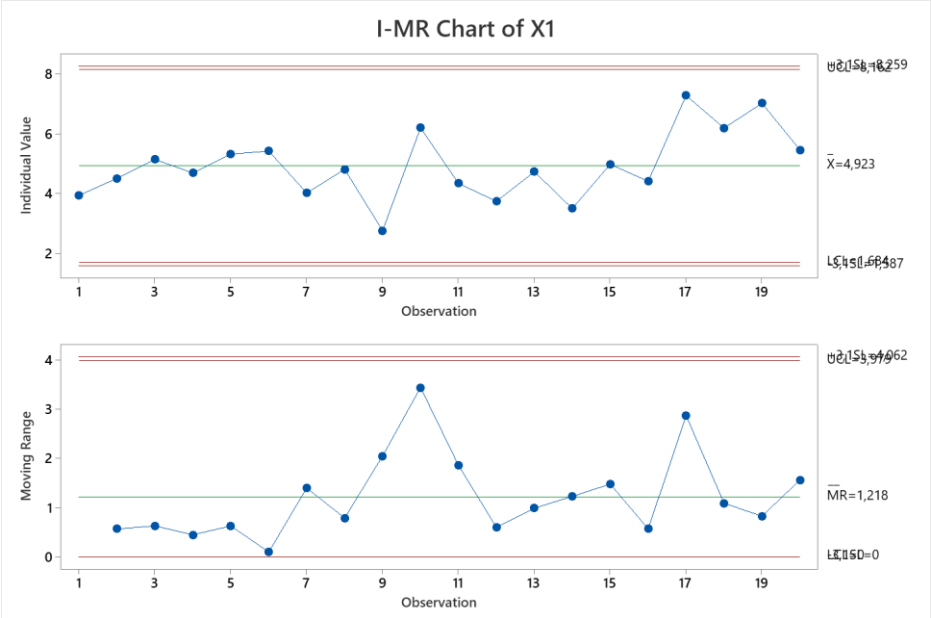
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

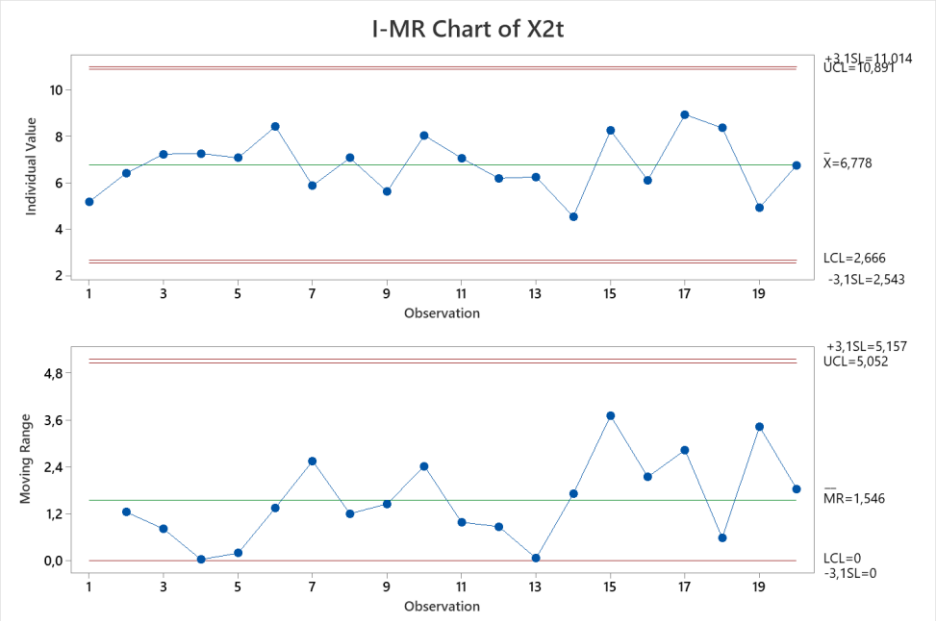
Number of Runs

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).

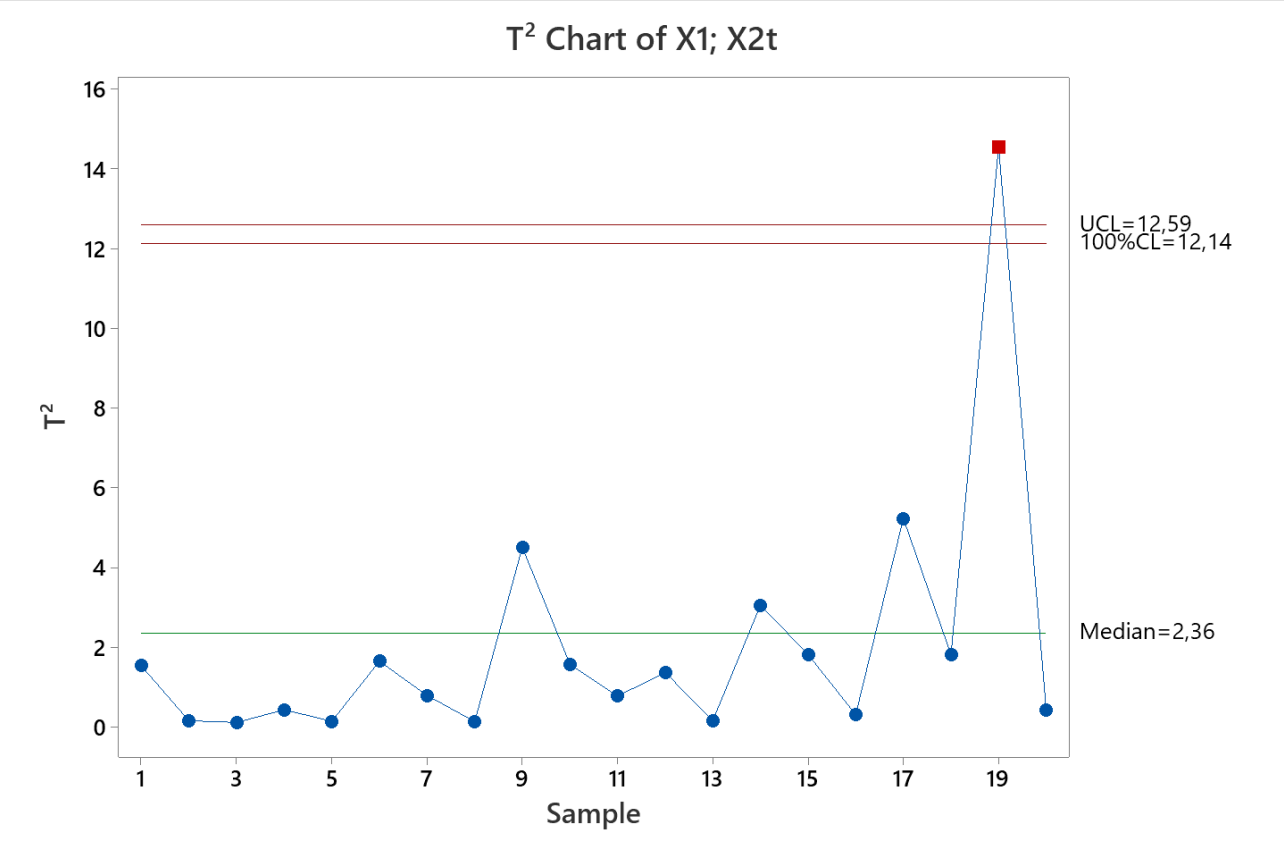




Apart from a small sustained shift in the last four observations of the first variable (which may possibly deserve some attention), no violation of the control limits is present.

b)

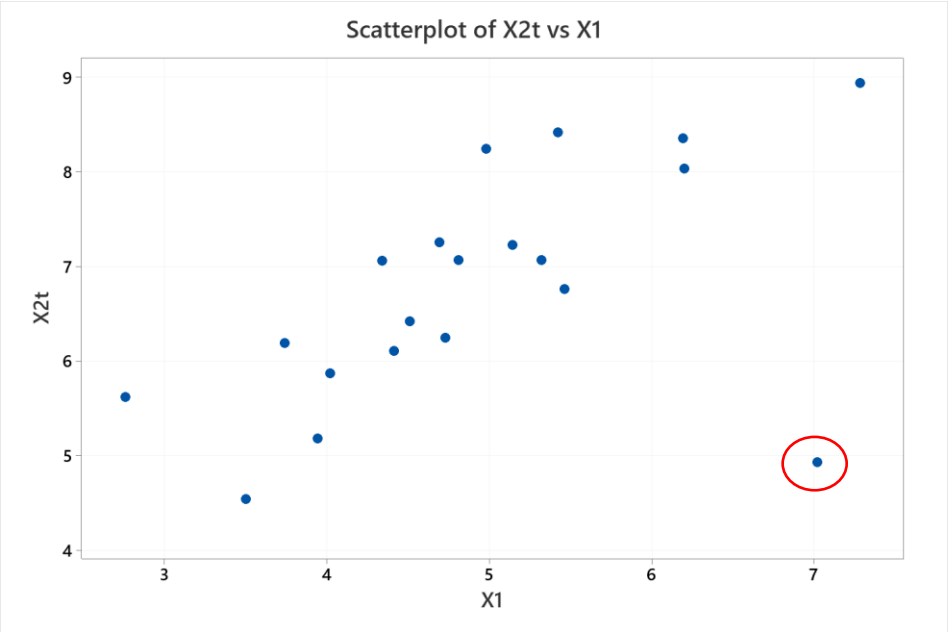
It is possible to design a T2 control chart with $ARL_0 = 250$, and hence $\alpha = 0.004$. The control chart is the following:



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

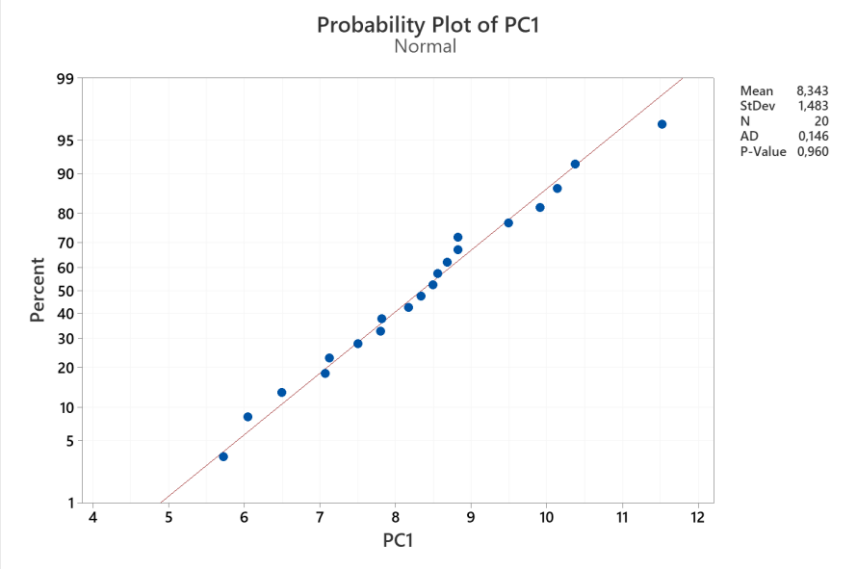
Eigenvalue	2,2004	0,6078
Proportion	0,784	0,216
Cumulative	0,784	1,000

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.



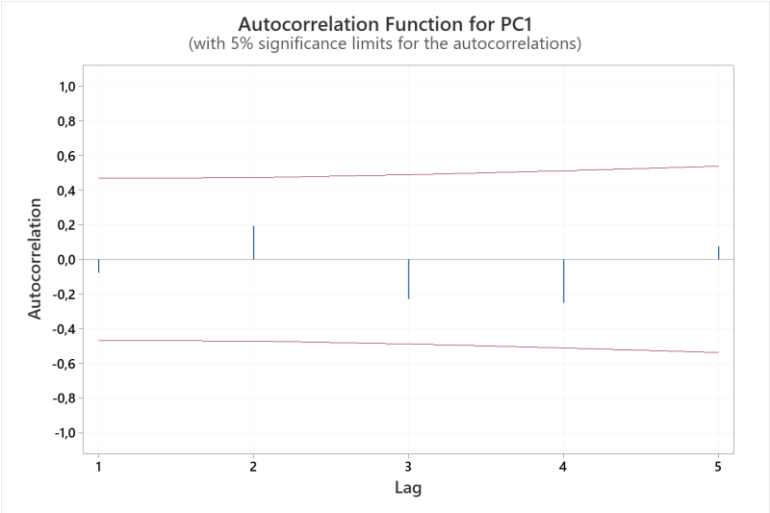
Test

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

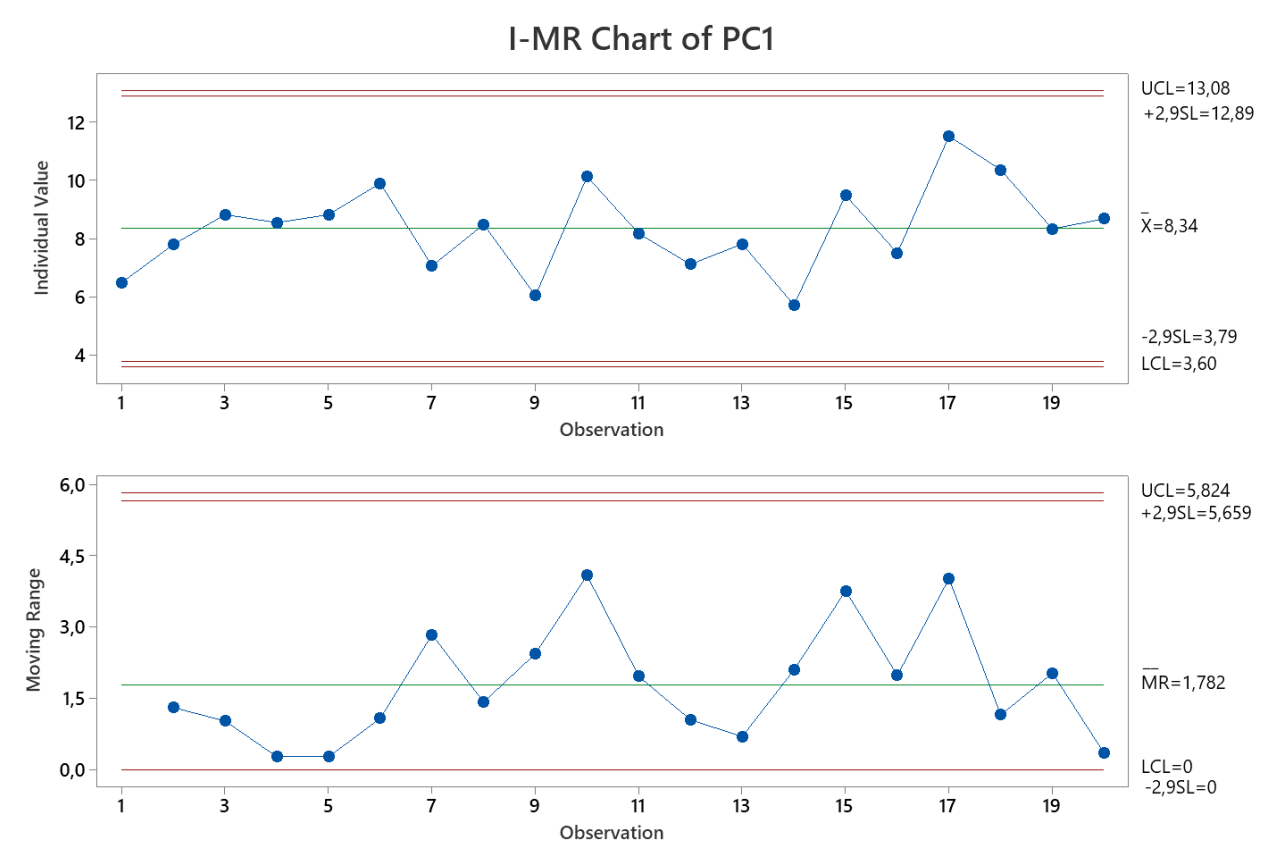
Number of Runs

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.

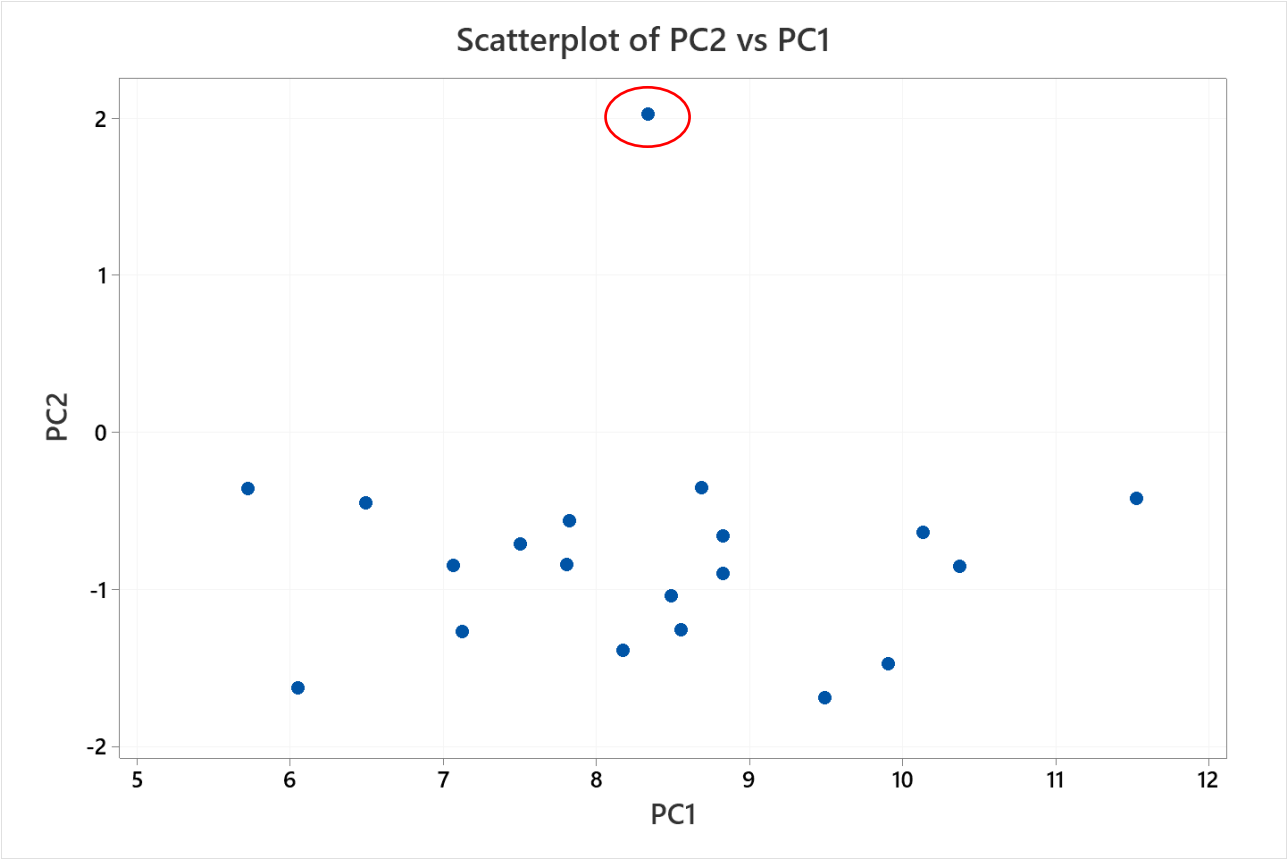


The I-MR control chart on the first PC with $ARL_0 = 250$, and hence $z_{\alpha/2} = 2.878$, is the following (ignore the limits at K=3):



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:

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

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

QUALITY DATA ANALYSIS

16/01/2023

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
- **MULTICHANCE STUDENTS SHALL SKIP: Exercise 1) point c, Exercise 2) point d.**

Exercise 1 (15 points)

A company produces titanium impellers for the oil and gas sector. To meet sustainability targets, the company started monitoring the spindle power consumption during machining operations. The production of each impeller consists of four consecutive machining steps: step 1, 2 and 3 are roughing operations, while step 4 is the finishing operation. Table 1 includes power consumption data gathered during the production of ten consecutive impellers.

Table 1

Step	X (kW)	Step	X (kW)	Step	X (kW)	Step	X (kW)
1	12,05	3	12,12	1	11,79	3	12,09
2	12,2	4	11,84	2	12	4	11,68
3	11,86	1	11,96	3	11,95	1	12,14
4	11,81	2	12,05	4	11,76	2	12,07
1	12,24	3	11,95	1	12,03	3	12,16
2	12,17	4	11,67	2	12,07	4	12,2
3	12,05	1	11,93	3	11,87	1	11,98
4	11,79	2	11,88	4	11,59	2	11,86
1	11,92	3	11,87	1	11,91	3	11,98
2	12,19	4	11,63	2	11,97	4	11,75

a) Find a suitable model for power consumption data in Table 1.
b) The head of the quality department is interested in using spindle power data to monitor the stability of the process. Based on the result at point a) design a suitable control chart with $ARL_0 = 200$. Assume the existence of assignable cause if out of control observations are present.
c) By using a suitable statistical test, determine whether the power consumption of the finishing operation is statistically lower than the power consumption during the roughing phase (exclude out of control observations identified in point b), if any).

Exercise 2 (15 points)

A wine producer decides to apply statistical process monitoring tools to keep under control the quality of his production. During the barrel aging phase, he periodically measures four quality variables, x1, x2, x3, x4 taking a wine sample from randomly selected barrels. Data collected in successive samples are reported in Table 2.

Table 2

Sample	X1	X2	X3	X4
1	30,7	15,2	294,6	75,6
2	32,2	16,1	292,5	76,9
3	27,2	14,7	295,9	77,5
4	31,1	16,7	299,3	79,2
5	29,4	16,4	293,8	87,2
6	28,4	14,7	302,1	73,5
7	29,5	13,7	286,6	75,4
8	30,4	15,8	294,8	74,6
9	34,4	18,7	305,5	75
10	33,4	17,2	290,4	78
11	28,3	15	296,1	78,8
12	33,7	17,6	295,6	76
13	30,9	15,2	293,7	76,6
14	29,9	15,2	295	75,5
15	30,9	14,8	286,3	78,4

a) The wine maker is interested in using the PCA to analyze these data. Would it be more appropriate to use the sample variance-covariance matrix or the sample correlation matrix to estimate the principal components?
b) Estimate the PCA model for data in Table 2 by retaining the number of principal components required to capture at least 75% of the overall variability (report the eigenvalues and eigenvectors of retained PCs).
c) Based on the result of point b), design a Hotelling's T^2 control chart for the wine data with $ARL_0 = 200$. Can we conclude that the barrel aging process is stable and in-control?
d) How do the result of point c) changes if the Hotelling's T^2 control chart is designed using m+1 principal components, where m is the number of PCs used in point c)? Discuss the results.
e) The wine maker decides to extend the data collection for a longer period. Based on the collection of 100 samples, he estimates the following sample mean and variance-covariance matrix:

$$\bar{\boldsymbol{x}} = [28.8 \;\; 12 \;\; 288 \;\; 75], \mathbf{S} = \begin{bmatrix} 1.4 & 0.75 & 0.1 & 0.5 \\ 0.75 & 1.3 & 1.7 & 0.5 \\ 0.1 & 1.7 & 6.6 & 0.3 \\ 0.5 & 0.5 & 0.3 & 3.6 \end{bmatrix}$$

Design a statistical test to determine if the variances explained, respectively, by the first and second PC estimated from the new data are statistically different from the ones estimated from data in Table 2 (use a familywise confidence level $\alpha = 0.05$; assume that the new sample is random, normal and independent from the data sample in Table 2).

Exercise 3 (3 points)

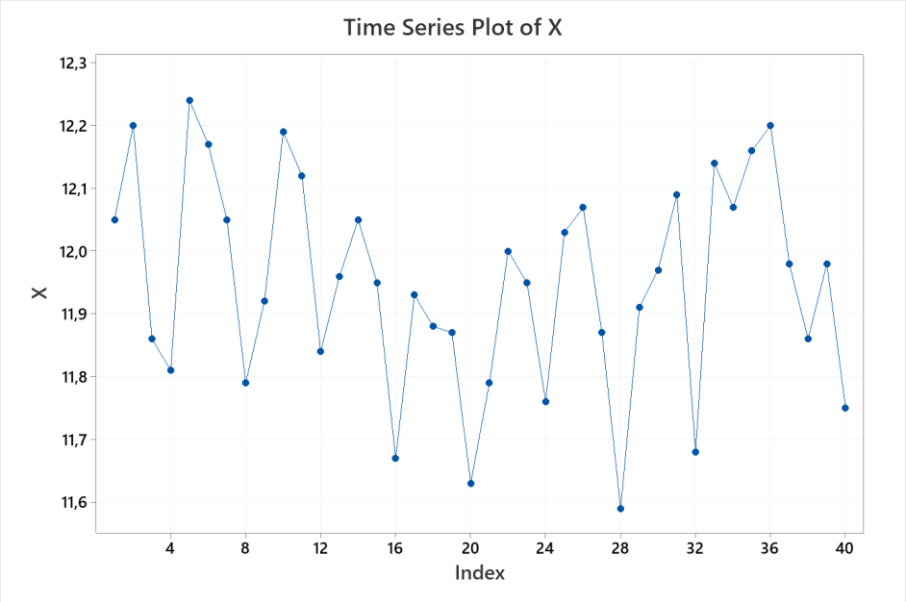
Using the sample statistics defined in point e) of Exercise 2, report the eigenvalue and eigenvector corresponding to the first PC. Show that the variance explained by this first PC is higher than the variance explained by a simple linear combination of the four variables where equal weight is given to all the variables (for sake of comparison, remind the normalization constraint $\boldsymbol{a}'\boldsymbol{a} = 1$). Discuss the result.

Solutions

Exercise 1

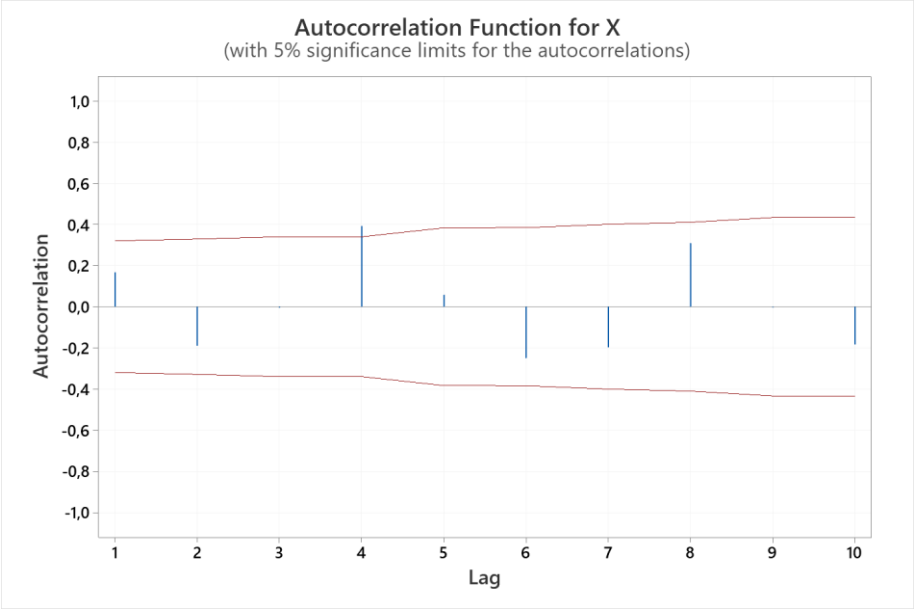
a)

Time series plot of the spindle power data:

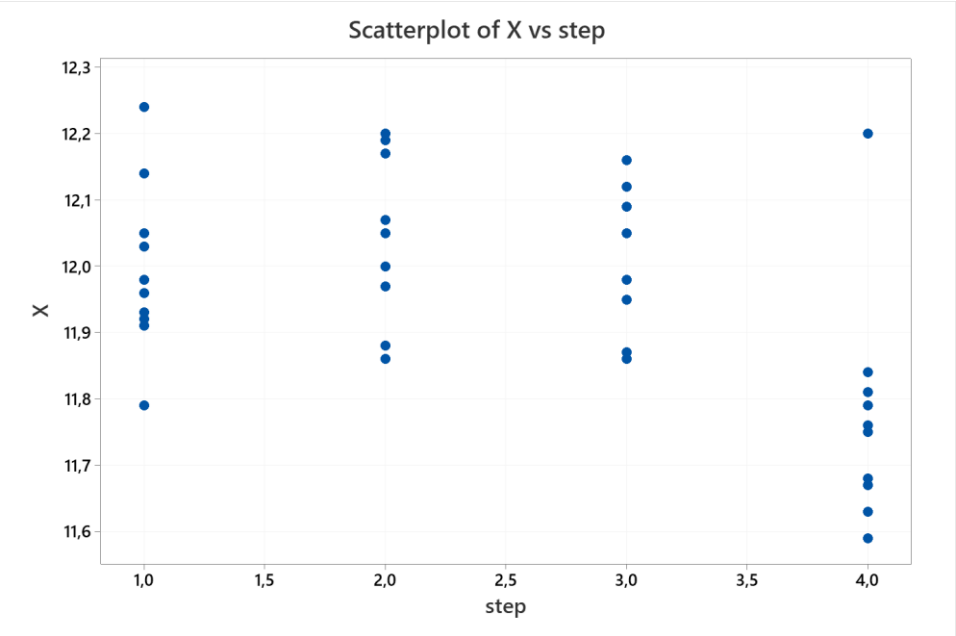


There is a meandering pattern with a seasonal drop of the spindle power in correspondence of step 4.

This lag 4 effect is also visible from the SACF:



The way in which the power varies along the different process steps is shown in the following scatter plot:



A part from one apparent outlying value, step 4 (finishing) yields a lower power consumption, as expected.

Based on this, it would be possible to fit a model of the spindle power using as regressor the categorical variable “step” as follows:

ESE1

Regression Analysis: X versus step

Method

Categorical predictor coding (1; 0)

Regression Equation

X = 11,9950 + 0,0 step_1 + 0,0510 step_2 - 0,0050 step_3 - 0,2230 step_4

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	11,9950	0,0424	282,80	0,000	
step					
2	0,0510	0,0600	0,85	0,401	1,50
3	-0,0050	0,0600	-0,08	0,934	1,50
4	-0,2230	0,0600	-3,72	0,001	1,50

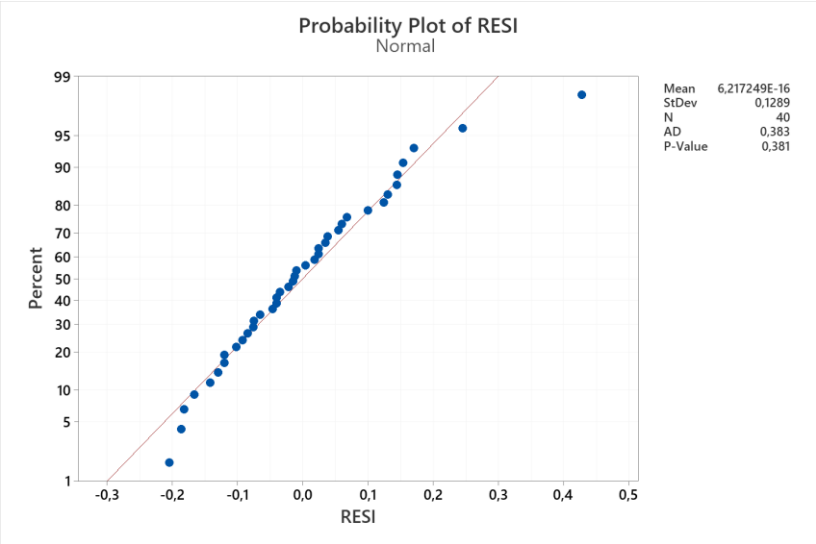
Model Summary

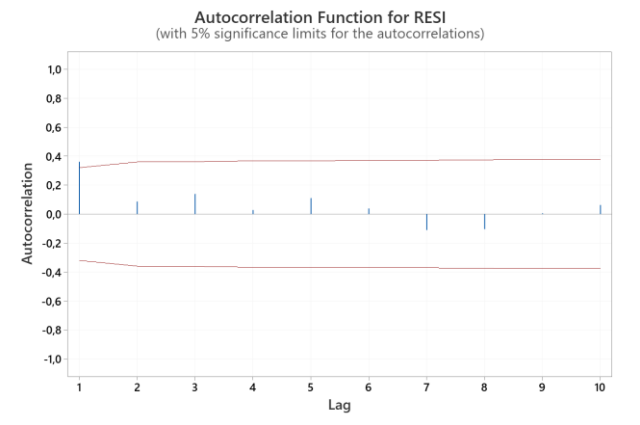
S	R-sq	R-sq(adj)	R-sq(pred)
0,134128	40,74%	35,80%	26,84%

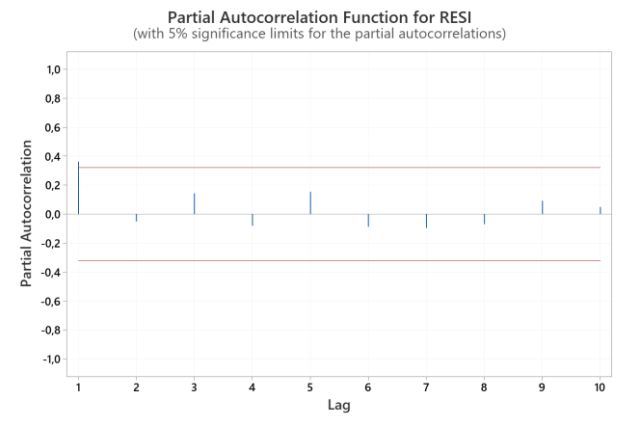
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	0,4452	0,14841	8,25	0,000
step	3	0,4452	0,14841	8,25	0,000
Error	36	0,6477	0,01799		
Total	39	1,0929			

The residuals are normal but not independent:







Bartlett's test at lag = 1 (95% confidence):

$$|r_k| = 0.361$$

$$\frac{z_{\alpha/2}}{\sqrt{n}} = 0.31$$

The autocorrelation at lag 1 is significant.
Therefore, it is possible to refit the model by including an autoregressive term AR(1):

ESE1

Regression Analysis: X versus AR1; step

Method

Categorical predictor coding	(1; 0)
Rows unused	1

Regression Equation

step	
1	X = 7,73 + 0,361 AR1
2	X = 7,71 + 0,361 AR1
3	X = 7,64 + 0,361 AR1
4	X = 7,44 + 0,361 AR1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7,73	1,88	4,12	0,000	
AR1	0,361	0,160	2,27	0,030	1,62
step					
2	-0,0226	0,0687	-0,33	0,744	2,13
3	-0,0970	0,0732	-1,33	0,194	2,42
4	-0,2948	0,0682	-4,32	0,000	2,10

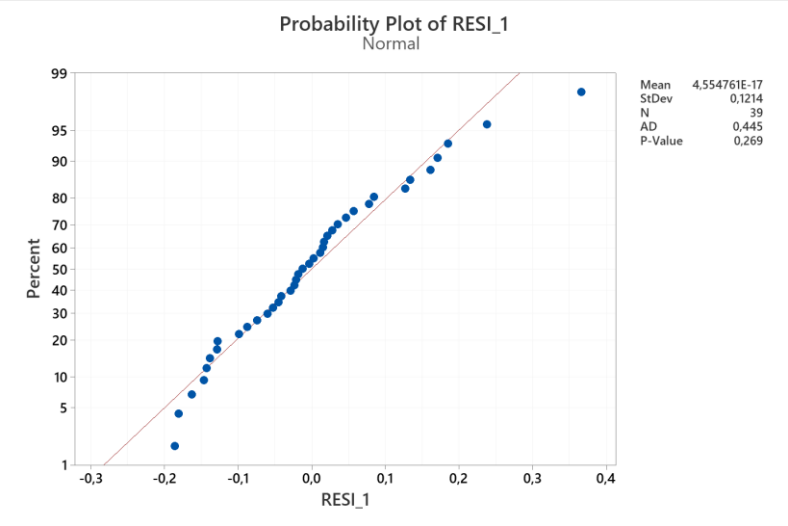
Model Summary

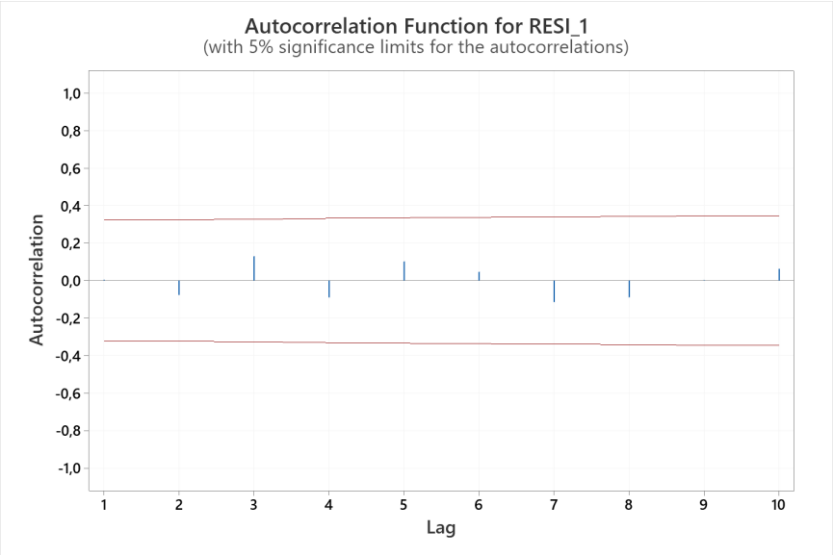
S	R-sq	R-sq(adj)	R-sq(pred)
0,128313	48,30%	42,22%	29,23%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0,52299	0,13075	7,94	0,000
AR1	1	0,08450	0,08450	5,13	0,030
step	3	0,49107	0,16369	9,94	0,000
Error	34	0,55979	0,01646		
Lack-of-Fit	31	0,51289	0,01654	1,06	0,569
Pure Error	3	0,04690	0,01563		
Total	38	1,08277			

Model residuals are normal and independent. The model is appropriate.





Test

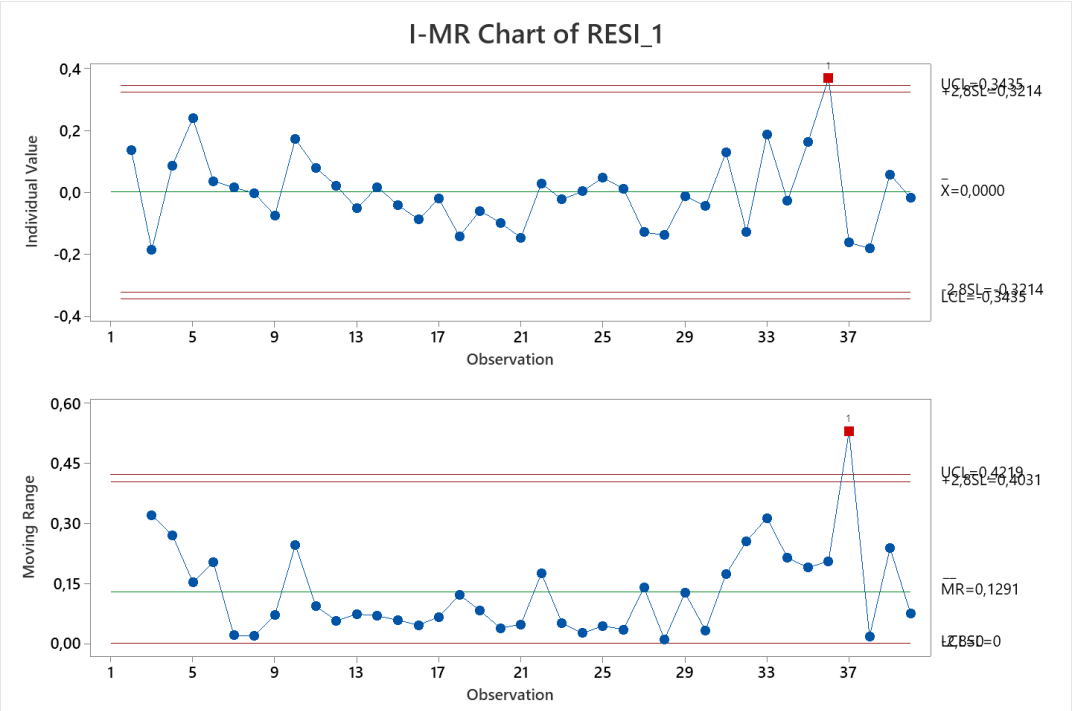
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs

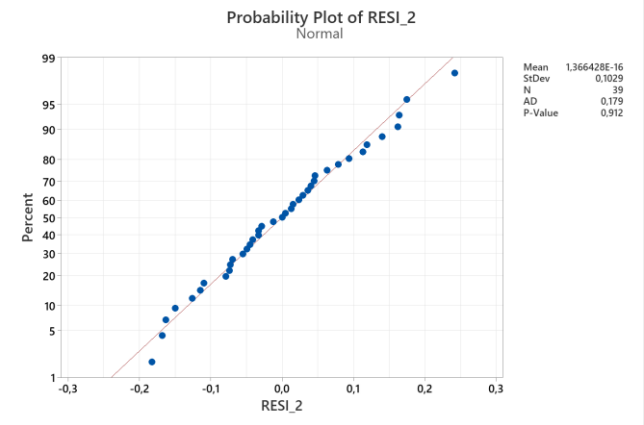
Observed	Expected	P-Value
20	20,38	0,900

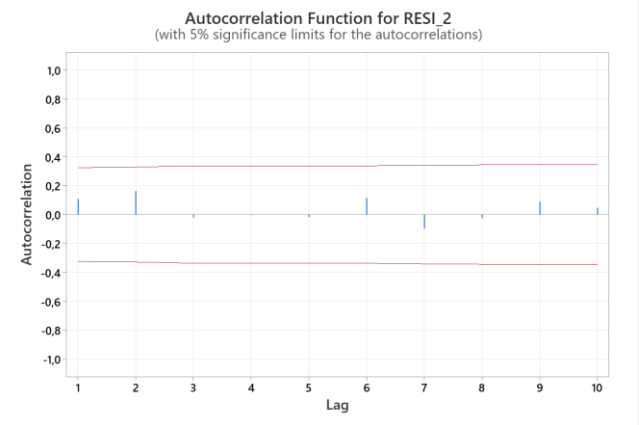
b)

With $ARL_0 = 200$, $z_{\alpha/2} = 2.807$. The I-MR control charts for model residuals are the following (do not consider the limits at k=3 in the figure).

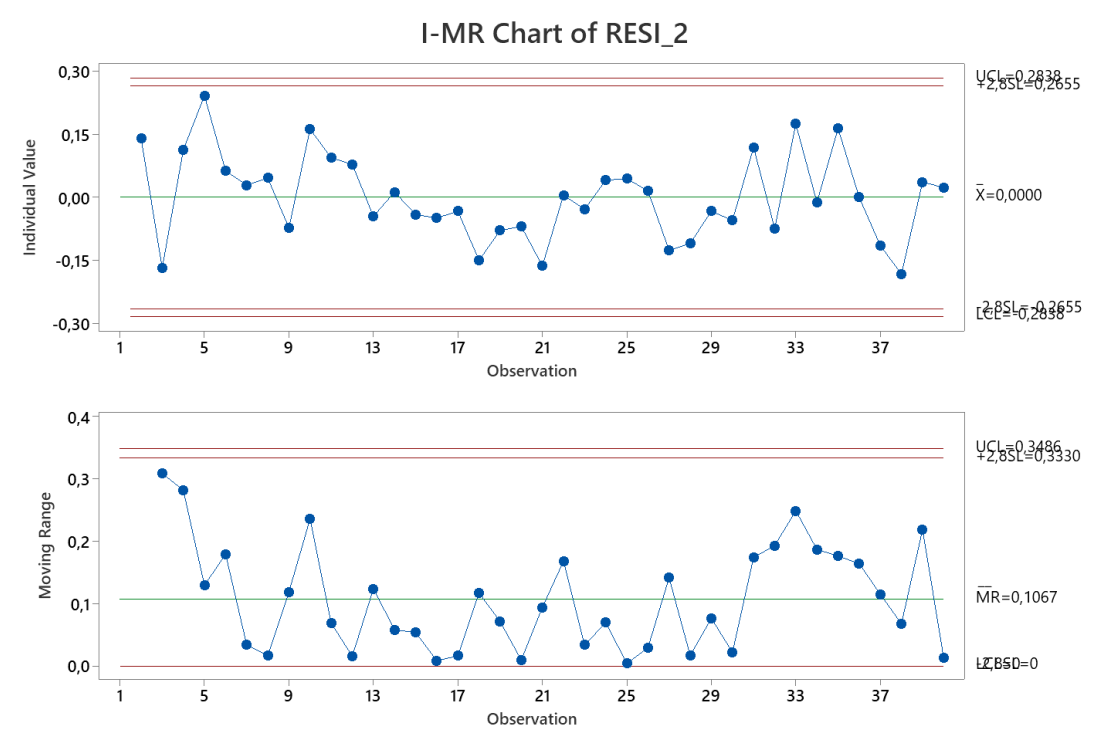


Observation 36 violates the control limits of both charts. Assuming the existence of an assignable cause, it is possible to introduce a dummy variable that is equal to 1 for this observation and 0 elsewhere. The new model is still appropriate, with normal and independent residuals:





The resulting control chart is the following:

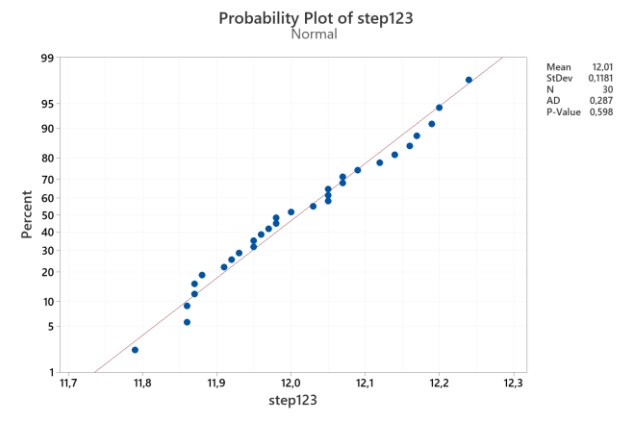


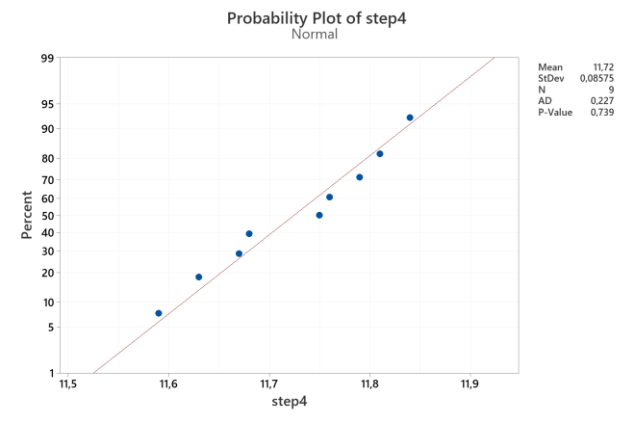
No further violation is present. The design phase is over.

c)

To make a test, it is possible to split the data into two vectors, one for spindle power measurements in the roughing operation (step 1, 2 and 3) and one for the measurements in the finishing operation (step 4). The out-of-control observation has been removed.

The two samples are normal and independent (although for the second sample the power of tests is quite low due to the small size of the sample).





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	15,93	0,142

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
5	5,44	0,748

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

We shall first test for equality of variances:

ESE1

Test and CI for Two Variances: step123; step4

Method

σ_1: standard deviation of step123
σ_2: standard deviation of step4
Ratio: σ_1/σ_2
The Bonett and Levene's methods are valid for any continuous distribution.

Descriptive Statistics

Variable	N	StDev	Variance	95% CI for σ
step123	30	0,118	0,014	(0,098; 0,152)
step4	9	0,086	0,007	(0,060; 0,156)

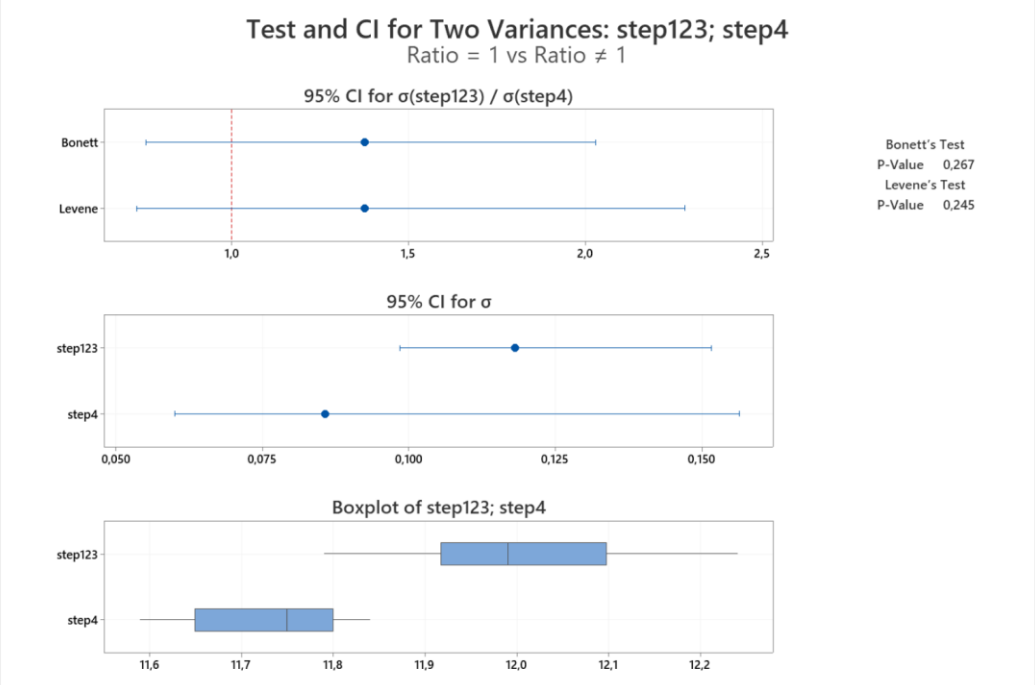
Ratio of Standard Deviations

Estimated Ratio	95% CI for Ratio using Bonett	95% CI for Ratio using Levene
1,37731	(0,758; 2,027)	(0,733; 2,279)

Test

Null hypothesis	H_0: $\sigma_1 / \sigma_2 = 1$
Alternative hypothesis	H_1: $\sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0,05$

Method	Test Statistic	DF1	DF2	P-Value
Bonett	*			0,267
Levene	1,39	1	37	0,245



There is no statistical difference between the two variances. Thus, it is possible to make the following two-sample t test with equal variances:

ESE1

Two-Sample T-Test and CI: step123; step4

Method

μ_1: population mean of step123
μ_2: population mean of step4
Difference: $\mu_1 - \mu_2$

Equal variances are assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
step123	30	12,010	0,118	0,022
step4	9	11,7244	0,0857	0,029

Estimation for Difference

Difference	Pooled StDev	95% Lower Bound for Difference
0,2859	0,1119	0,2141

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 > 0$

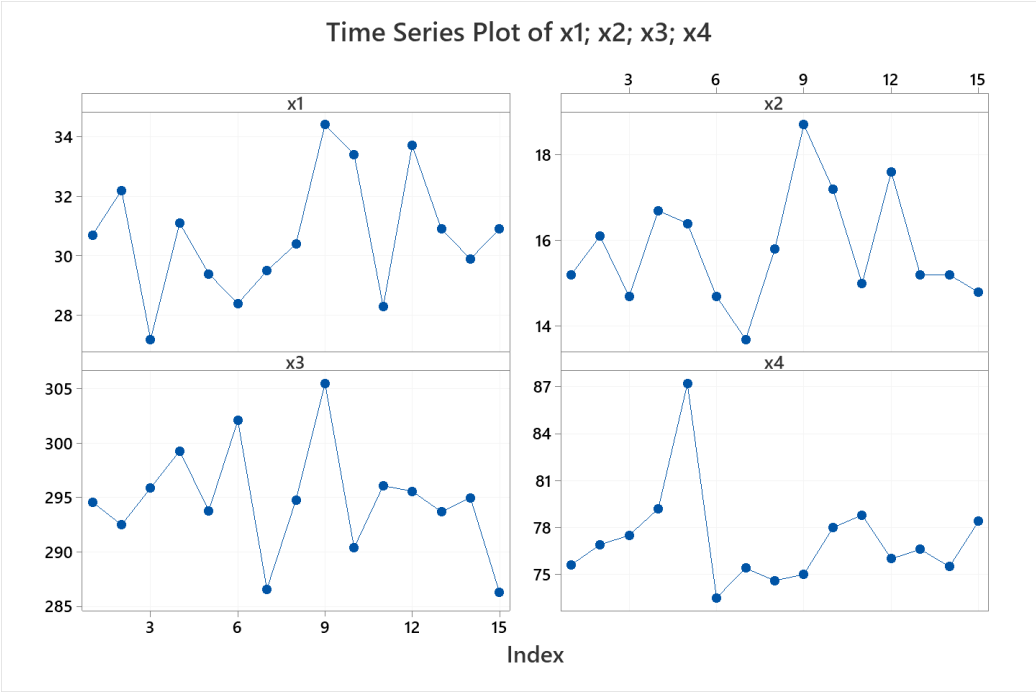
T-Value	DF	P-Value
6,72	37	0,000

The test confirms that the power consumption during the finishing operation is statistically lower than the one in the roughing operation.

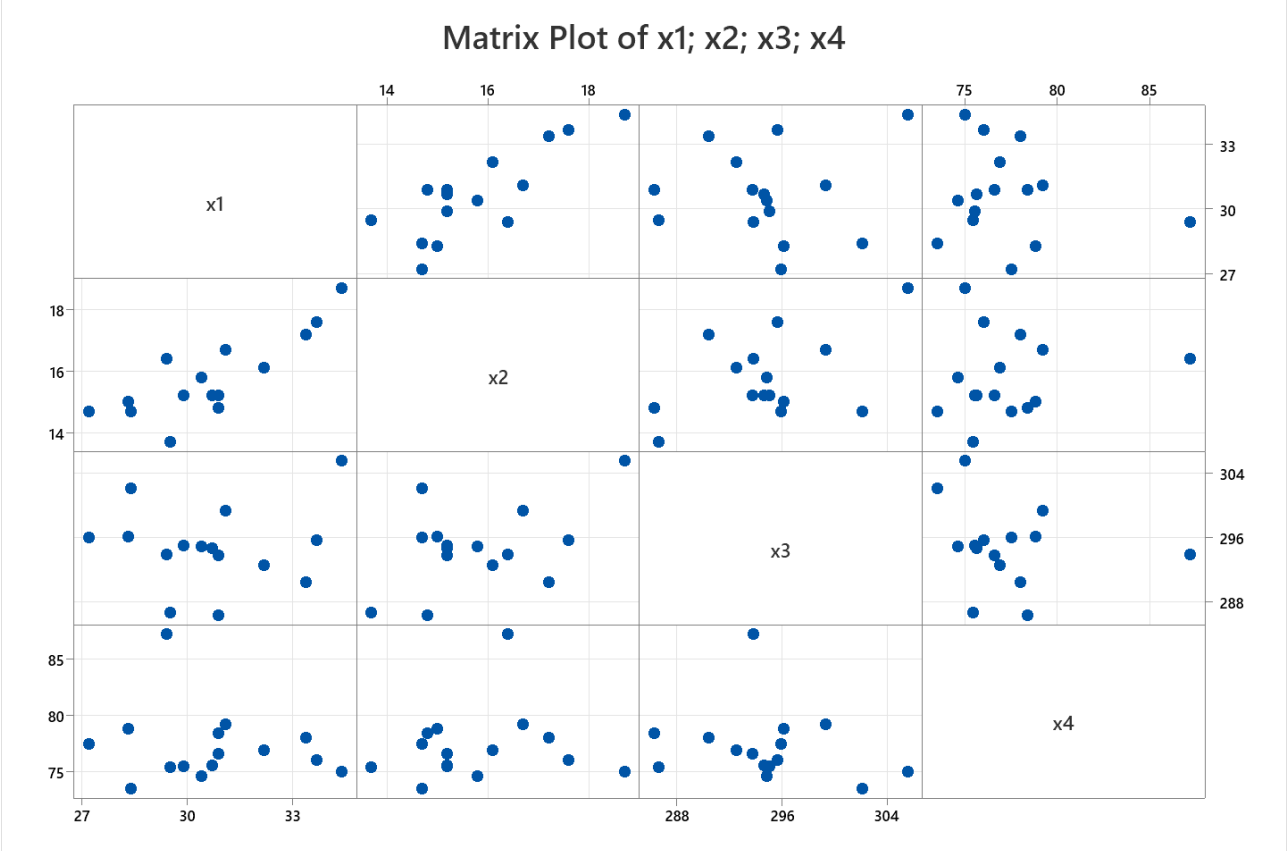
Exercise 2

a)

The time series plot of the four variables:



Their scatterplot:



The four variables are on different scales with different variances, as shown below. In this case, the PCA on the correlation matrix is the appropriate choice.

Statistics

Variable	Mean	StDev
x1	30,693	2,064
x2	15,800	1,321
x3	294,81	5,06
x4	77,213	3,214

b)

PCA on the sample correlation matrix:

ESE2

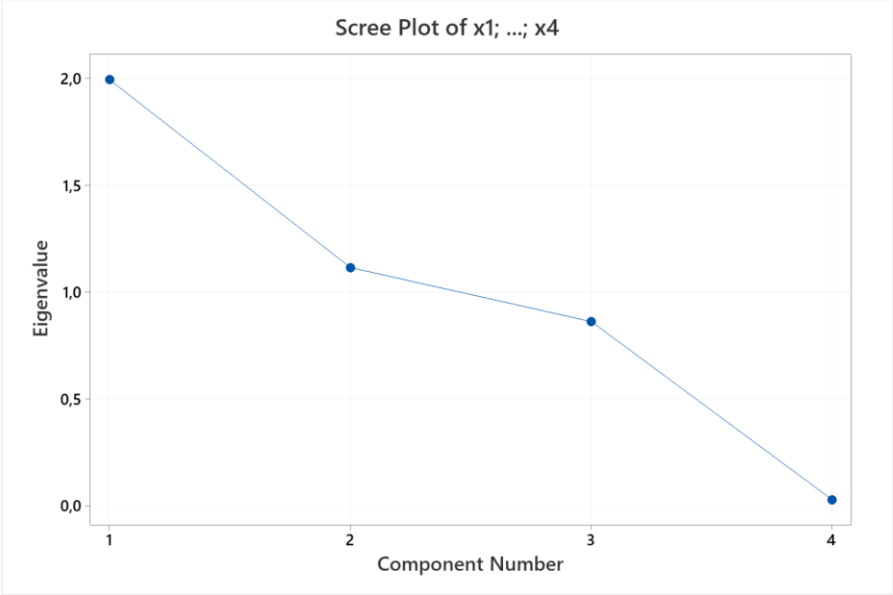
Principal Component Analysis: x1; x2; x3; x4

Eigenanalysis of the Correlation Matrix

Eigenvalue	1,9945	1,1151	0,8622	0,0282
Proportion	0,499	0,279	0,216	0,007
Cumulative	0,499	0,777	0,993	1,000

Eigenvectors

Variable	PC1	PC2	PC3	PC4
x1	0,605	0,159	0,518	0,583
x2	0,679	0,234	-0,088	-0,690
x3	0,406	-0,439	-0,725	0,342
x4	-0,091	0,852	-0,446	0,257

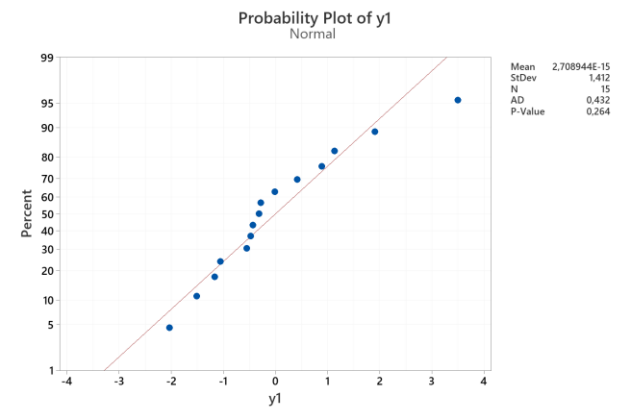


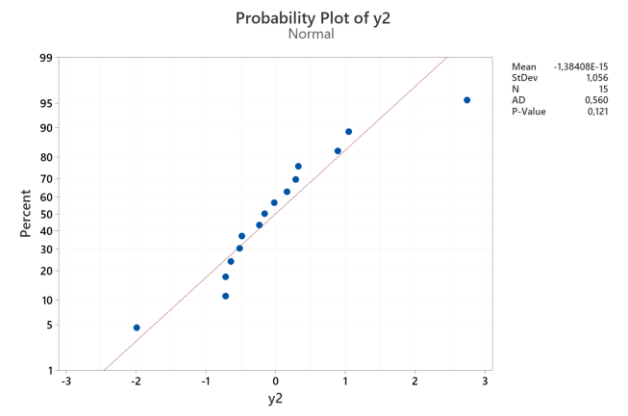
The number of PCs to retain to explain at least 75% of the overall data variability is 2.

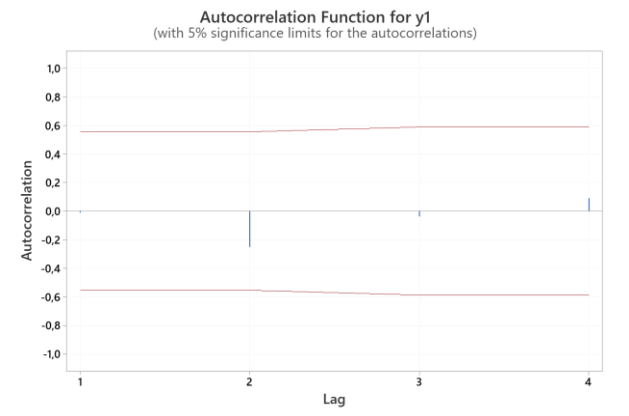
c)

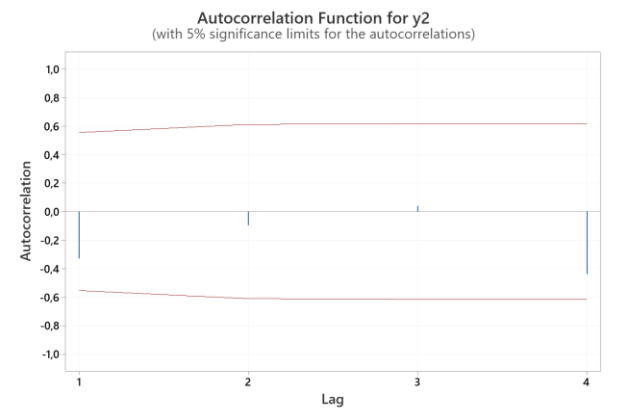
Before designing the T^2 control chart on the scores of the first 2 PCs, assumptions shall checked (marginal normality is assumed as a sufficient condition for multivariate normality).

The scores of the two PCs are normal and independent.









Test

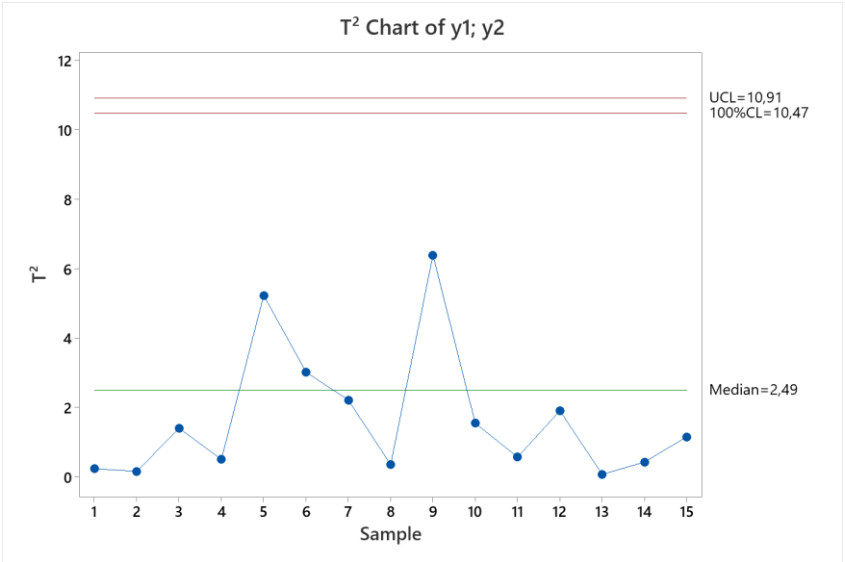
Null hypothesis H_0: The order of the data is random
Alternative hypothesis H_1: The order of the data is not random

Number of Runs

Variable	Observed	Expected	P-Value
y1	9	7,67	0,417
y2	10	8,20	0,313

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

With $ARL_0 = 200$, $\alpha = 0{,}005$. The T^2 control chart is the following (ignore the control limit at k=3).

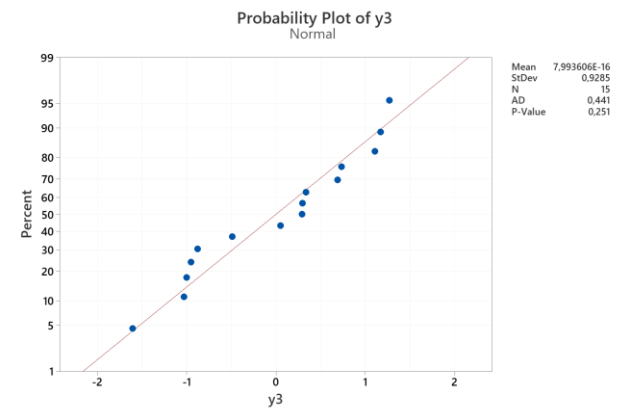


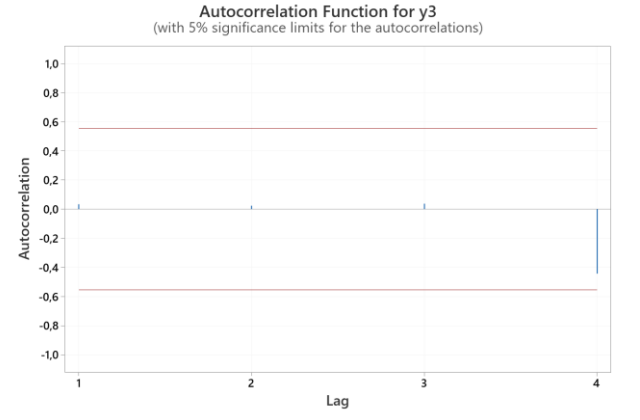
The process is in-control.

d)

By adding the third PC, about 99% of the overall variability is explained. Before re-designing the control chart it is necessary to check assumptions for the scores of the third PC as well.

They are normal and independent.





Test

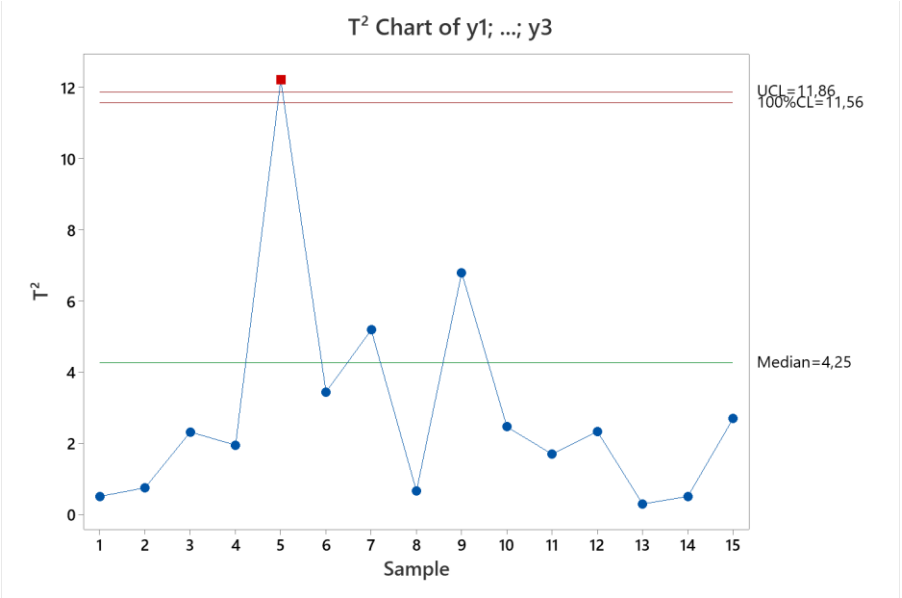
Null hypothesis H_0: The order of the data is random
Alternative hypothesis H_1: The order of the data is not random

Number of Runs

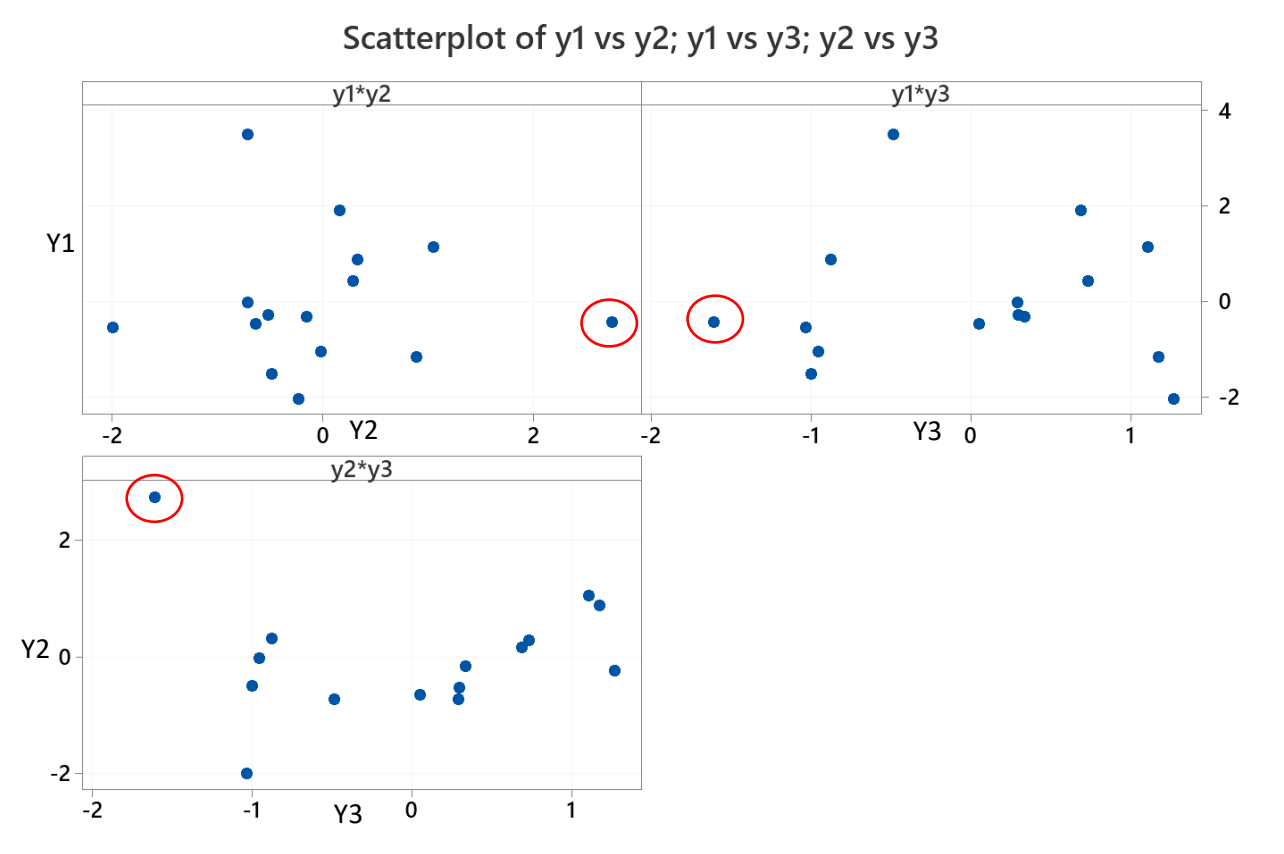
Observed	Expected	P-Value
7	8,20	0,502

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:



Now, sample 5 is out of control. The anomaly in sample 5 corresponds to a peak in variable x4. When only the first 2 PCs are monitored, sample 5 is a bit outlying along PC2, as shown in the figure below where sample 5 is highlighted in red (note that PC2 associates a high weight to variable x4), but not enough to signal an alarm. When the third PC is included, the sample 5 anomaly is emphasized in the space spanned by PC2 and PC3, as shown below.



Looking at the eigenvectors, PC2 associates a high weight to variable x4 and contrasts it against variable x3. PC3 is instead a contrast between variables x3 and x4 on one side, and variable x1 on the other side. PC1 associated a very low weight to variable x4, instead.

Eigenvectors

Variable	PC1	PC2	PC3	PC4
x1	0,605	0,159	0,518	0,583
x2	0,679	0,234	-0,088	-0,690
x3	0,406	-0,439	-0,725	0,342
x4	-0,091	0,852	-0,446	0,257

In the presence of the violation of the control limit at sample 5, a search for assignable causes shall be carried out. In the absence of further information, the control chart design phase is over.

e)

The new dataset consists of 100 samples and its sample variance-covariance matrix is the following:

$$\mathbf{S} = \begin{bmatrix} 1.4 & 0.75 & 0.1 & 0.5 \\ 0.75 & 1.3 & 1.7 & 0.5 \\ 0.1 & 1.7 & 6.6 & 0.3 \\ 0.5 & 0.5 & 0.3 & 3.6 \end{bmatrix}$$

In order to compare the new PCs with the old ones in terms of explained variance, we shall compute the sample correlation matrix, reminding that:

$$\rho_{ij} = \frac{\text{cov}(x_i x_j)}{\sqrt{V(x_i)V(x_j)}}$$

The correlation matrix is the following:

1	0,56	0,033	0,22
0,56	1	0,58	0,23
0,033	0,58	1	0,06
0,22	0,23	0,06	1

The eigenvalues of this matrix are:

λ1 = 1,92269

λ2 = 1,04890

λ3 = 0,81935

λ4 = 0,20906

The eigenvalues estimated with Table 2 data were:

λ1 = 1,99454

λ2 = 1,11505

λ3 = 0,86217

λ4 = 0,02823

Considering a familywise $\alpha = 0.05$ and assuming that normality and randomness hold for the new data as well, and that the two samples are independent, the following two tests can be designed to compare the explained variances of PC1 and PC2:

PC1:

ESE2

Test and CI for Two Variances

Method

σ_1^2: variance of Sample 1
σ_2^2: variance of Sample 2
Ratio: σ_1^2/σ_2^2
F method was used. This method is accurate for normal data only.

Descriptive Statistics

Sample	N	StDev	Variance	97,5% CI for σ
Sample 1	10	1,412	1,995	(0,924; 2,844)
Sample 2	100	1,387	1,923	(1,195; 1,647)

Ratio of Standard Deviations

Estimated Ratio	97,5% CI for Ratio using F
1,01851	(0,643; 2,075)

Test

Null hypothesis	H_0: $\sigma_1 / \sigma_2 = 1$
Alternative hypothesis	H_1: $\sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0,025$

Method	Test Statistic	DF1	DF2	P-Value
F	1,04	9	99	0,832

PC2:

ESE2

Test and CI for Two Variances

Method

σ_1^2: variance of Sample 1
σ_2^2: variance of Sample 2
Ratio: σ_1^2/σ_2^2
F method was used. This method is accurate for normal data only.

Descriptive Statistics

Sample	N	StDev	Variance	97,5% CI for σ
Sample 1	10	1,056	1,115	(0,691; 2,126)
Sample 2	100	1,024	1,049	(0,883; 1,216)

Ratio of Standard Deviations

Estimated Ratio	97,5% CI for Ratio using F
1,03105	(0,651; 2,101)

Test

Null hypothesis	H_0: $\sigma_1 / \sigma_2 = 1$
Alternative hypothesis	H_1: $\sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0,025$

Method	Test Statistic	DF1	DF2	P-Value
F	1,06	9	99	0,794

There is no statistical difference between the variances of the first 2 PCs estimated with data in Table 2 and data from the new extended measurement campaign.

Exercise 3

The eigenvalues and eigenvectors of the correlation matrix for the given data are:

Eigenvalues:

$\lambda 1 = 1,92269$

$\lambda 2 = 1,04890$

$\lambda 3 = 0,81935$

$\lambda 4 = 0,20906$

Eigenvectors:

Matrix EIG100

0,496346	0,456492	-0,563462	0,477249
0,667311	-0,141026	-0,135166	-0,718706
0,457667	-0,675402	0,283056	0,504234
0,314447	0,561746	0,764278	0,037997

Let $\boldsymbol{a}' = [a_1\ a_2\ a_3\ a_4]$ be a 4x1 vector of weights of a linear combination $\boldsymbol{a}'\mathbf{x}$ where $a_1 = a_2 = a_3 = a_4$.

Under the constraint $\boldsymbol{a}'\boldsymbol{a} = 1$, $a_1 = a_2 = a_3 = a_4 = 0.5$.

The variance of such linear combination is $\boldsymbol{V}(\boldsymbol{a}'\mathbf{x}) = \boldsymbol{a}'\mathbf{R}\boldsymbol{a}$, where **R** is the sample correlation matrix. The variance is $\boldsymbol{V}(\boldsymbol{a}'\mathbf{x}) = 1.8415$, which is lower than λ1 = 1,92269. Indeed, the first PC is, among all possible linear combinations, the one that maximizes the explained variance.

QUALITY DATA ANALYSIS

16/01/2023

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
- **MULTICHANCE STUDENTS SHALL SKIP: Exercise 1) point d, Exercise 3) point c.**

Exercise 1 (15 points)

A wine producer decides to apply statistical process monitoring tools to keep under control the quality of his production. During the barrel aging phase, he periodically measures four quality variables, x1, x2, x3, x4 taking a wine sample from randomly selected barrels. Data collected in successive samples are reported in Table 1.

Table 1

Sample	X1	X2	X3	X4
1	30,7	15,2	294,6	75,6
2	32,2	16,1	292,5	76,9
3	27,2	14,7	295,9	77,5
4	31,1	16,7	299,3	79,2
5	29,4	16,4	293,8	87,2
6	28,4	14,7	302,1	73,5
7	29,5	13,7	286,6	75,4
8	30,4	15,8	294,8	74,6
9	34,4	18,7	305,5	75
10	33,4	17,2	290,4	78
11	28,3	15	296,1	78,8
12	33,7	17,6	295,6	76
13	30,9	15,2	293,7	76,6
14	29,9	15,2	295	75,5
15	30,9	14,8	286,3	78,4

a) The wine maker is interested in using the PCA to analyze these data. Would it be more appropriate to use the sample variance-covariance matrix or the sample correlation matrix to estimate the principal components?
b) Estimate the PCA model for data in Table 1 by retaining the number of principal components required to capture at least 75% of the overall variability (report the eigenvalues and eigenvectors of retained PCs).
c) Based on the result of point b), design a Hotelling's T^2 control chart for the wine data with $ARL_0 = 200$. Can we conclude that the barrel aging process is stable and in-control?
d) How do the result of point c) changes if the Hotelling's T^2 control chart is designed using m+1 principal components, where m is the number of PCs used in point c)? Discuss the results.

e) The wine maker decides to extend the data collection for a longer period. Based on the collection of 100 samples, he estimates the following sample mean and variance-covariance matrix:

$$\bar{\boldsymbol{x}} = [28.8 \;\; 12 \;\; 288 \;\; 75], \mathbf{S} = \begin{bmatrix} 1.4 & 0.75 & 0.1 & 0.5 \\ 0.75 & 1.3 & 1.7 & 0.5 \\ 0.1 & 1.7 & 6.6 & 0.3 \\ 0.5 & 0.5 & 0.3 & 3.6 \end{bmatrix}$$

Design a statistical test to determine if the variances explained, respectively, by the first and second PC estimated from the new data are statistically different from the ones estimated from data in Table 1 (use a familywise confidence level $\alpha = 0.05$; assume that the new sample is random, normal and independent from the data sample in Table 1).

Exercise 2 (3 points)

Using the sample statistics defined in point e) of Exercise 1, report the eigenvalue and eigenvector corresponding to the first PC. Show that the variance explained by this first PC is higher than the variance explained by a simple linear combination of the four variables where equal weight is given to all the variables (for sake of comparison, remind the normalization constraint $\boldsymbol{a}'\boldsymbol{a} = 1$). Discuss the result.

Exercise 3 (15 points)

A company produces titanium impellers for the oil and gas sector. To meet sustainability targets, the company started monitoring the spindle power consumption during machining operations. The production of each impeller consists of four consecutive machining steps: step 1, 2 and 3 are roughing operations, while step 4 is the finishing operation. Table 2 includes power consumption data gathered during the production of ten consecutive impellers.

Table 2

Step	X (kW)	Step	X (kW)	Step	X (kW)	Step	X (kW)
1	12,05	3	12,12	1	11,79	3	12,09
2	12,2	4	11,84	2	12	4	11,68
3	11,86	1	11,96	3	11,95	1	12,14
4	11,81	2	12,05	4	11,76	2	12,07
1	12,24	3	11,95	1	12,03	3	12,16
2	12,17	4	11,67	2	12,07	4	12,2
3	12,05	1	11,93	3	11,87	1	11,98
4	11,79	2	11,88	4	11,59	2	11,86
1	11,92	3	11,87	1	11,91	3	11,98
2	12,19	4	11,63	2	11,97	4	11,75

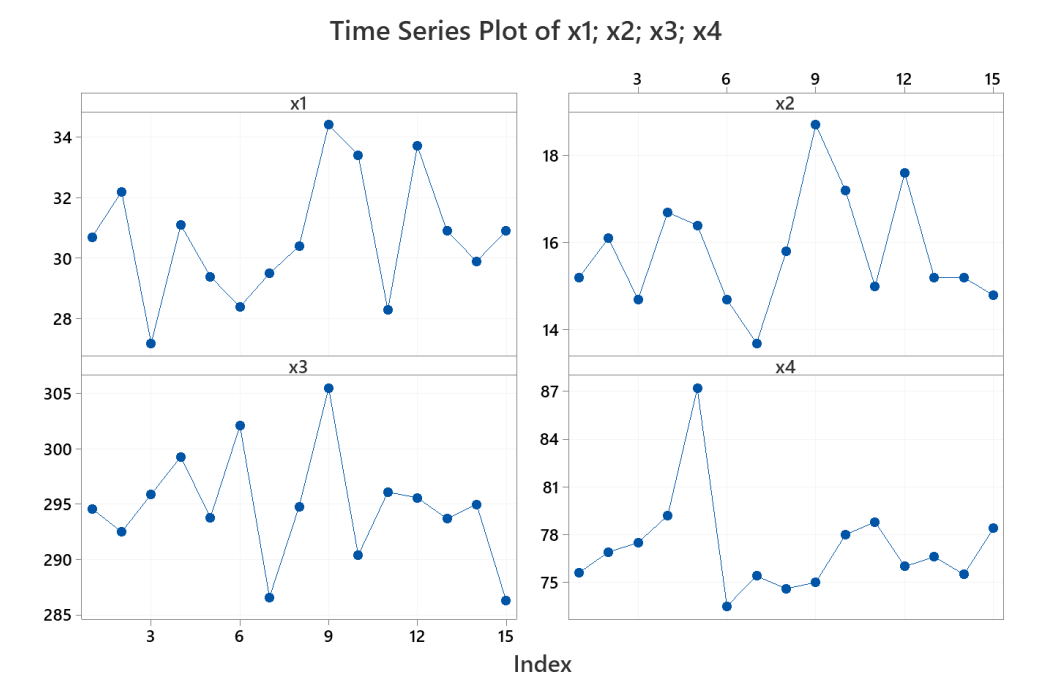
a) Find a suitable model for power consumption data in Table 2.
b) The head of the quality department is interested in using spindle power data to monitor the stability of the process. Based on the result at point a) design a suitable control chart with $ARL_0 = 200$. Assume the existence of assignable cause if out of control observations are present.
c) By using a suitable statistical test, determine whether the power consumption of the finishing operation is statistically lower than the power consumption during the roughing phase (exclude out of control observations identified in point b), if any).

Solutions

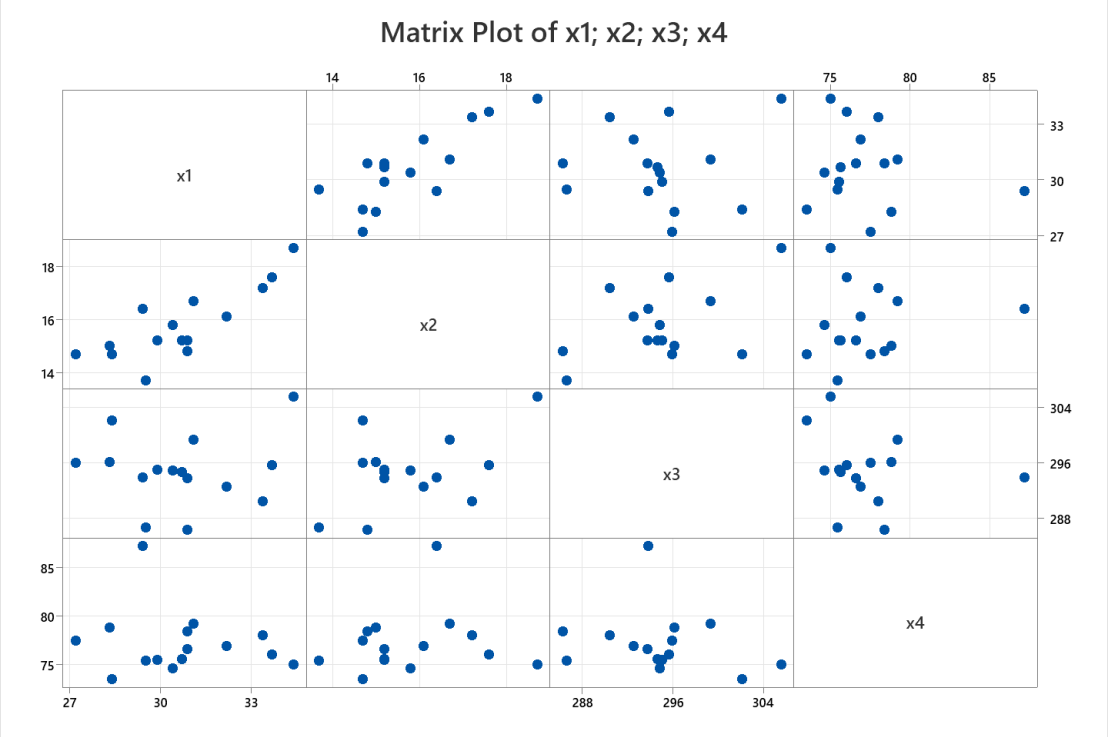
Exercise 1

a)

The time series plot of the four variables:



Their scatterplot:



The four variables are on different scales with different variances, as shown below. In this case, the PCA on the correlation matrix is the appropriate choice.

Statistics

Variable	Mean	StDev
x1	30,693	2,064
x2	15,800	1,321
x3	294,81	5,06
x4	77,213	3,214

b)

PCA on the sample correlation matrix:

ESE2

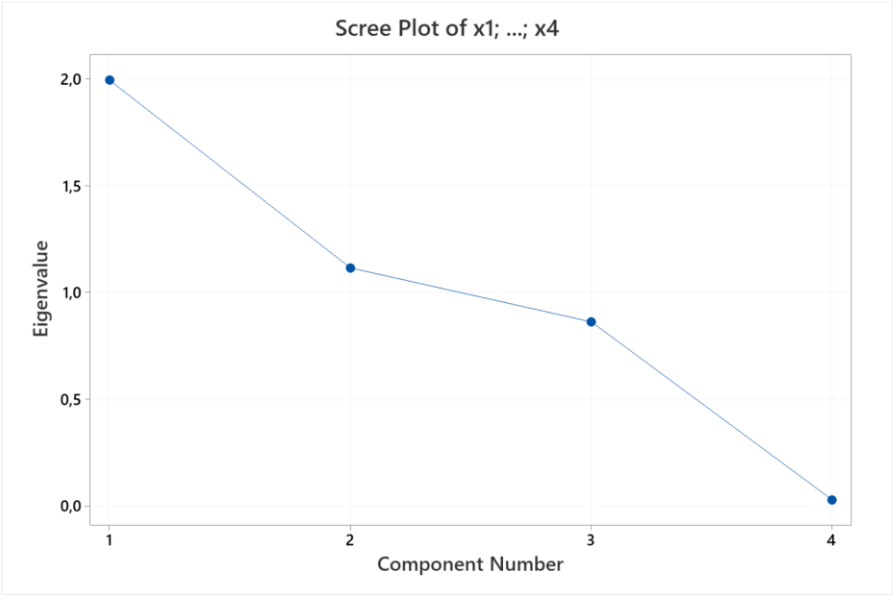
Principal Component Analysis: x1; x2; x3; x4

Eigenanalysis of the Correlation Matrix

Eigenvalue	1,9945	1,1151	0,8622	0,0282
Proportion	0,499	0,279	0,216	0,007
Cumulative	0,499	0,777	0,993	1,000

Eigenvectors

Variable	PC1	PC2	PC3	PC4
x1	0,605	0,159	0,518	0,583
x2	0,679	0,234	-0,088	-0,690
x3	0,406	-0,439	-0,725	0,342
x4	-0,091	0,852	-0,446	0,257

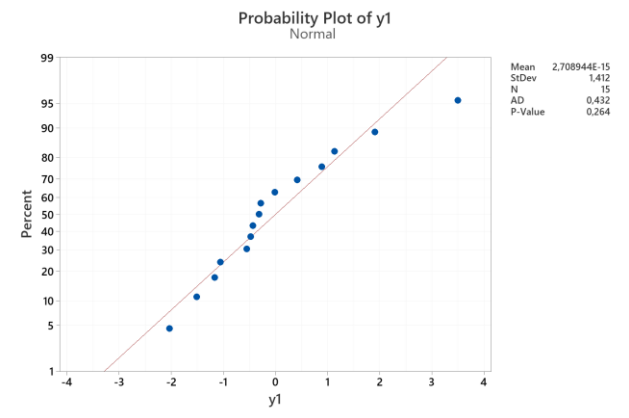


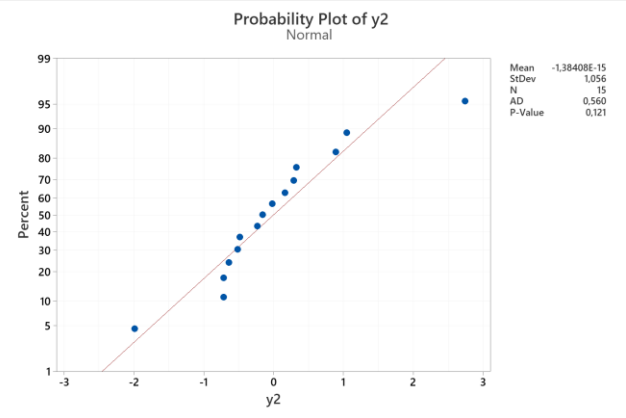
The number of PCs to retain to explain at least 75% of the overall data variability is 2.

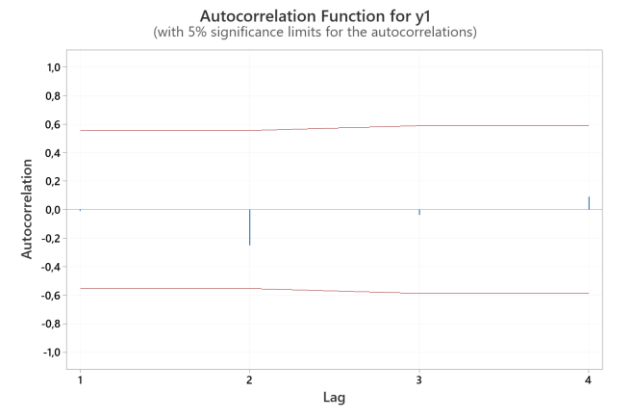
c)

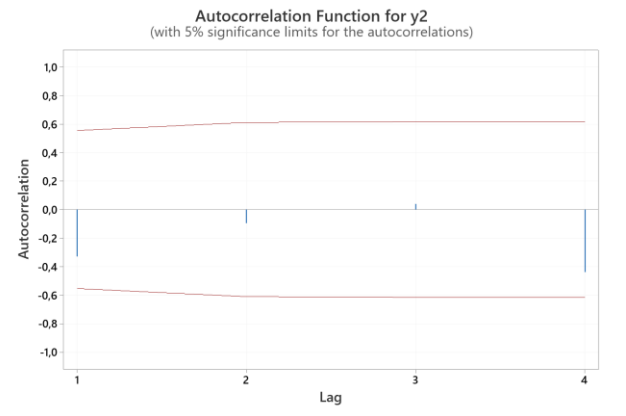
Before designing the T^2 control chart on the scores of the first 2 PCs, assumptions shall checked (marginal normality is assumed as a sufficient condition for multivariate normality).

The scores of the two PCs are normal and independent.









Test

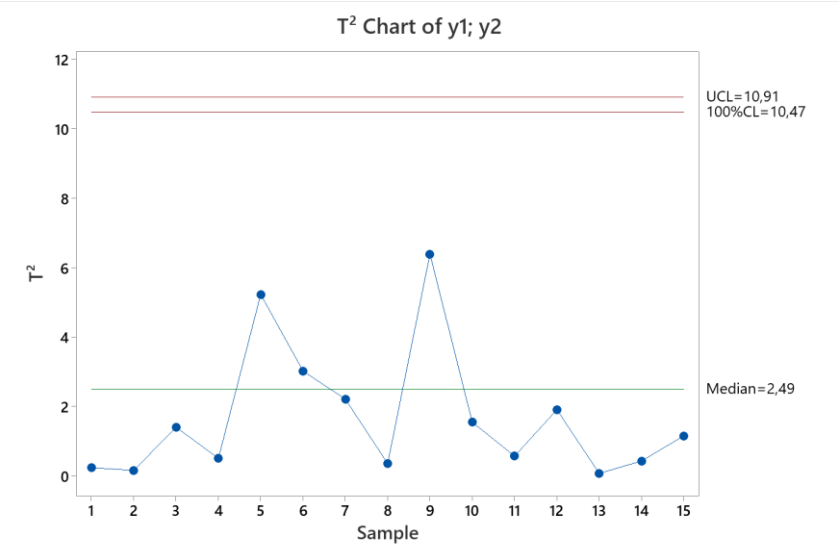
Null hypothesis H_0: The order of the data is random
Alternative hypothesis H_1: The order of the data is not random

Number of Runs

Variable	Observed	Expected	P-Value
y1	9	7,67	0,417
y2	10	8,20	0,313

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

With $ARL_0 = 200$, $\alpha = 0{,}005$. The T^2 control chart is the following (ignore the control limit at k=3).

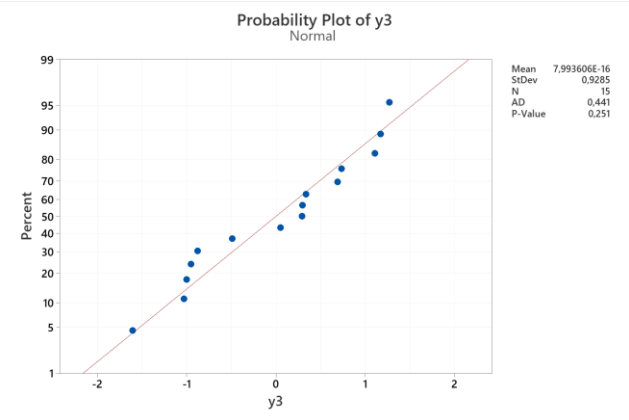


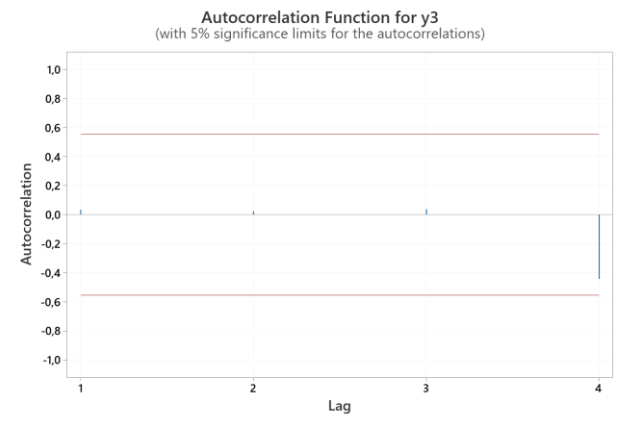
The process is in-control.

d)

By adding the third PC, about 99% of the overall variability is explained. Before re-designing the control chart it is necessary to check assumptions for the scores of the third PC as well.

They are normal and independent.





Test

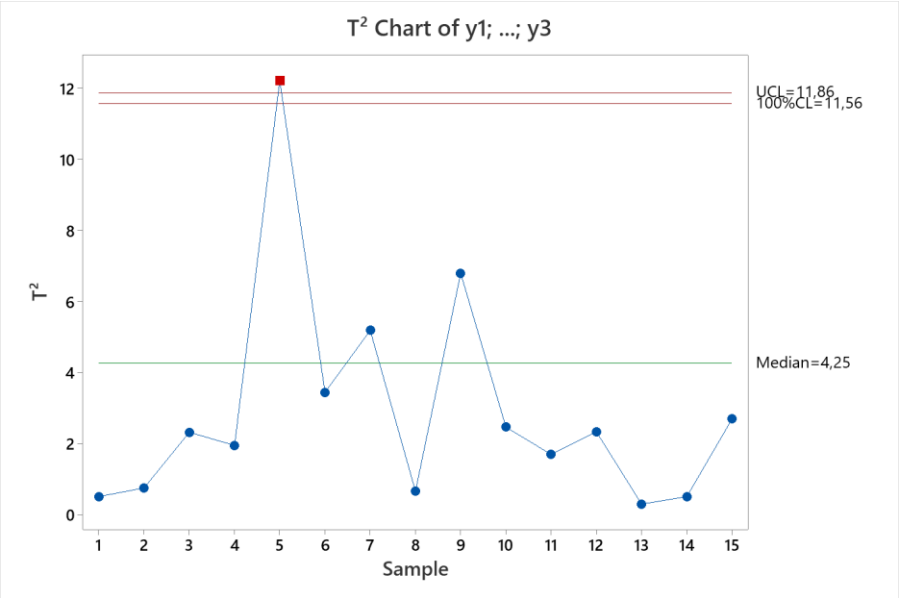
Null hypothesis H_0: The order of the data is random
Alternative hypothesis H_1: The order of the data is not random

Number of Runs

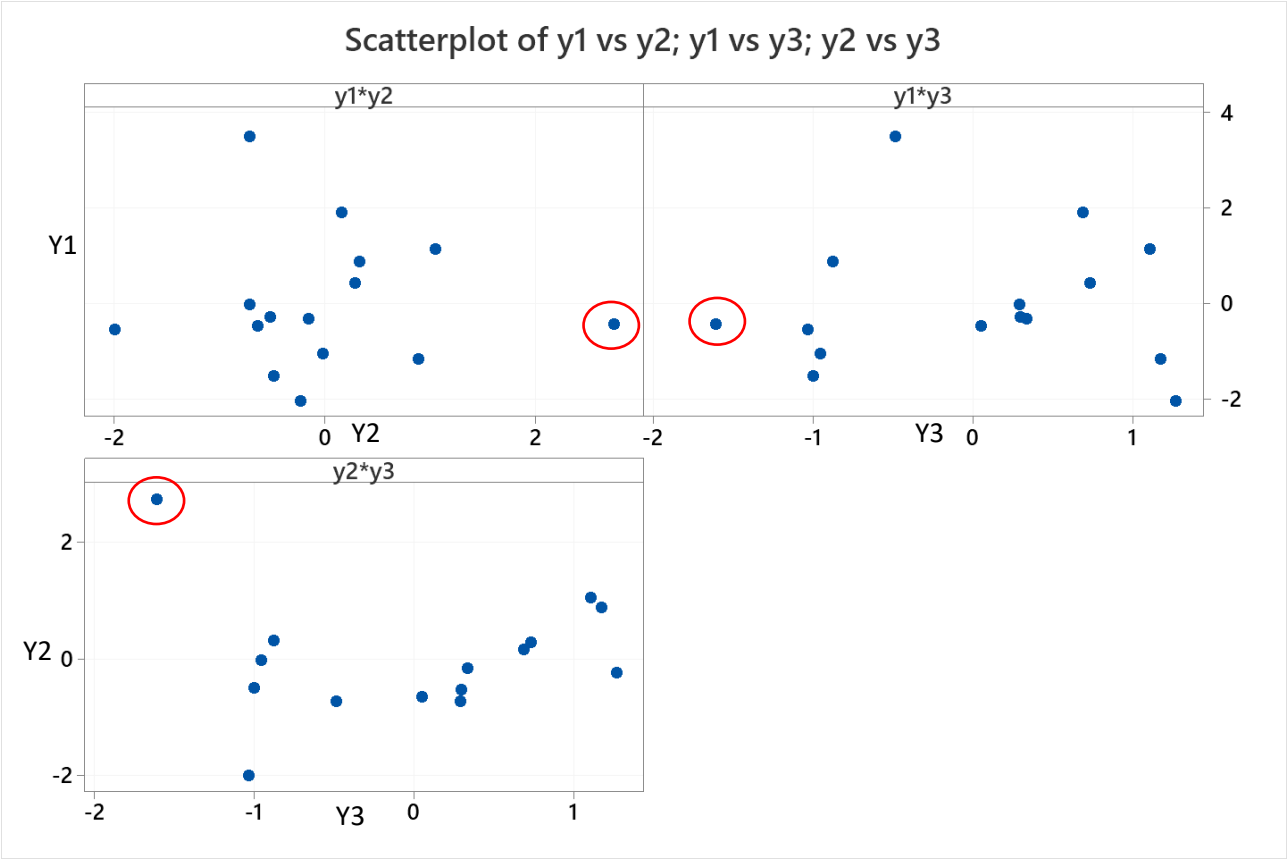
Observed	Expected	P-Value
7	8,20	0,502

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:



Now, sample 5 is out of control. The anomaly in sample 5 corresponds to a peak in variable x4. When only the first 2 PCs are monitored, sample 5 is a bit outlying along PC2, as shown in the figure below where sample 5 is highlighted in red (note that PC2 associates a high weight to variable x4), but not enough to signal an alarm. When the third PC is included, the sample 5 anomaly is emphasized in the space spanned by PC2 and PC3, as shown below.



Looking at the eigenvectors, PC2 associates a high weight to variable x4 and contrasts it against variable x3. PC3 is instead a contrast between variables x3 and x4 on one side, and variable x1 on the other side. PC1 associated a very low weight to variable x4, instead.

Eigenvectors

Variable	PC1	PC2	PC3	PC4
x1	0,605	0,159	0,518	0,583
x2	0,679	0,234	-0,088	-0,690
x3	0,406	-0,439	-0,725	0,342
x4	-0,091	0,852	-0,446	0,257

In the presence of the violation of the control limit at sample 5, a search for assignable causes shall be carried out. In the absence of further information, the control chart design phase is over.

e)

The new dataset consists of 100 samples and its sample variance-covariance matrix is the following:

$$\mathbf{S} = \begin{bmatrix} 1.4 & 0.75 & 0.1 & 0.5 \\ 0.75 & 1.3 & 1.7 & 0.5 \\ 0.1 & 1.7 & 6.6 & 0.3 \\ 0.5 & 0.5 & 0.3 & 3.6 \end{bmatrix}$$

In order to compare the new PCs with the old ones in terms of explained variance, we shall compute the sample correlation matrix, reminding that:

$$\rho_{ij} = \frac{\text{cov}(x_i x_j)}{\sqrt{V(x_i)V(x_j)}}$$

The correlation matrix is the following:

1	0,56	0,033	0,22
0,56	1	0,58	0,23
0,033	0,58	1	0,06
0,22	0,23	0,06	1

The eigenvalues of this matrix are:

λ1 = 1,92269

λ2 = 1,04890

λ3 = 0,81935

λ4 = 0,20906

The eigenvalues estimated with Table 2 data were:

λ1 = 1,99454

λ2 = 1,11505

λ3 = 0,86217

λ4 = 0,02823

Considering a familywise $\alpha = 0.05$ and assuming that normality and randomness hold for the new data as well, and that the two samples are independent, the following two tests can be designed to compare the explained variances of PC1 and PC2:

PC1:

ESE2

Test and CI for Two Variances

Method

σ_1^2: variance of Sample 1
σ_2^2: variance of Sample 2
Ratio: σ_1^2/σ_2^2
F method was used. This method is accurate for normal data only.

Descriptive Statistics

Sample	N	StDev	Variance	97,5% CI for σ
Sample 1	10	1,412	1,995	(0,924; 2,844)
Sample 2	100	1,387	1,923	(1,195; 1,647)

Ratio of Standard Deviations

Estimated Ratio	97,5% CI for Ratio using F
1,01851	(0,643; 2,075)

Test

Null hypothesis	H_0: $\sigma_1 / \sigma_2 = 1$
Alternative hypothesis	H_1: $\sigma_1 / \sigma_2 \neq 1$
Significance level	α = 0,025

Method	Test Statistic	DF1	DF2	P-Value
F	1,04	9	99	0,832

PC2:

ESE2

Test and CI for Two Variances

Method

σ_1^2: variance of Sample 1
σ_2^2: variance of Sample 2
Ratio: σ_1^2/σ_2^2
F method was used. This method is accurate for normal data only.

Descriptive Statistics

Sample	N	StDev	Variance	97,5% CI for σ
Sample 1	10	1,056	1,115	(0,691; 2,126)
Sample 2	100	1,024	1,049	(0,883; 1,216)

Ratio of Standard Deviations

Estimated Ratio	97,5% CI for Ratio using F
1,03105	(0,651; 2,101)

Test

Null hypothesis	H_0: $\sigma_1 / \sigma_2 = 1$
Alternative hypothesis	H_1: $\sigma_1 / \sigma_2 \neq 1$
Significance level	α = 0,025

Method	Test Statistic	DF1	DF2	P-Value
F	1,06	9	99	0,794

There is no statistical difference between the variances of the first 2 PCs estimated with data in Table 2 and data from the new extended measurement campaign.

Exercise 2

The eigenvalues and eigenvectors of the correlation matrix for the given data are:

Eigenvalues:

λ1 = 1,92269

λ2 = 1,04890

λ3 = 0,81935

λ4 = 0,20906

Eigenvectors:

Matrix EIG100

0,496346	0,456492	-0,563462	0,477249
0,667311	-0,141026	-0,135166	-0,718706
0,457667	-0,675402	0,283056	0,504234
0,314447	0,561746	0,764278	0,037997

Let $\boldsymbol{a}' = [a_1\ a_2\ a_3\ a_4]$ be a 4x1 vector of weights of a linear combination $\boldsymbol{a}'\mathbf{x}$ where $a_1 = a_2 = a_3 = a_4$.

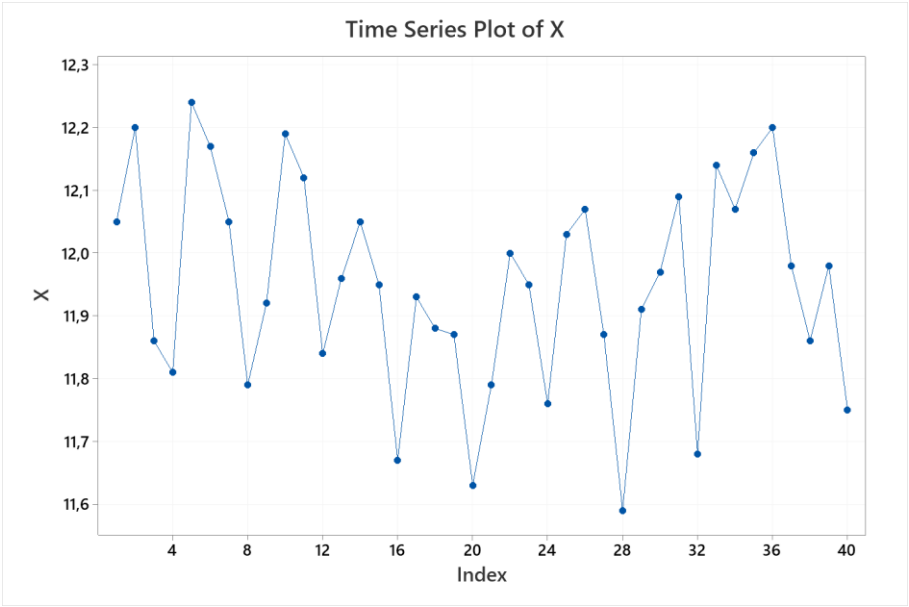
Under the constraint $\boldsymbol{a}'\boldsymbol{a} = 1$, $a_1 = a_2 = a_3 = a_4 = 0.5$.

The variance of such linear combination is $\boldsymbol{V}(\boldsymbol{a}'\mathbf{x}) = \boldsymbol{a}'\mathbf{R}\boldsymbol{a}$, where **R** is the sample correlation matrix. The variance is $\boldsymbol{V}(\boldsymbol{a}'\mathbf{x}) = 1.8415$, which is lower than λ1 = 1,92269. Indeed, the first PC is, among all possible linear combinations, the one that maximizes the explained variance.

Exercise 3

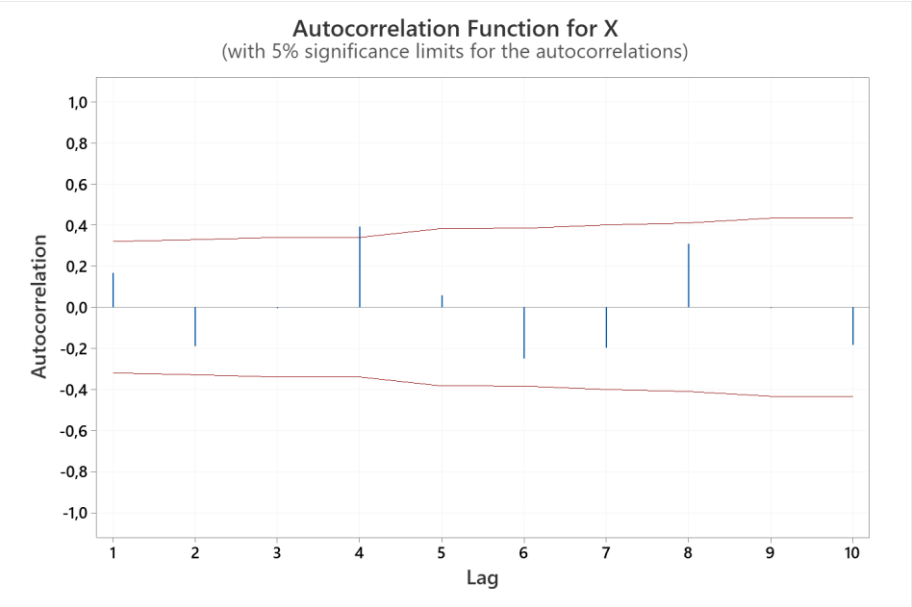
a)

Time series plot of the spindle power data:

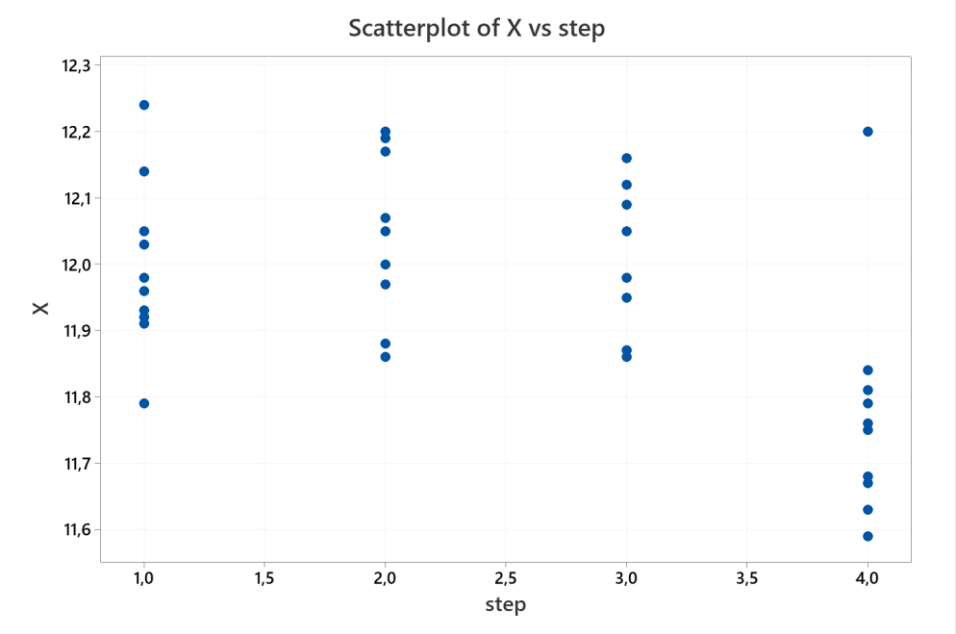


There is a meandering pattern with a seasonal drop of the spindle power in correspondence of step 4.

This lag 4 effect is also visible from the SACF:



The way in which the power varies along the different process steps is shown in the following scatter plot:



A part from one apparent outlying value, step 4 (finishing) yields a lower power consumption, as expected.

Based on this, it would be possible to fit a model of the spindle power using as regressor the categorical variable “step” as follows:

ESE1

Regression Analysis: X versus step

Method

Categorical predictor coding (1; 0)

Regression Equation

X = 11,9950 + 0,0 step_1 + 0,0510 step_2 - 0,0050 step_3 - 0,2230 step_4

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	11,9950	0,0424	282,80	0,000	
step					
2	0,0510	0,0600	0,85	0,401	1,50
3	-0,0050	0,0600	-0,08	0,934	1,50
4	-0,2230	0,0600	-3,72	0,001	1,50

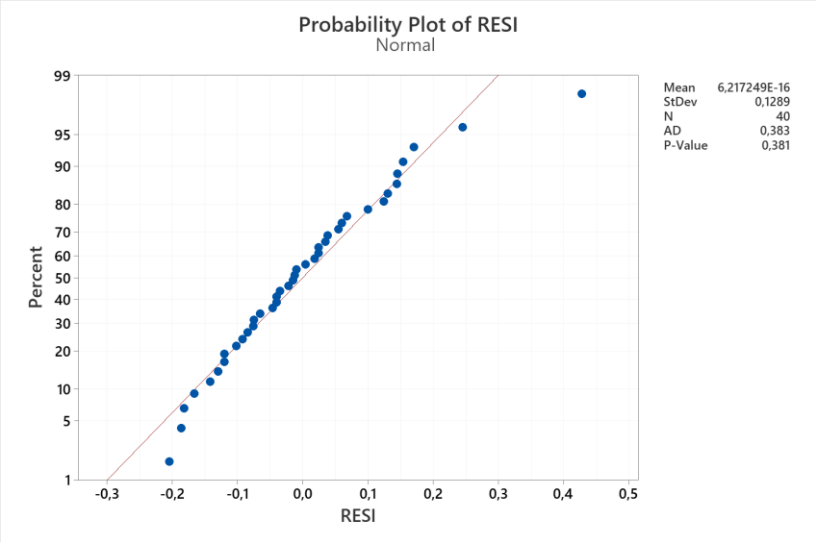
Model Summary

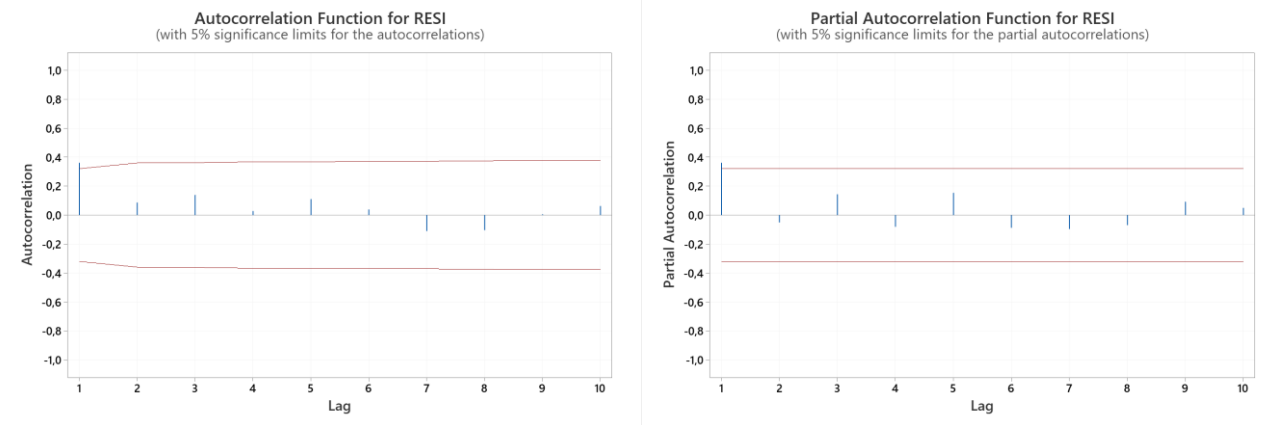
S	R-sq	R-sq(adj)	R-sq(pred)
0,134128	40,74%	35,80%	26,84%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	0,4452	0,14841	8,25	0,000
step	3	0,4452	0,14841	8,25	0,000
Error	36	0,6477	0,01799		
Total	39	1,0929			

The residuals are normal but not independent:





Bartlett's test at lag = 1 (95% confidence):

$$|r_k| = 0.361$$

$$\frac{z_{\alpha/2}}{\sqrt{n}} = 0.31$$

The autocorrelation at lag 1 is significant.
Therefore, it is possible to refit the model by including an autoregressive term AR(1):

ESE1

Regression Analysis: X versus AR1; step

Method

Categorical predictor coding	(1; 0)
Rows unused	1

Regression Equation

step	
1	X = 7,73 + 0,361 AR1
2	X = 7,71 + 0,361 AR1
3	X = 7,64 + 0,361 AR1
4	X = 7,44 + 0,361 AR1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7,73	1,88	4,12	0,000	
AR1	0,361	0,160	2,27	0,030	1,62
step					
2	-0,0226	0,0687	-0,33	0,744	2,13
3	-0,0970	0,0732	-1,33	0,194	2,42
4	-0,2948	0,0682	-4,32	0,000	2,10

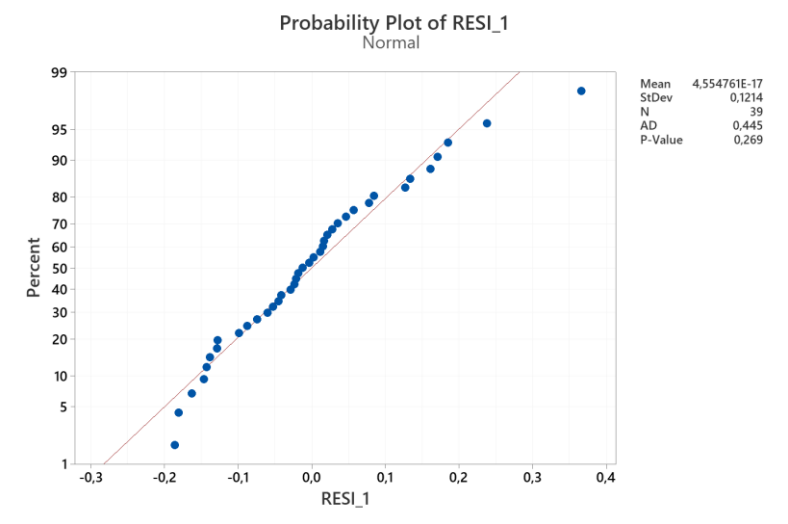
Model Summary

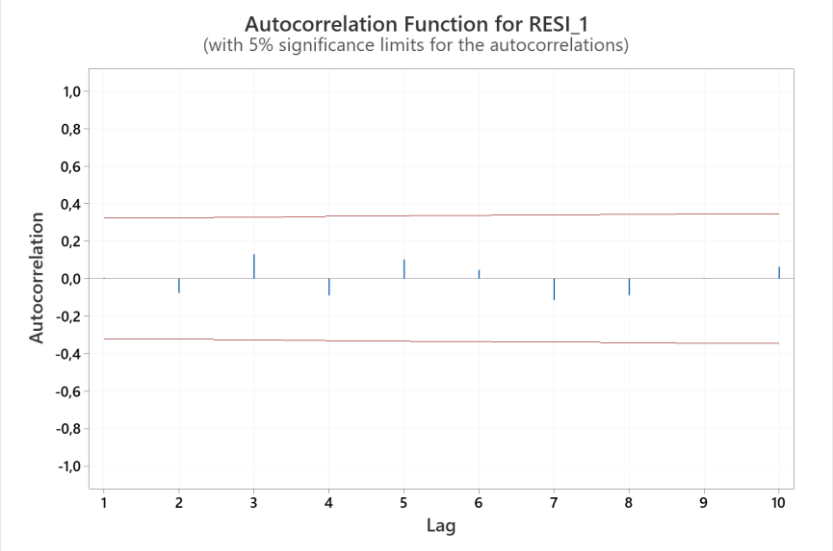
S	R-sq	R-sq(adj)	R-sq(pred)
0,128313	48,30%	42,22%	29,23%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0,52299	0,13075	7,94	0,000
AR1	1	0,08450	0,08450	5,13	0,030
step	3	0,49107	0,16369	9,94	0,000
Error	34	0,55979	0,01646		
Lack-of-Fit	31	0,51289	0,01654	1,06	0,569
Pure Error	3	0,04690	0,01563		
Total	38	1,08277			

Model residuals are normal and independent. The model is appropriate.





Test

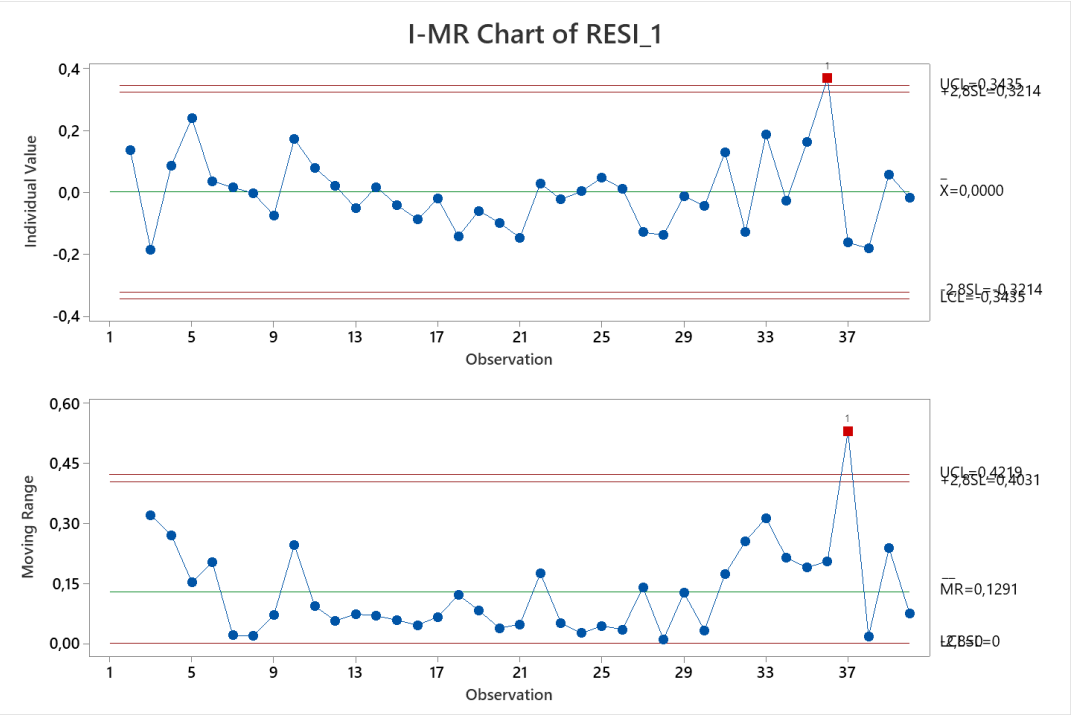
Null hypothesis	H₀: The order of the data is random
Alternative hypothesis	H₁: The order of the data is not random

Number of Runs

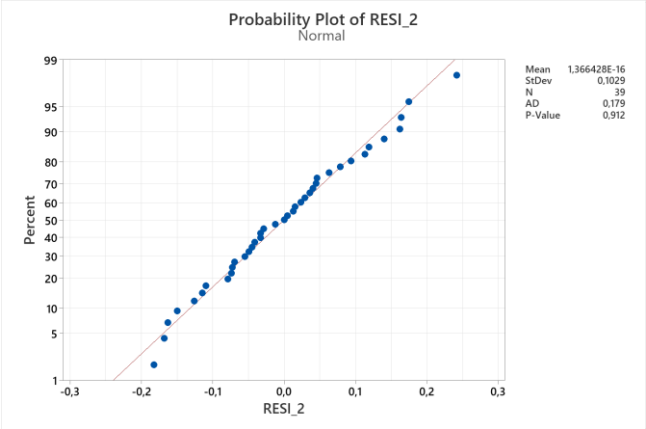
Observed	Expected	P-Value
20	20,38	0,900

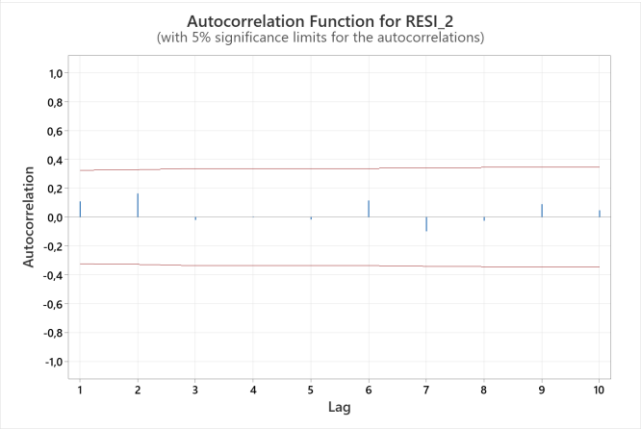
b)

With $ARL_0 = 200$, $z_{\alpha/2} = 2.807$. The I-MR control charts for model residuals are the following (do not consider the limits at k=3 in the figure).

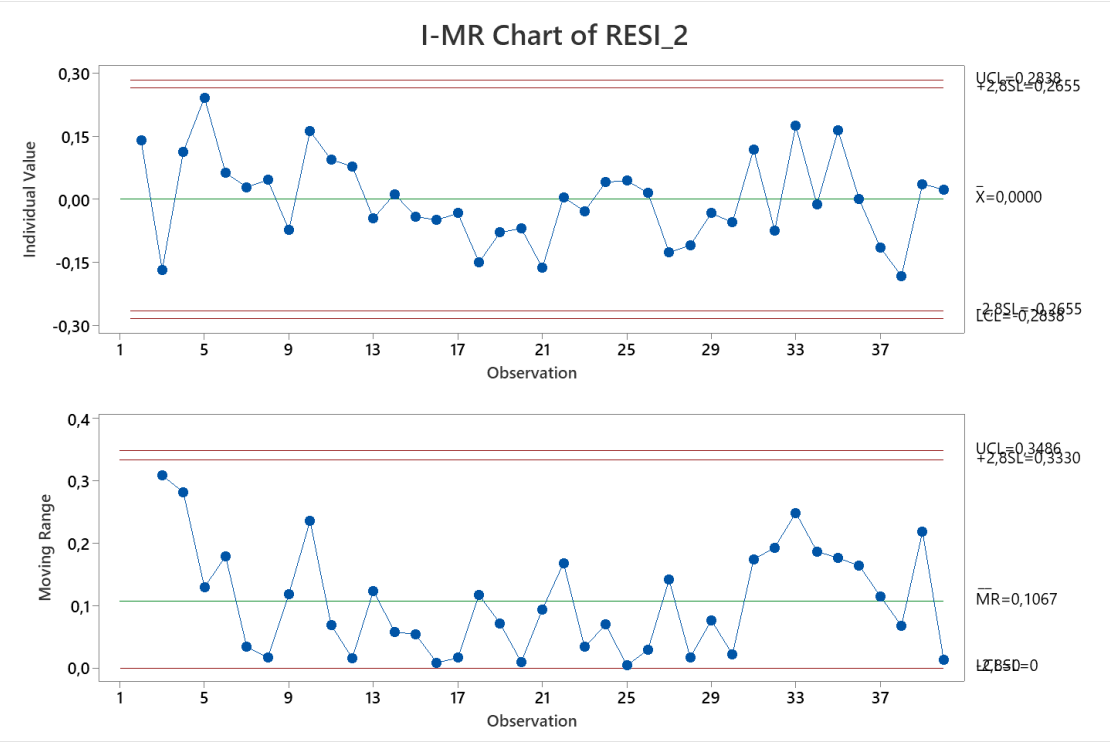


Observation 36 violates the control limits of both charts. Assuming the existence of an assignable cause, it is possible to introduce a dummy variable that is equal to 1 for this observation and 0 elsewhere. The new model is still appropriate, with normal and independent residuals:





The resulting control chart is the following:

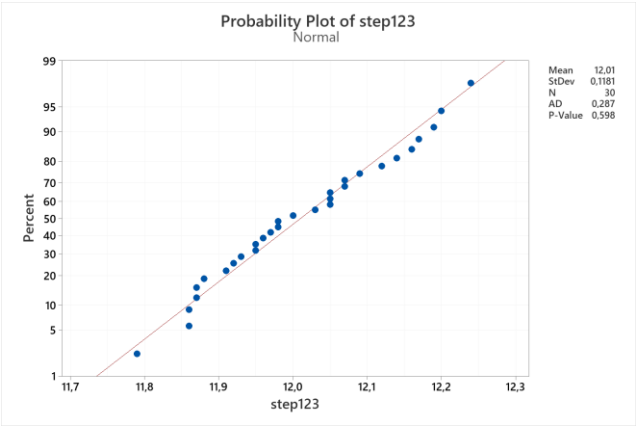


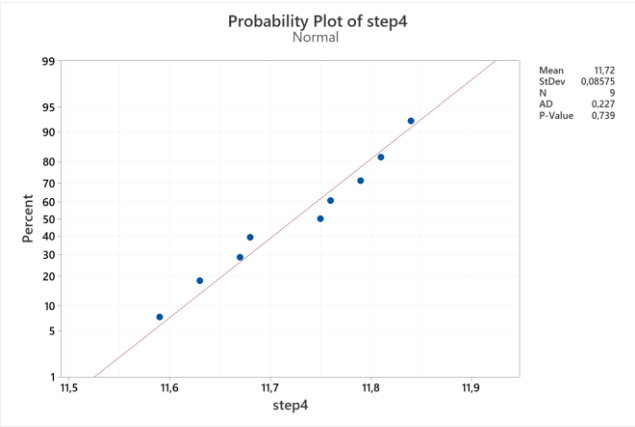
No further violation is present. The design phase is over.

c)

To make a test, it is possible to split the data into two vectors, one for spindle power measurements in the roughing operation (step 1, 2 and 3) and one for the measurements in the finishing operation (step 4). The out-of-control observation has been removed.

The two samples are normal and independent (although for the second sample the power of tests is quite low due to the small size of the sample).





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	15,93	0,142

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
5	5,44	0,748

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

We shall first test for equality of variances:

ESE1

Test and CI for Two Variances: step123; step4

Method

σ_1: standard deviation of step123
σ_2: standard deviation of step4
Ratio: σ_1/σ_2
The Bonett and Levene's methods are valid for any continuous distribution.

Descriptive Statistics

Variable	N	StDev	Variance	95% CI for σ
step123	30	0,118	0,014	(0,098; 0,152)
step4	9	0,086	0,007	(0,060; 0,156)

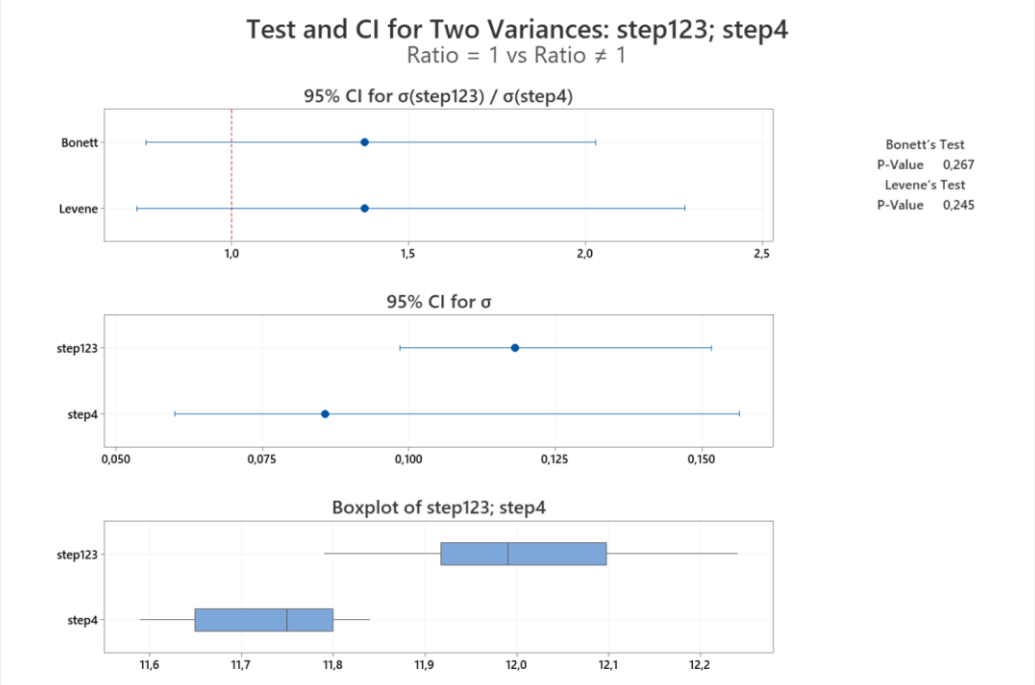
Ratio of Standard Deviations

Estimated Ratio	95% CI for Ratio using Bonett	95% CI for Ratio using Levene
1,37731	(0,758; 2,027)	(0,733; 2,279)

Test

Null hypothesis	H_0: $\sigma_1 / \sigma_2 = 1$
Alternative hypothesis	H_1: $\sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0,05$

Method	Test Statistic	DF1	DF2	P-Value
Bonett	*			0,267
Levene	1,39	1	37	0,245



There is no statistical difference between the two variances. Thus, it is possible to make the following two-sample t test with equal variances:

ESE1

Two-Sample T-Test and CI: step123; step4

Method

μ_1: population mean of step123
μ_2: population mean of step4
Difference: $\mu_1 - \mu_2$

Equal variances are assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
step123	30	12,010	0,118	0,022
step4	9	11,7244	0,0857	0,029

Estimation for Difference

Difference	Pooled StDev	95% Lower Bound for Difference
0,2859	0,1119	0,2141

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 > 0$

T-Value	DF	P-Value
6,72	37	0,000

The test confirms that the power consumption during the finishing operation is statistically lower than the one in the roughing operation.

QUALITY DATA ANALYSIS

09/02/2024

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 measured diameters of the shafts produced over two shifts are reported in `diameter_phase1.csv`. The columns of the table report a sequential index ('idx' column), the measurements in mm ('diam' column) and the shift ('shift' column) at which the data were collected.

1) Find an adequate model to fit the data.
2) Design the appropriate control charts to monitor the diameter of the shafts (use K = 3). *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 for the following 20 shafts produced during shift 2 (stored in `diameter_phase2.csv`) are in control. Report the index of the OOC points, if any.

Exercise 2 (15 points)

A German company operating in the aerospace sector needs to monitor the manufacturing process for a new type of titanium bracket. The quality characteristic of interest is the Brinell hardness. Four hardness measurements are performed in four pre-defined locations of the component, and parts are randomly picked up from the shop floor and inspected every two hours. Data to be used for control chart design are reported in the file `AERO_phase1.csv`. Each column refers to one location where the hardness measurement is performed. *Assume the measurements within each sample were performed following the same order shown in the provided table.*

1) Check the assumptions and discuss the result.
2) Design a statistical test to check if the hardness in location 1 is statistically higher than the hardness in location 2. Discuss the result.
3) Design a suitable univariate control charting method for these data, using K = 3. *In case of violation of control limits, assume no assignable cause is known.*
4) Using the control chart designed in point 3) determine if the new Brinell hardness measurements in `AERO_phase2.csv` are in control or not. Discuss the result.

Exercise 3 (4 points)

Question 1)

The Shewhart control chart for the mean of a process, where the data are known to be normally distributed, sets the control limits (i.e., LCL and UCL) to be K standard deviations away from the center line, for some constant, $K > 0$. If a user will decide to replace K by $K_1 > K$, then which of the following statements will be valid for the control chart performance:

a) The false alarms will increase.
b) The false alarms will decrease.
c) The out-of-control detection power will increase.
d) We cannot tell from the above information only.

Question 2)

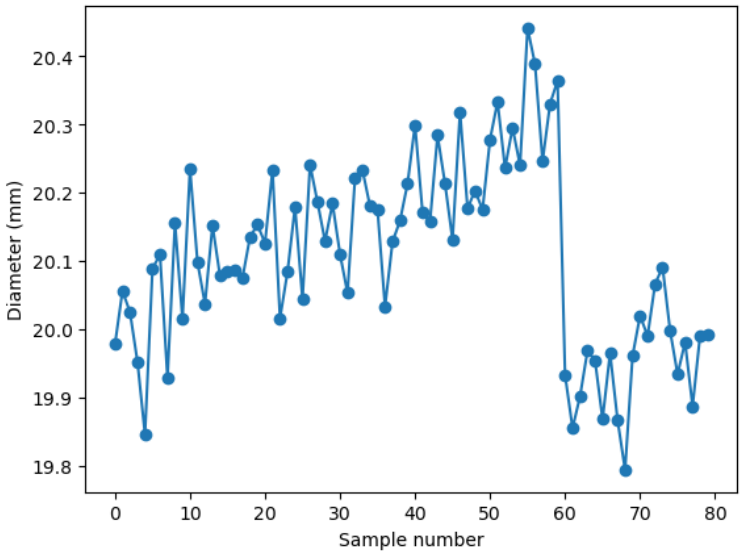
On a data set, we run the linear model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$ and we derive the ANOVA table for the fitted model. If the overall model's F-ratio has a p-value that is smaller than the predetermined level of significance (alpha) for this problem, then which of the following statements will be valid?

a) All model coefficients $(\beta_1, \beta_2, \dots, \beta_k)$ are statistically significantly different from zero.
b) All model coefficients $(\beta_1, \beta_2, \dots, \beta_k)$ are not statistically significantly different from zero.
c) At least one of the model coefficients $(\beta_1, \beta_2, \dots, \beta_k)$ is statistically significantly different from zero.
d) At most one of the model coefficients $(\beta_1, \beta_2, \dots, \beta_k)$ is statistically significantly different from zero.

Exercise 1 solution

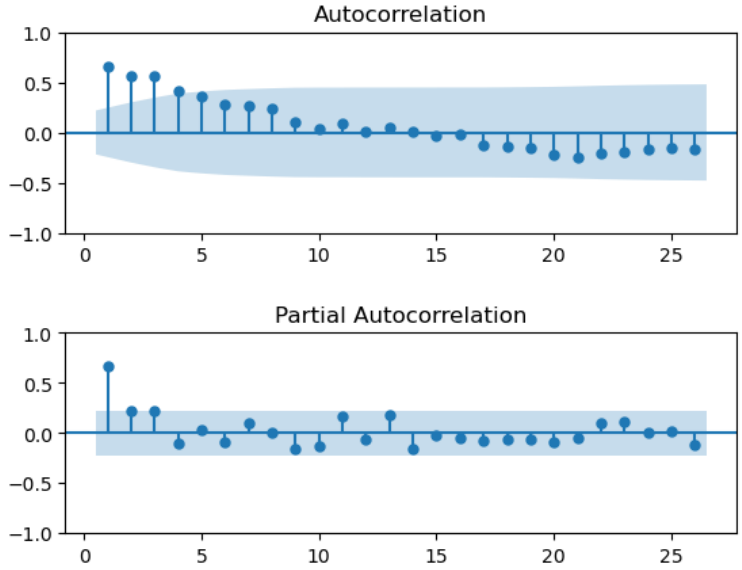
1)

Import the data and plot them.



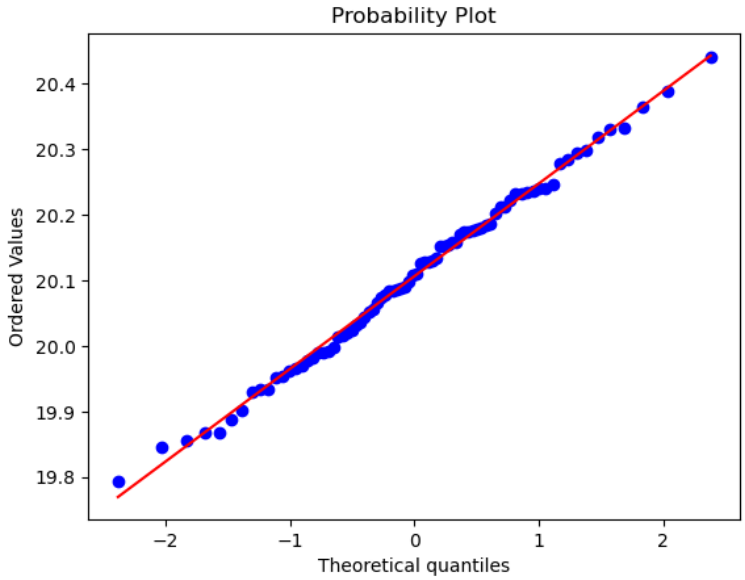
A trend and a shift in the mean seems to be present in the data. The change corresponds to the switch from shift 1 to shift 2.

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.

Let's check the normality.



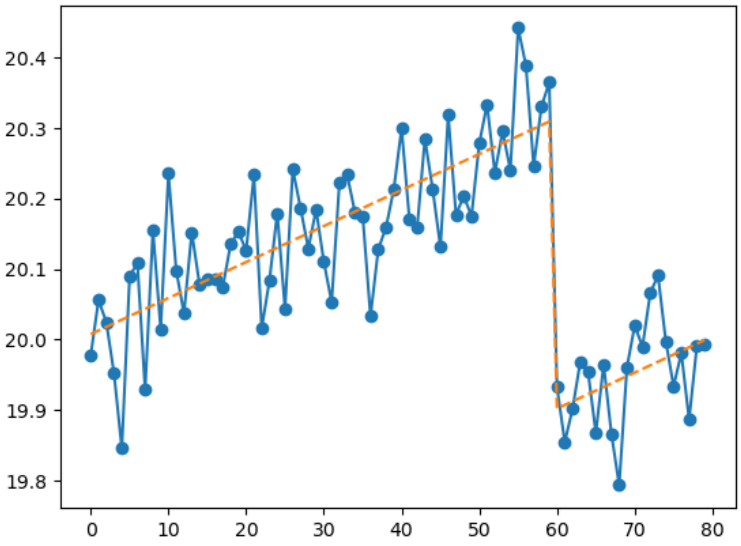
The Shapiro-Wilk test returns a p-value of 0.947, we cannot reject normality hypothesis.

Let's use the index and the shift as regressors and try creating a model:

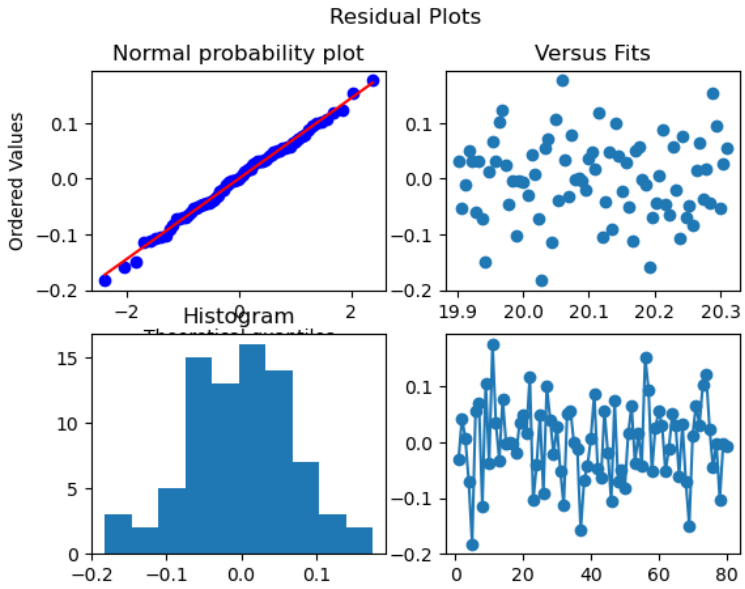
```
REGRESSION EQUATION
-------------------
diam =  + 20.414 const + 0.005 idx -0.411 shift

COEFFICIENTS
------------
 Term    Coef  SE Coef  T-Value     P-Value
const 20.4140   0.0251 811.8836 3.5828e-153
  idx  0.0051   0.0005   9.6542  6.6269e-15
shift -0.4112   0.0282 -14.6030  6.8483e-24

MODEL SUMMARY
-------------
     S   R-sq  R-sq(adj)
0.0721 0.7382     0.7314

ANALYSIS OF VARIANCE
--------------------
    Source   DF    Adj SS    Adj MS     F-Value     P-Value
Regression  2.0    1.1297    0.5648    108.5523  3.9136e-23
     const  1.0 3429.8021 3429.8021 659155.0434 3.5828e-153
       idx  1.0    0.4850    0.4850     93.2026  6.6269e-15
     shift  1.0    1.1096    1.1096    213.2472  6.8483e-24
     Error 77.0    0.4007    0.0052         NaN         NaN
     Total 79.0    1.5303       NaN         NaN         NaN
```

The runs test p-value on the residuals is 0.653 (OK) and the normality test p-value on the residuals is 0.993 (OK), so both normality and randomness on the residuals are verified.

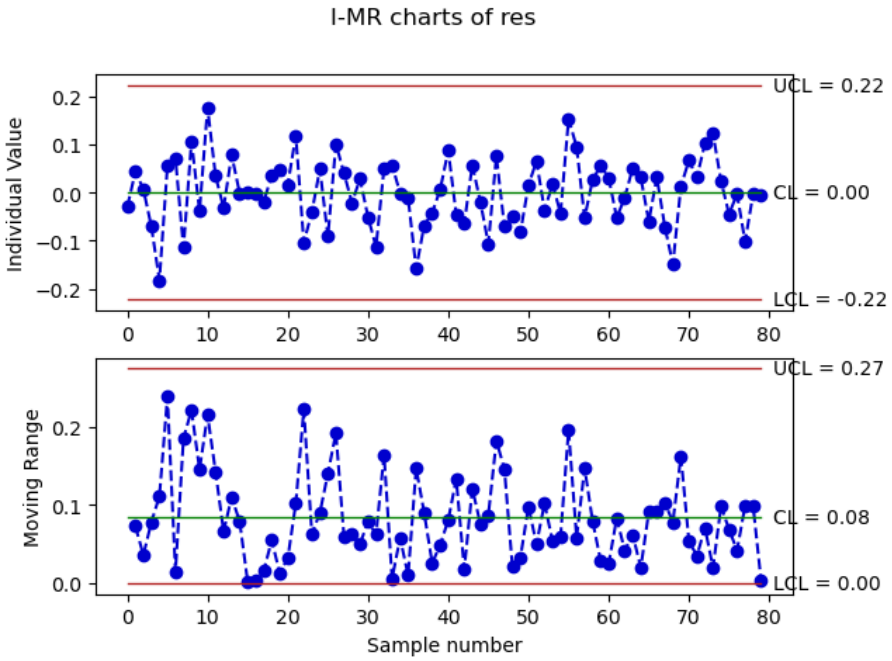
The fitted values seem to match closely the datapoints (R2_adj = 0.7314).



No particular trends are found in the residuals.



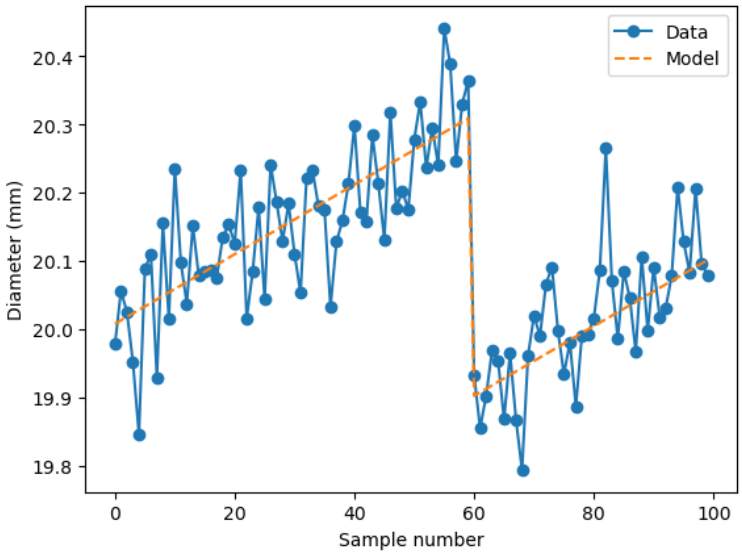
2) We can design a special cause control chart, using the residuals computed from the regression model to design an IMR control chart. K is set to 3.



No OOC in the control chart. The design phase (phase 1) is concluded.

3)

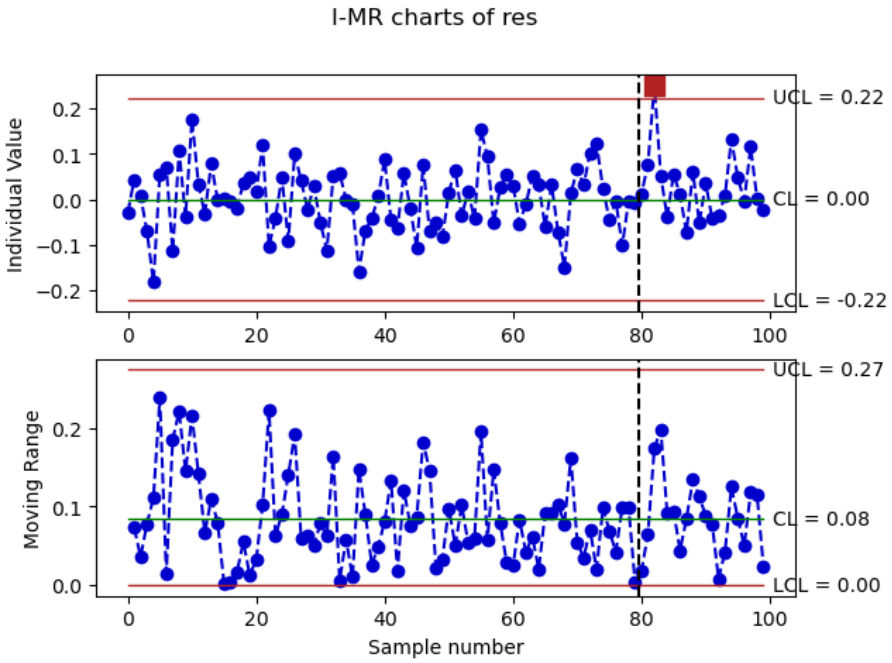
Let's use the model fitted in phase 1 to model the new data.



The model fitted in phase 1 seems to be a good fit for the phase 2 datapoints as well.

The residuals are still normal (SW test p-value = 0.730) and random (runs test p-value = 0.685).

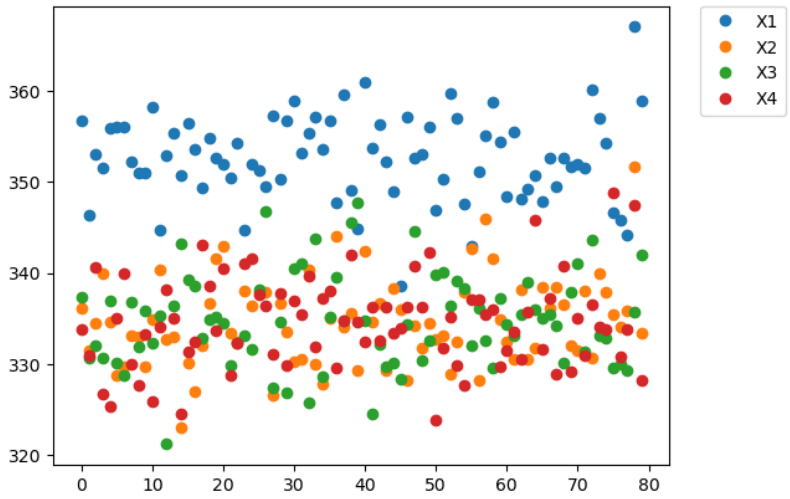
Let's apply the CC. One out-of-control is present in the phase 2 dataset (idx = 83).


I-MR charts of res
Individual Value
UCL = 0.22
CL = 0.00
LCL = -0.22
Moving Range
UCL = 0.27
CL = 0.08
LCL = 0.00
Sample number

Exercise 2 solution

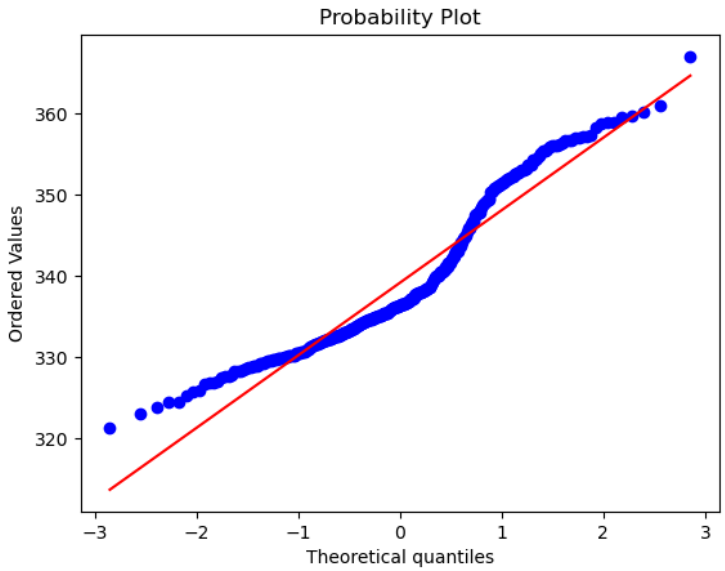
1)

Let's have a look at the data.



Hardness in location 1 is systematically higher than hardness in other locations.

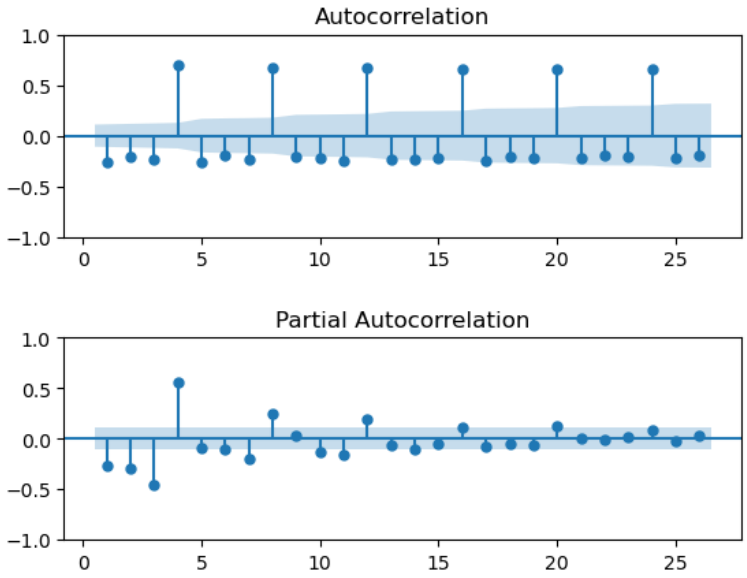
This affects the normality as shown below.



p-value of the Shapiro-Wilk test: 0.000

Since we know the time order of the data, we may also stack the data and check their autocorrelation pattern over time.

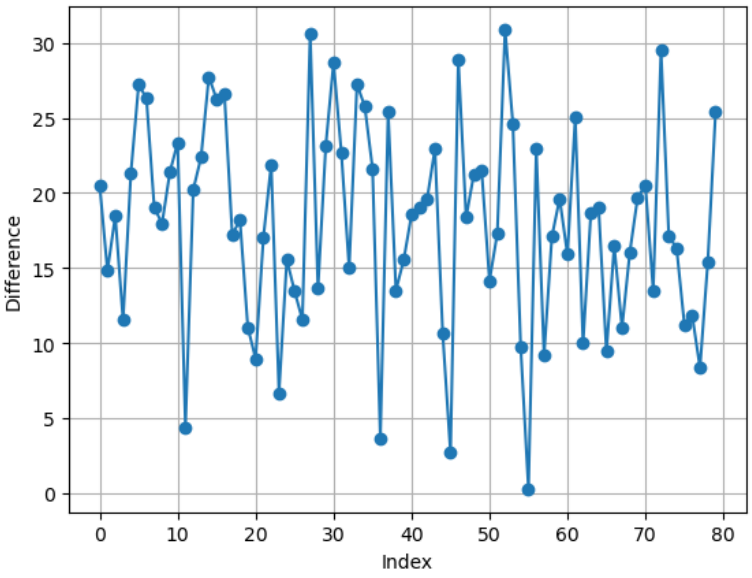
Runs test p-value = 0.000



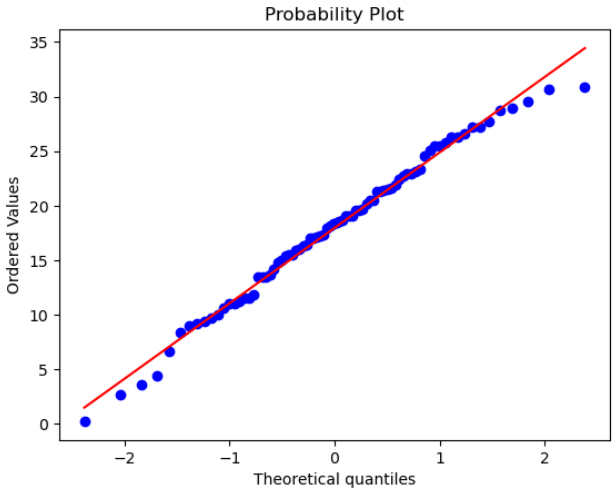
The lag 4 autocorrelation reflects the systematic pattern affecting location 1.

2)

We can perform a paired t-test.

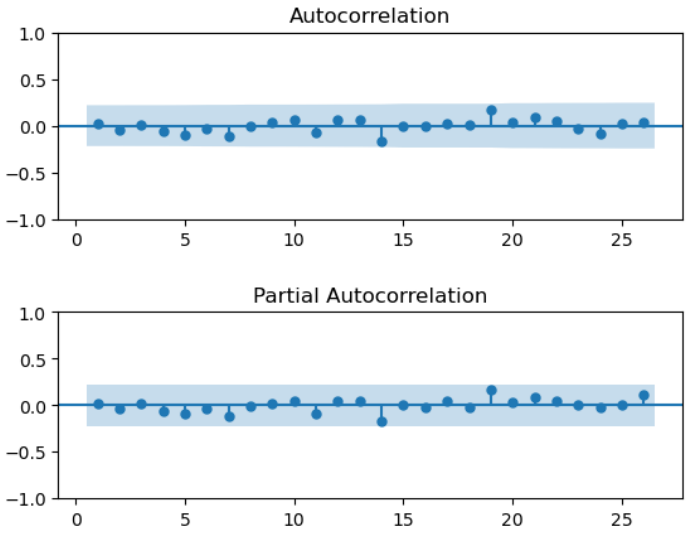


Let's check the normality of the difference with the Shapiro-Wilk test and the independence with the runs test and the ACF/PACF functions.



p-value of the Shapiro-Wilk test: 0.565

Runs test p-value = 0.982



Now that we know that the data are normally distributed, we can use the t-test to evaluate the following hypothesis:

$$H_0: \mu_d = 0$$

$$H_1: \mu_d > 0$$

The t-test statistic is:

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

Where $\bar{d}$ is the sample mean of the difference, s_d is the sample standard deviation of the difference and n is the number of observations.

We get:

t-statistic: 23.554

p-value: 0.000

There is statistical evidence that the hardness in location 1 is statistically greater than the hardness in location 2.

3)

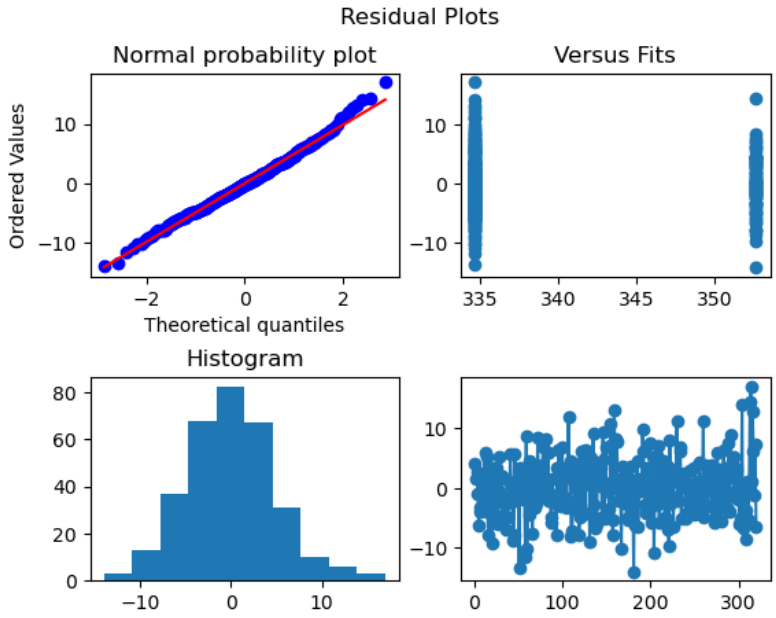
Due to the violation of the normality assumption caused by the statistically different hardness in location one, the most suitable approach would consist of fitting a model with a dummy regressor, and design a control chart for the residuals of the model. The dummy variable can be defined such that its value is 1 for location 1 and 0 for all other locations.

We get:

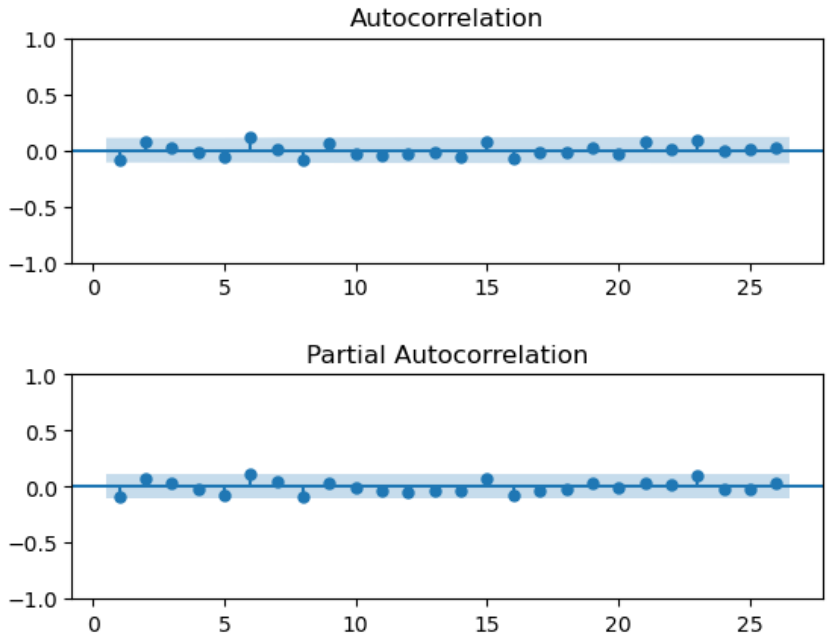
```
REGRESSION EQUATION
-------------------
data =  + 334.650 const + 17.979 loc

COEFFICIENTS
------------
 Term     Coef  SE Coef   T-Value    P-Value
const 334.6495   0.3183 1051.4040 0.0000e+00
  loc  17.9787   0.6366   28.2428 1.1257e-88

MODEL SUMMARY
-------------
     S  R-sq  R-sq(adj)
4.9309 0.715     0.7141

ANALYSIS OF VARIANCE
--------------------
    Source    DF     Adj SS     Adj MS    F-Value    P-Value
Regression   1.0 1.9394e+04 1.9394e+04 7.9765e+02 1.1257e-88
     const   1.0 2.6878e+07 2.6878e+07 1.1055e+06 0.0000e+00
       loc   1.0 1.9394e+04 1.9394e+04 7.9765e+02 1.1257e-88
     Error 318.0 7.7318e+03 2.4314e+01        NaN        NaN
     Total 319.0 2.7126e+04        NaN        NaN        NaN
```

Let’s check assumptions by analysing model residuals:

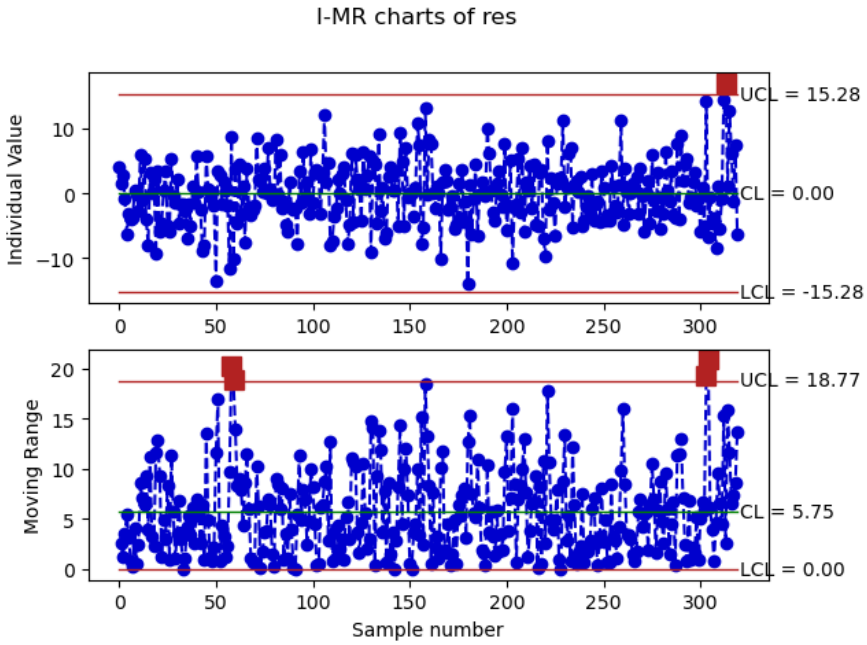
Shapiro-Wilk test p-value on the residuals = 0.198



Runs test p-value on the residuals = 0.140



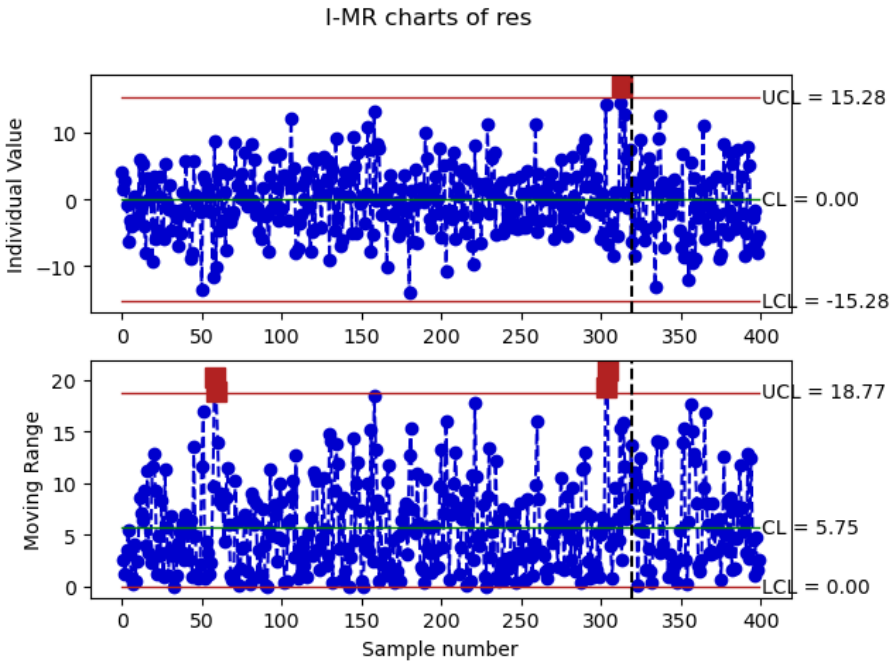
Let’s design the I-MR control chart on the residuals.



There is one violation of control limits in the I chart and two violations in the MR chart. Assuming no assignable cause for them, the control chart design is over.

4)

Phase 2 control chart. Let's fit the same model to the phase 2 data and apply the previously designed I-MR chart on model residuals of new data.



The new data are in control.

Exercise 3) Solution

Question 1)

Answer: b
Explanation: The area outside of the control limits is the type I error, i.e. describes the false alarm rate, which for the choice of K will correspond to the value α. Increasing K to K_1 will decrease the type I error to $\alpha_1 < \alpha$ and as a result the false alarm rate will decrease, so (b) is valid and therefore (a) and (d) will be invalid.
We also know that as type I error, α, decreases, then the type II error, β, will increase and given that power $= 1 - \beta$, we will have that power will decrease, so (c) is not valid as well.

Question 2)

Answer: c

Explanation: The ANOVA table for the linear model performs the following hypothesis testing: $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ versus $H_1: not\ H_0$.

As the p-value < alpha we can reject the null hypothesis H_0 and thus (b) is invalid. The alternative hypothesis H_1 is the complement of H_0, i.e. at least one of the model coefficients is not equal to zero, thus (c) is valid. Both (a) and (d) are invalid as none of them is expressing the alternative hypothesis.

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?

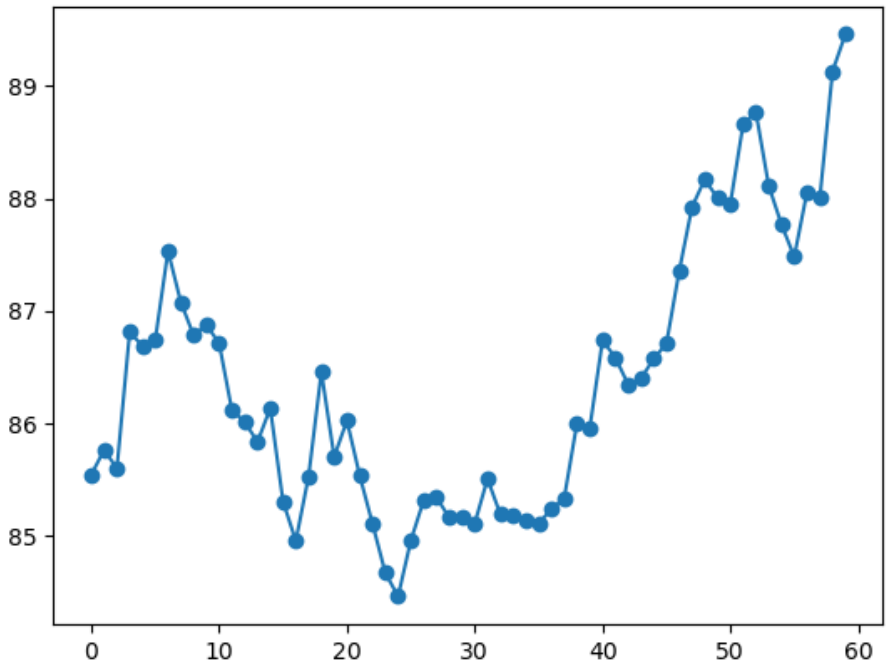
a. The type I error will decrease. T
b. The power will decrease. T
c. The in-control Average Run Length (ARL_0) will increase. T
d. The out-of-control Average Run Length (ARL_1) will decrease. F

il beta aumenta! perciò ARL1 che è il numero di campioni per detectare OOC dopo lo shift aumenta

Exercise 1 solution

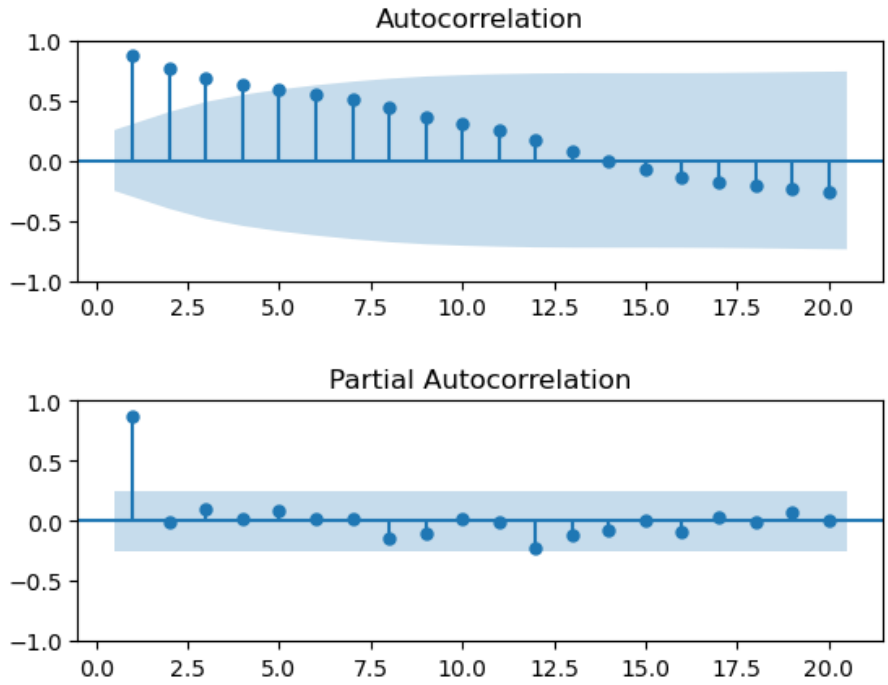
1)

Import the data and plot them.



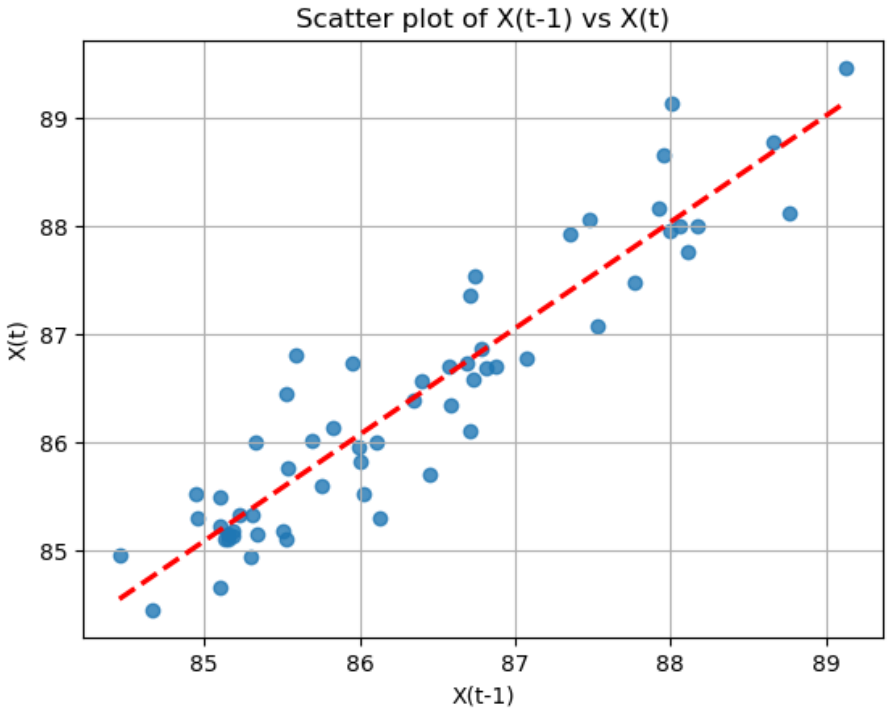
A meandering patters seems to be present in the data and the process does not seem to be stationary.

Let's check the randomness.

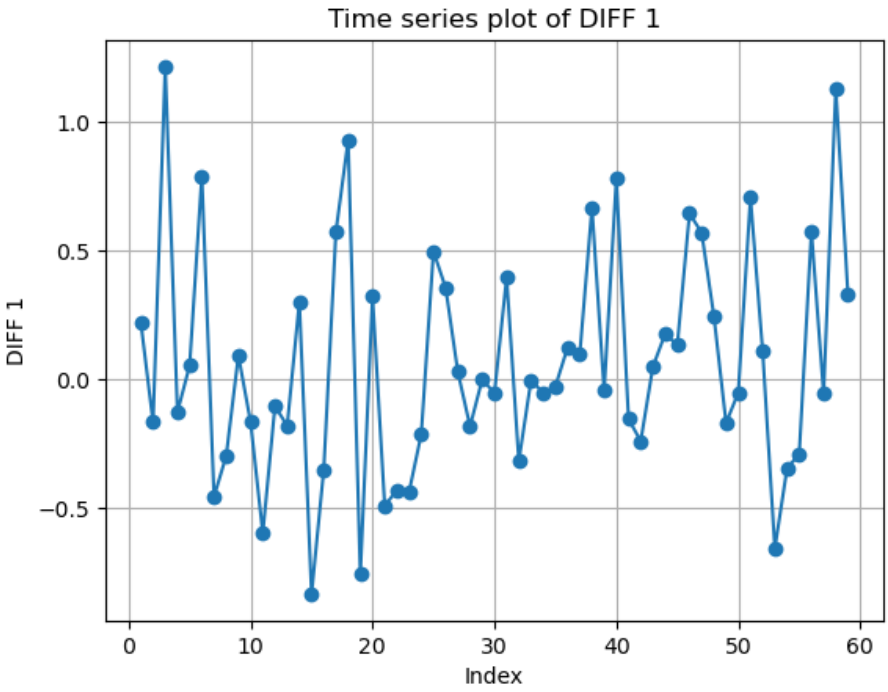


The runs test returns a p-value of 0.000, and the sample ACF shows a linearly decaying trend, which is typical of non-stationary time series. Based on ACF analysis, we can state that the process is non-stationary

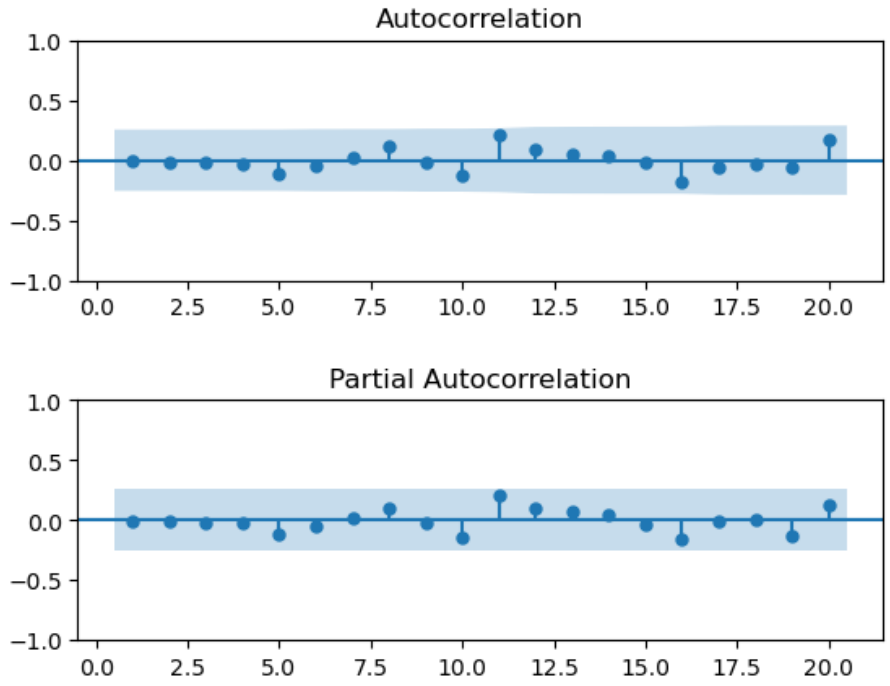
Moreover, we can observe with a scatterplot the correlation between X(t) and X(t-1).



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.

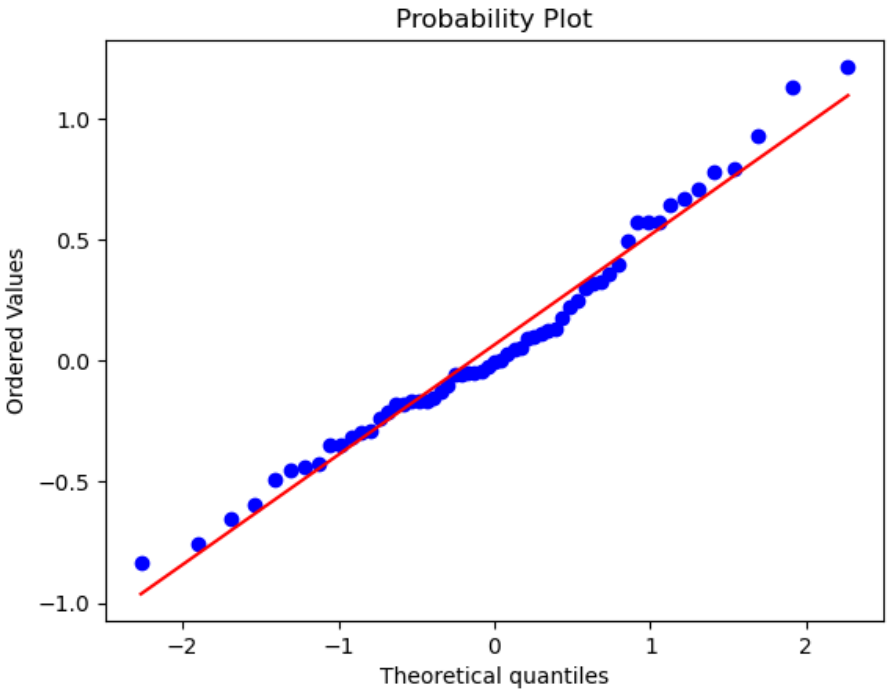


Now everything seems stationary. Let's check randomness.



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:

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

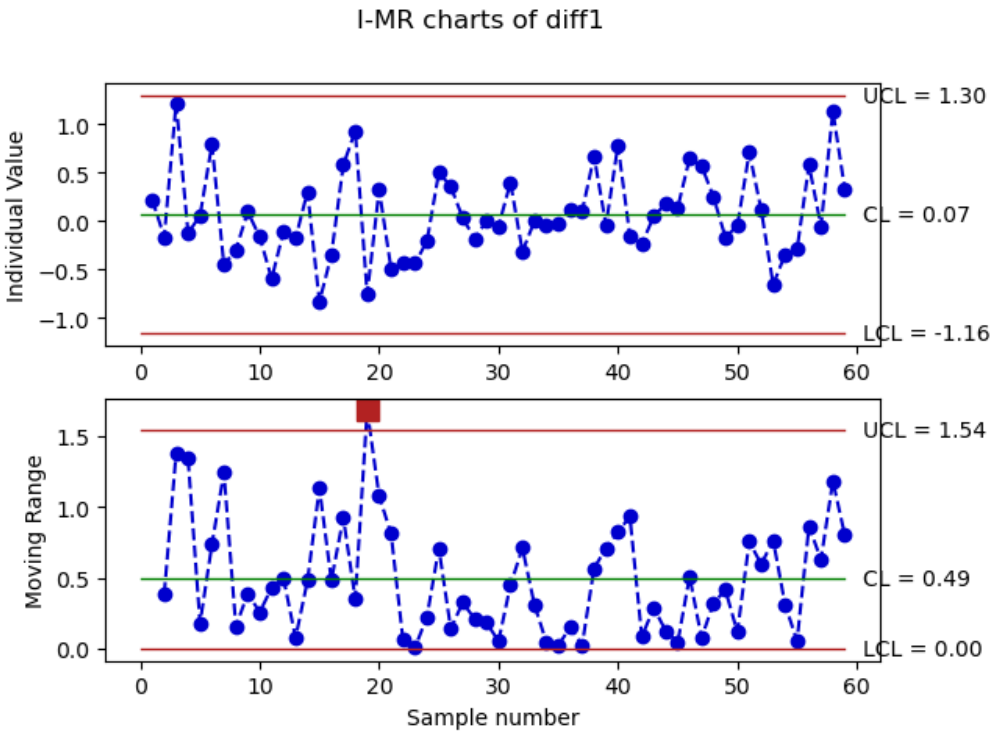
2)

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

The ARL0 is set to 200.

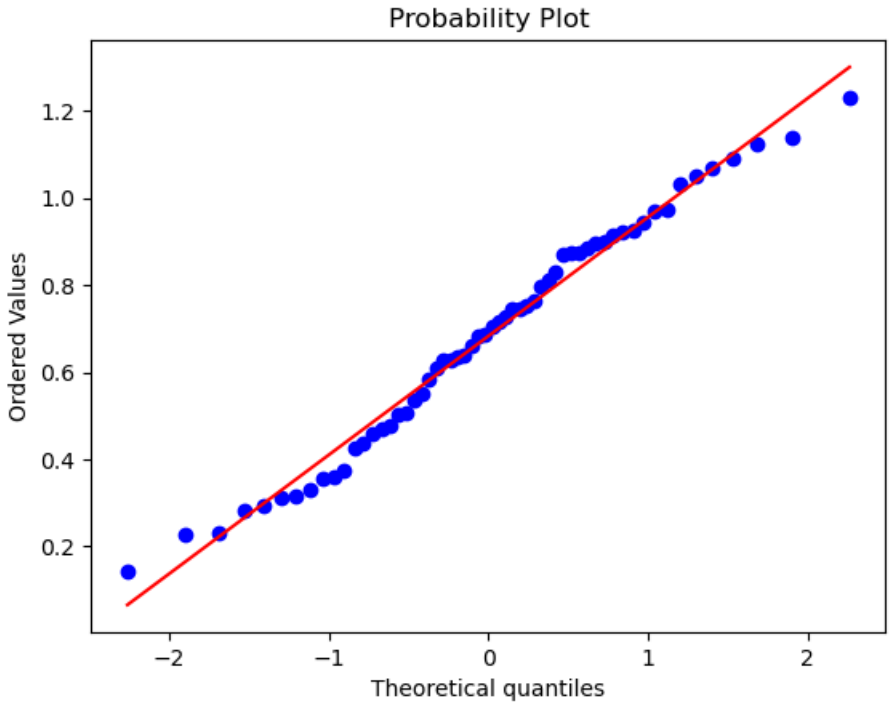
alpha = 1/ARL0 = 0.005

K = 2.807

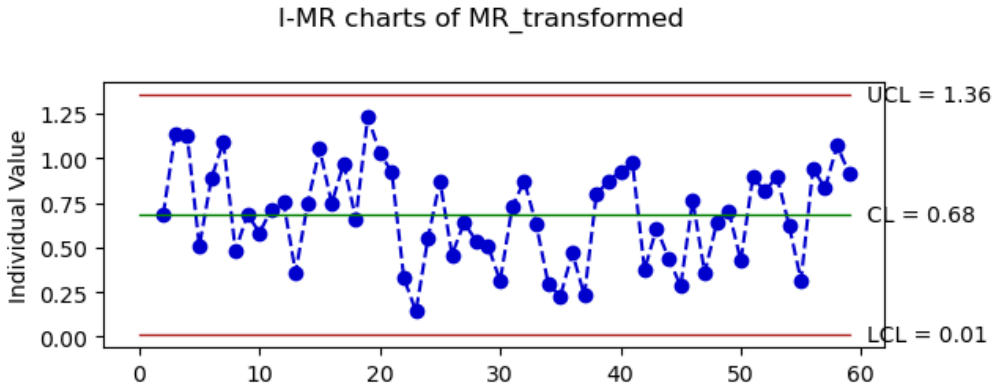


One moving range value seems to be OOC.

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).



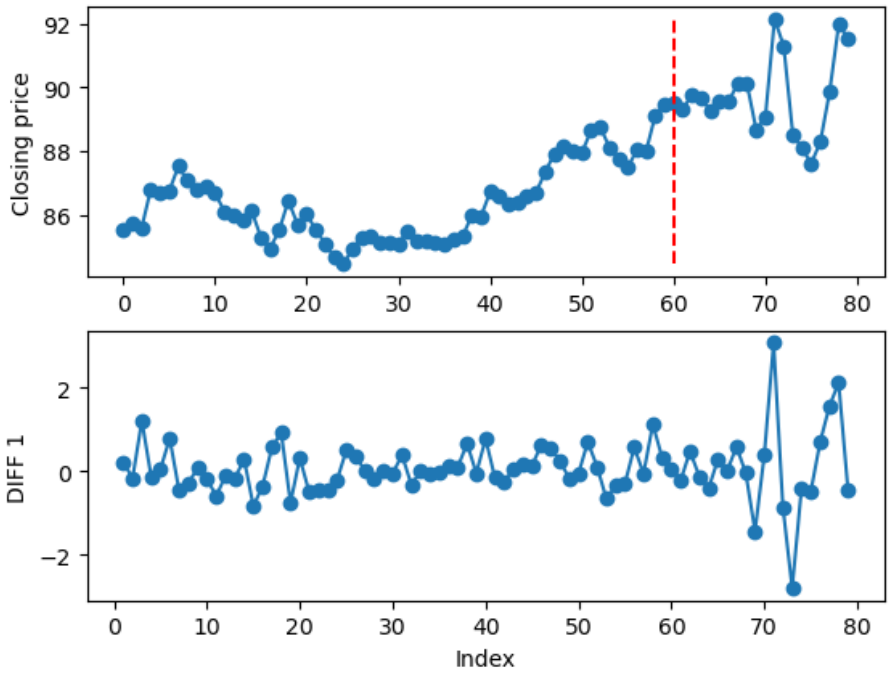
The individual control chart on the transformed variable can be designed.



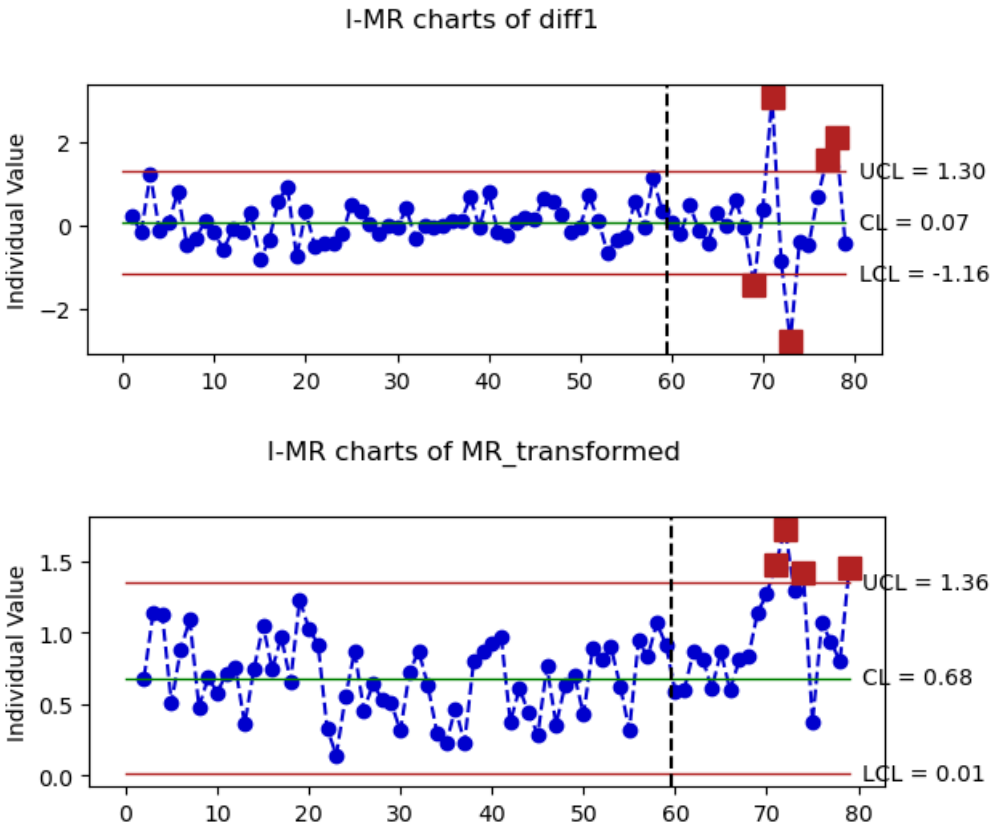
No points are out of control, the design phase is complete.

3)

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

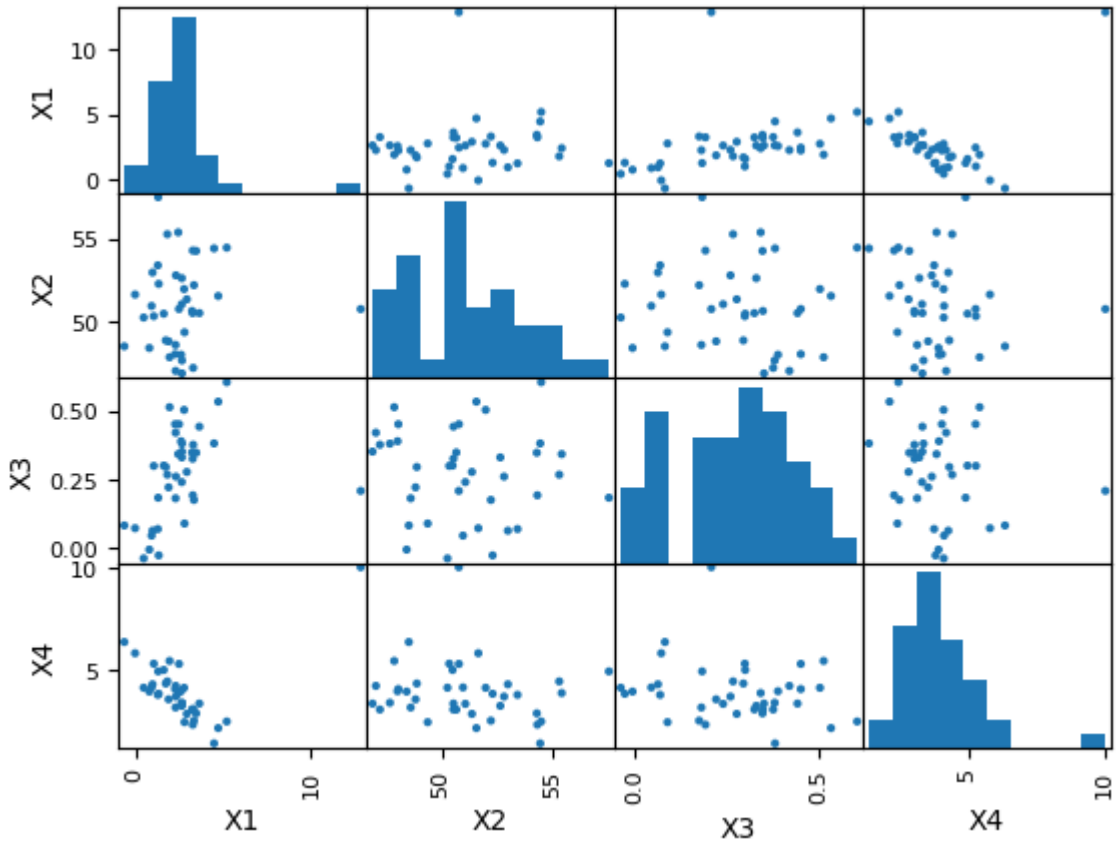


The new data seem to show a higher volatility, with larger fluctuations compared to the phase 1 data. The control chart confirms this and highlights a large number of OOC points in the I-chart (69, 71, 73, 77, 78) and in the MR chart after transformation (71, 72, 74, 79).



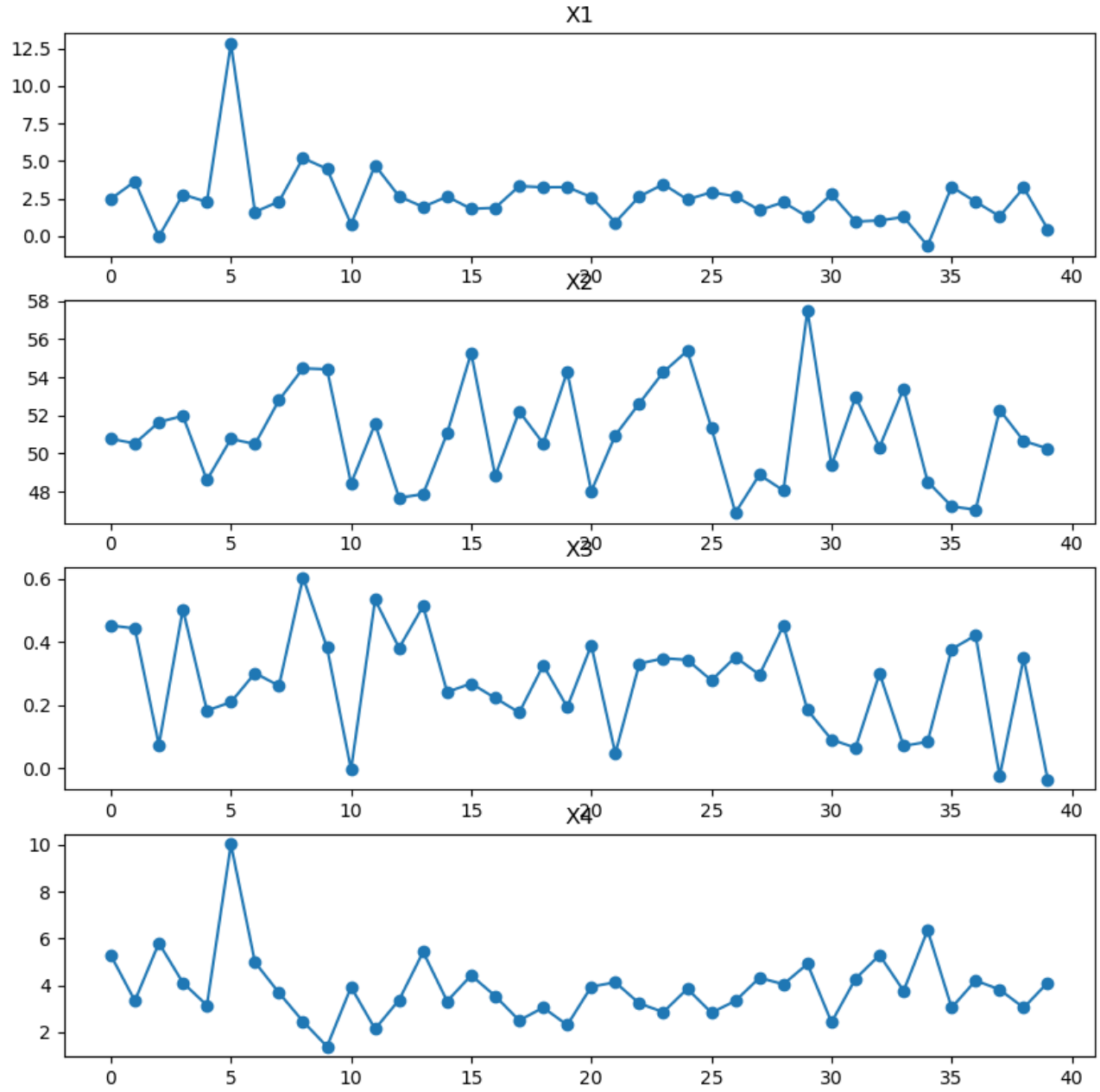
Exercise 2 solution

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]

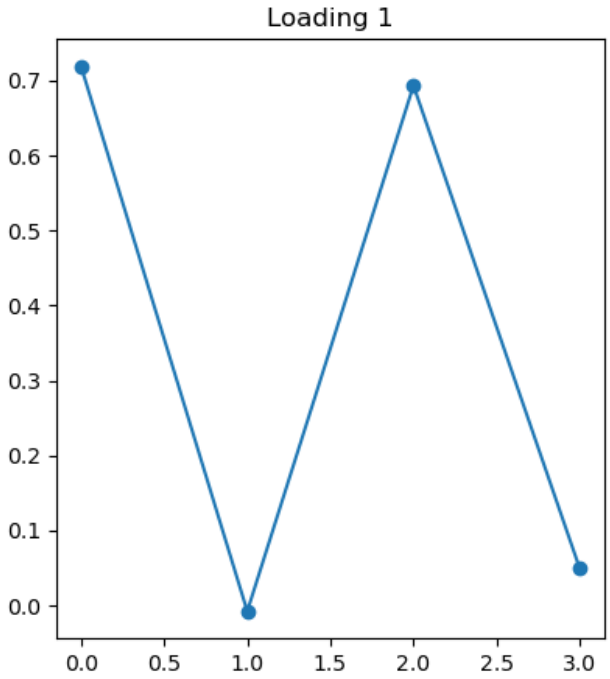
Cumulative explained variance ratio

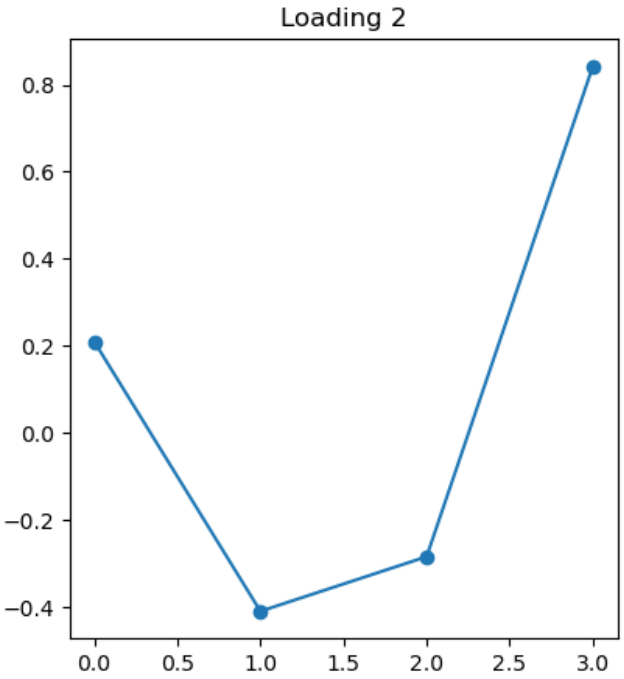
[0.34058614 0.63649607 0.89208659 1.]

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:

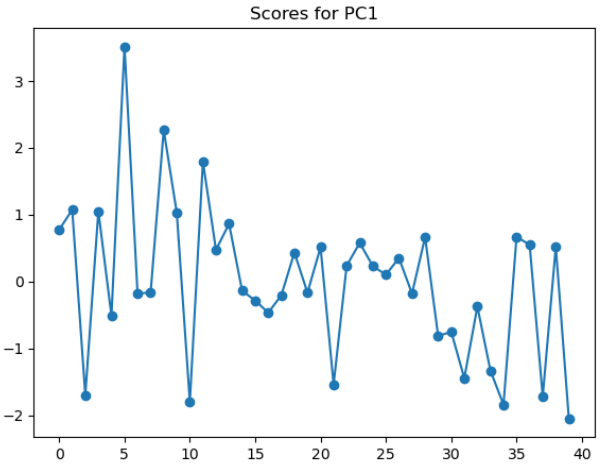


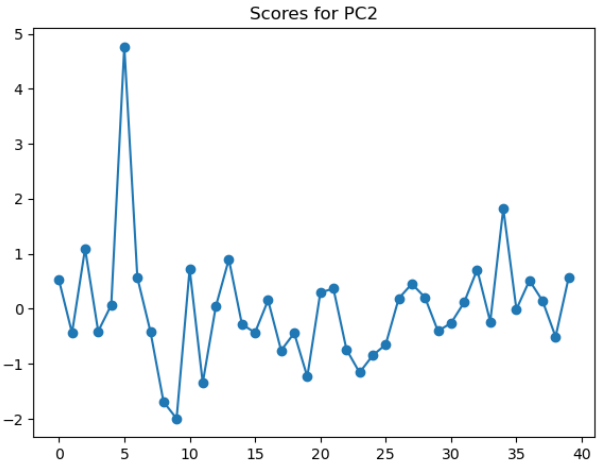


The first PC is mainly influenced by variables 1 and 3 (large positive weights), while variables 2 and 4 have a weight close to 0. The second PC is a contrast between variables 1 and 4 (positive weights) and variables 2 and 3 (negative weights), and variable 4 is the one with the largest weight.

3)

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

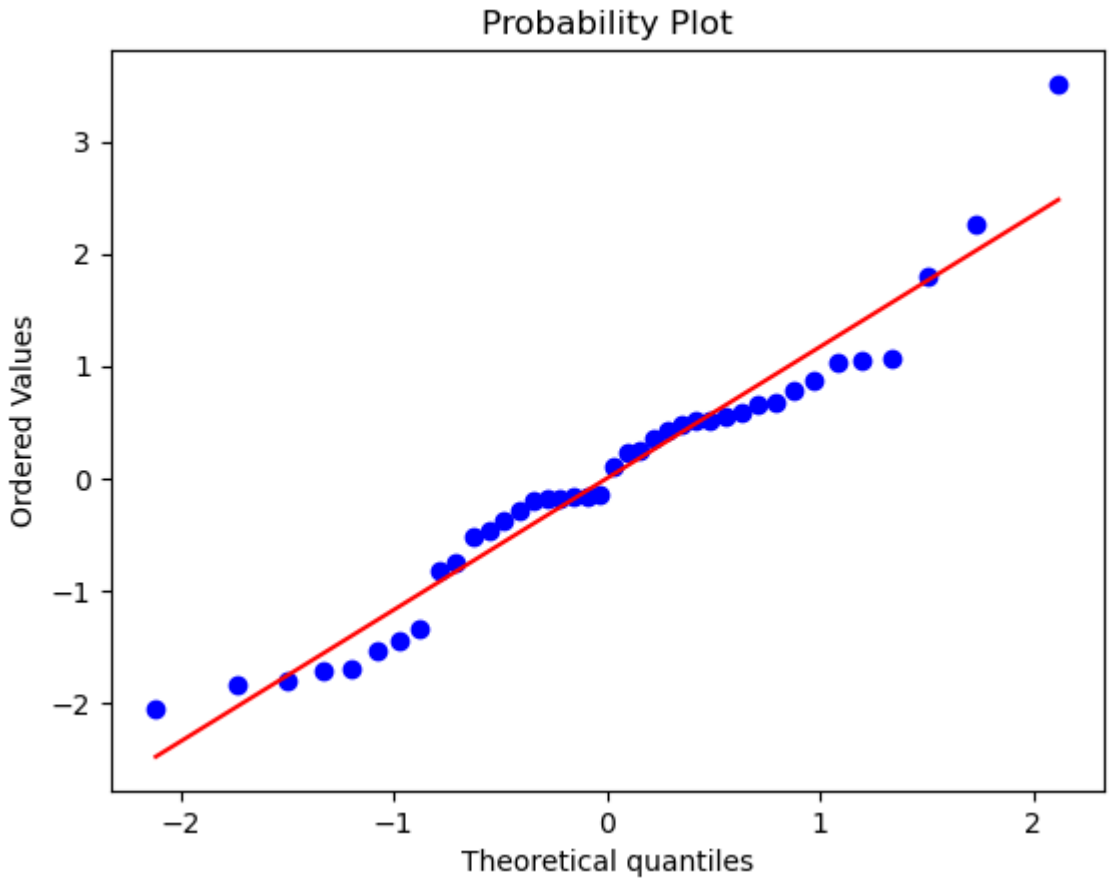




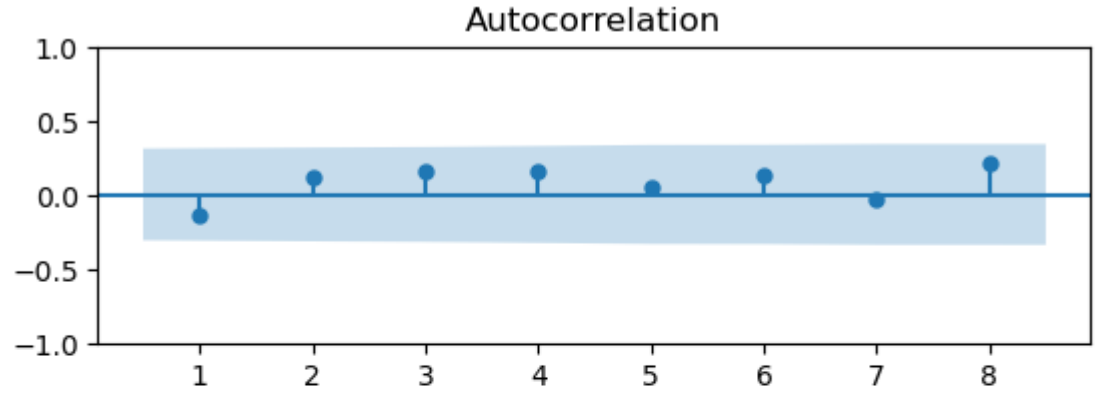
An outlier is possibly present along the second principal component. It is the same datapoint that was observed in the scatterplots of variables 1 and 4. This outlier mainly influences the second PC due to the linear combinations discussed above. Let's test for normality and independence.

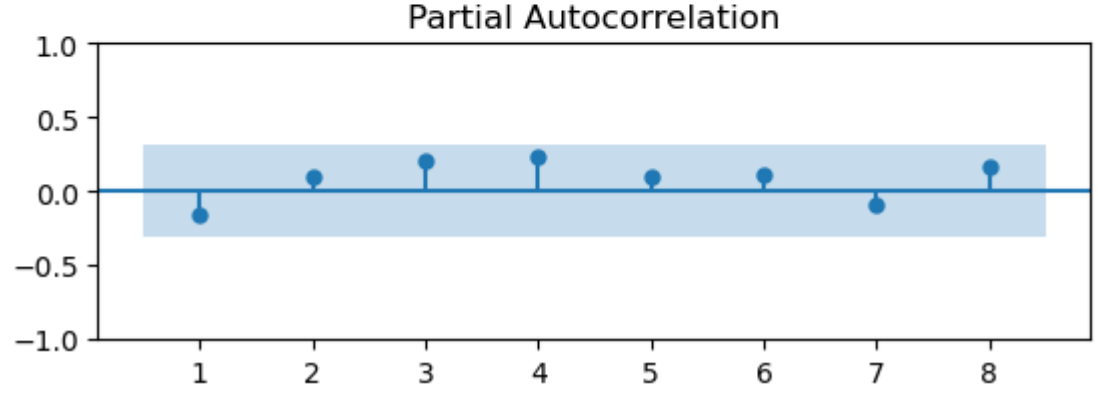
Tests for PC1.

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




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

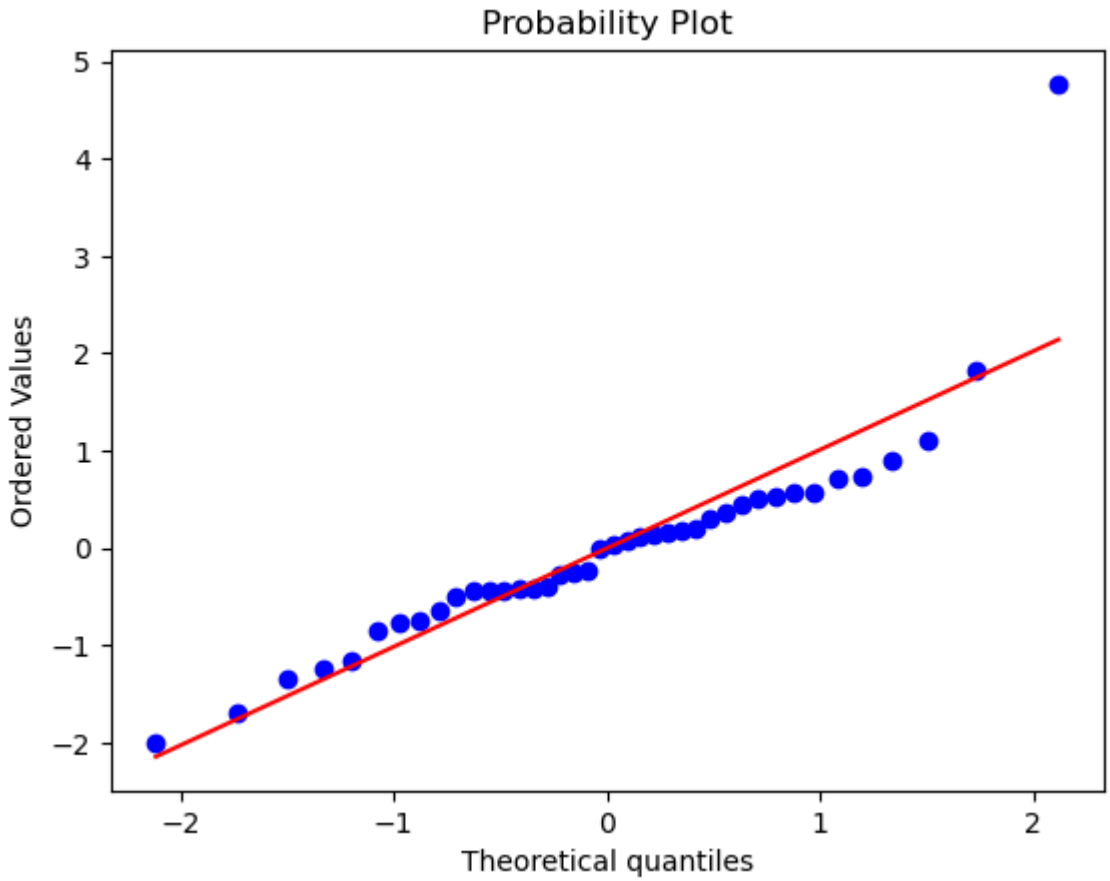




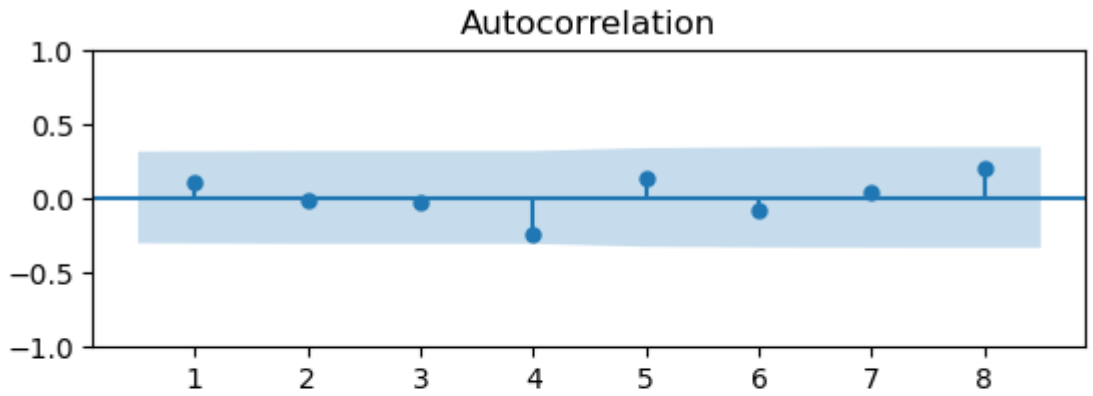
PC1 is normal and independent.

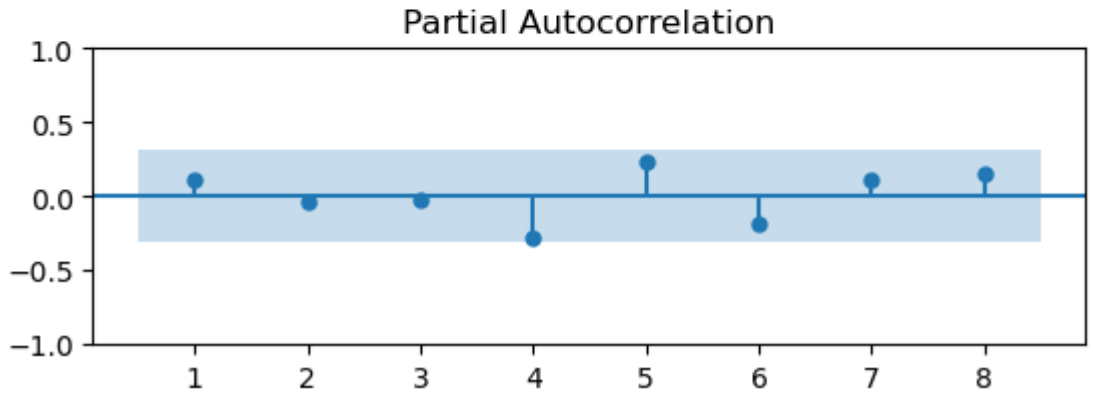
Tests for PC2:

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



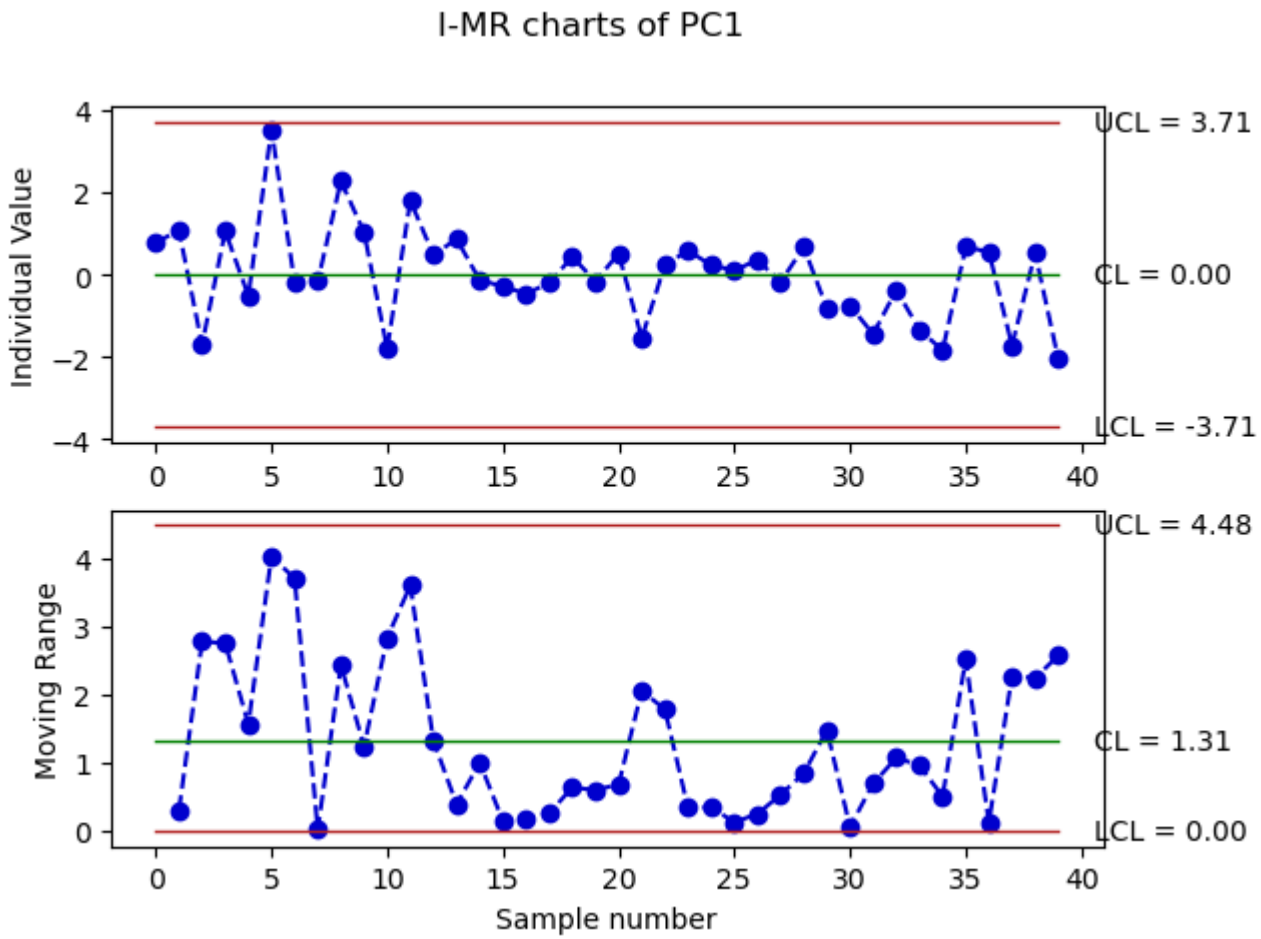

```
Runs test p-value = 0.631
```

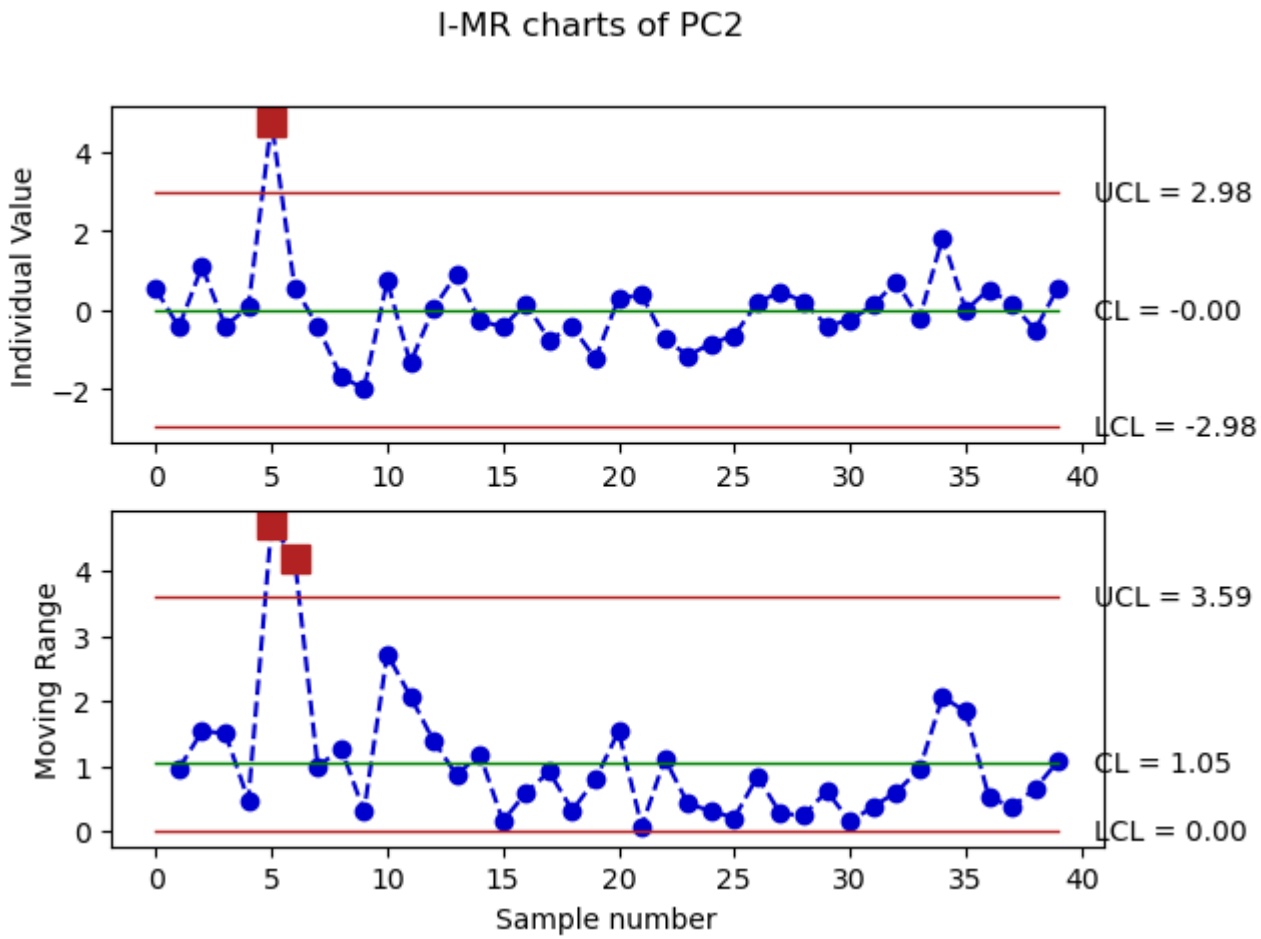




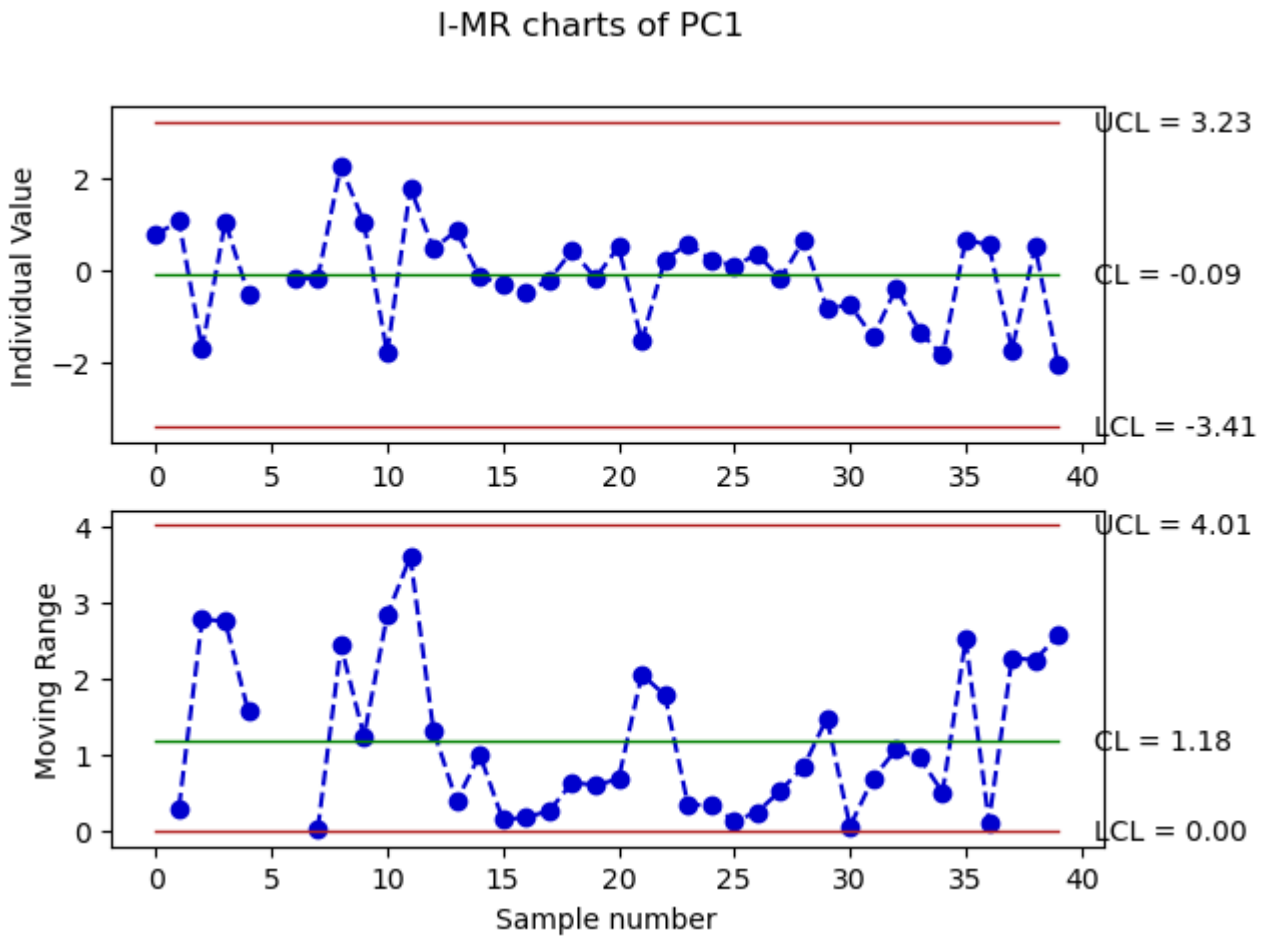
The second PC violates the normality assumption, but such violation is caused by the presence of the outlier. Indeed, it can be easily verified that removing the outlier, normality is met. Since this is the only violation of assumptions, we may design the I-MR control charts for the two PCs and verify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

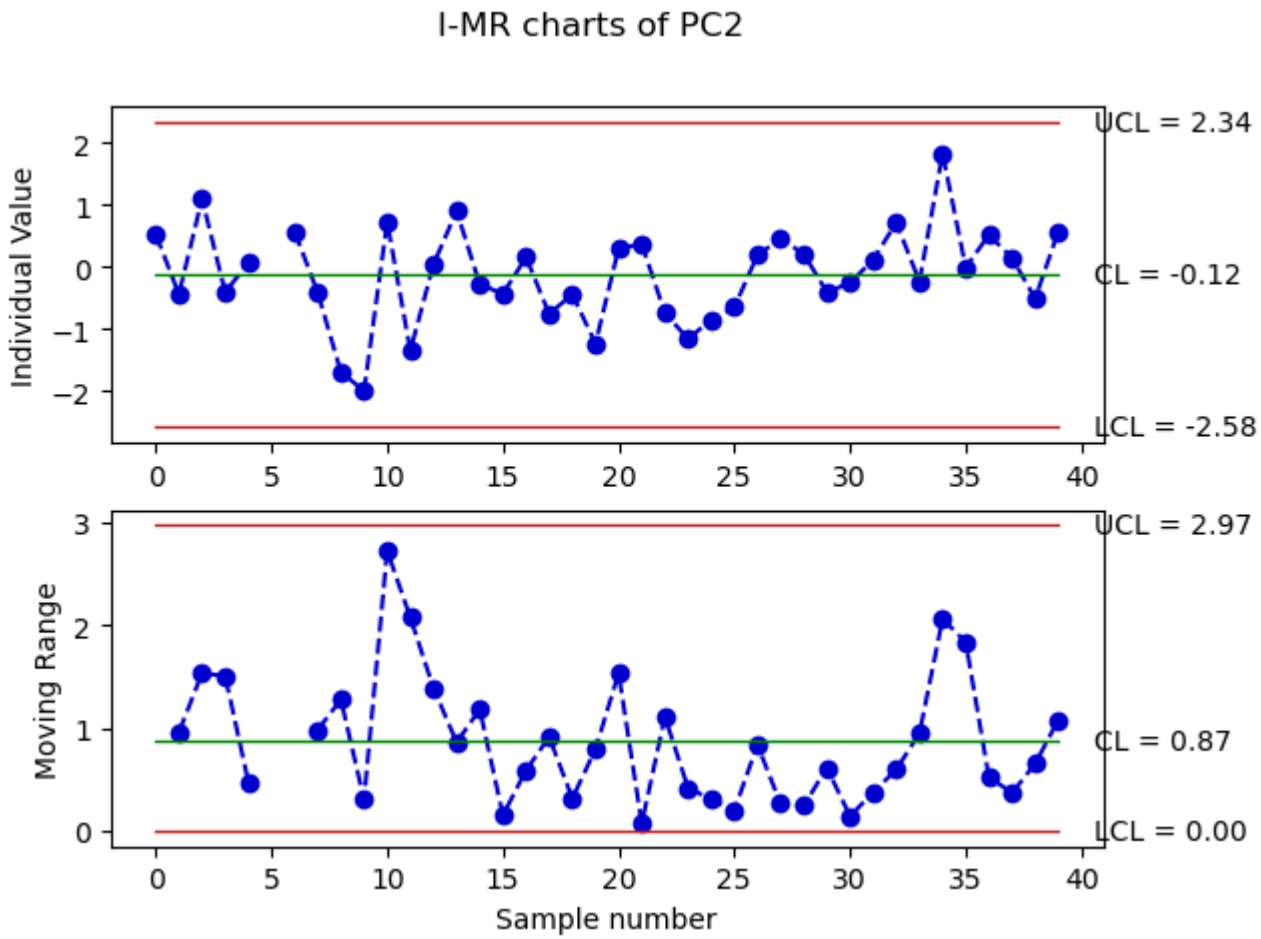
Design the I-MR control charts. Since we have two independent control variables we shall use the family-wise correction for independent data, which leads to K = 3.189.


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.

QUALITY DATA ANALYSIS

19/07/2023

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 2, point d) and Exercise 3, Question 2.**

Exercise 1 (14 points)

The oxygen concentration inside the build chamber of a metal 3D printer is measured at the end of each layer. The data of the first 100 layers is reported in the file `*oxygen_phase1.csv*`.

a) Find a suitable model for the data.
b) Predict the oxygen concentration at the end of the next layer.
c) Design a suitable control chart to monitor the oxygen concentration with a Type I error $\alpha = 0.0029$. Please show on the solution the formula of the control limits and their numerical values. *Note: in case of violations of the control limits, assume the existence of assignable causes.*
d) Using the control chart designed in point 3 (phase 1), test it on the data of the next 10 (stored in `*oxygen_phase2.csv*`). Please report all the values plotted in the control chart in Phase 2 against limits. Report then the index of the out-of-control observations (if any).

Exercise 2 (15 points)

In a process for the production of an energy drink, three quality variables are monitored, denoted by x_1, x_2 and x_3. Under in-control conditions they are known to be iid and to follow a multivariate normal distribution with $\mu_1 = 25.4,\ \mu_2 = 23.7,\ \mu_3 = 28.5,$ and $\sigma_1 = 1.483,\ \sigma_2 = 1.435,\ \sigma_3 = 1.530$. It is also known that the correlation between each pair of quality variables is $\rho_{12} = 0.799$, $\rho_{13} = 0.705$, and $\rho_{23} = 0.683$.

a) Design three univariate I control charts for the three distinct quality characteristics such that the familywise type I error is at most $\alpha = 0.01$. In the use phase of the control chart, the dataset included in *exe_pca.csv* is collected. Using the designed chart, determine if these new measurements are in control or not.
b) Estimate and draw the operating characteristic curve of the control chart for x_1 designed in point a) in the presence of a shift $\Delta\mu_1$ of the mean of x_1 with $\Delta\mu_1 \in [0, 10]$, and report the value of the Type II error for $\Delta\mu_1 = 3$.
c) The head of the quality department decides to use the Principal Component Analysis to monitor these data. The first principal component (PC) allows capturing at least 65% of the data variability and it is such that:

$$\boldsymbol{u}_1 = [-0.589, -0.561, -0.582]^T,\ \lambda_1 = 5.402$$

Design a univariate control chart to monitor the first PC with a type I error $\alpha = 0.01$. Using this control chart, determine if the new observations in *exe_pca.csv* are in control or not.

d) Estimate and draw the operating characteristic curve of the control chart designed in point c) in the presence of a shift $\Delta\mu_1$ of the mean of variable x_1 with $\Delta\mu_1 \in [0, 10]$, and report the value of the Type II error for $\Delta\mu_1 = 3$. Compare this result with the one in point b) and discuss the difference.

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 the stepwise regression we need to define two levels of significance: Alpha-to-Enter and Alpha-to-Remove. Which of the following choices is **not** a valid selection?

a) Alpha-to-Enter ≤ Alpha-to-Remove.
b) Alpha-to-Enter = Alpha-to-Remove.
c) Alpha-to-Enter < Alpha-to-Remove.
d) Alpha-to-Enter > Alpha-to-Remove.

Question 2 (2 points):
In a fitting the linear regression model of the continuous variable **Y** on the five predictors {$\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$, $\mathbf{X_4}$ and $\mathbf{X_5}$} we obtained the following p-values:

Predictor	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$
p-value in the full model	0.015	0.241	0.001	0.087	0.047

If we will apply on the above model the backward elimination procedure using the value of Alpha-to-Remove = 0.05, which is the first predictor that will be removed?

a) X_1
b) X_2
c) X_4
d) We cannot tell from the above output only.

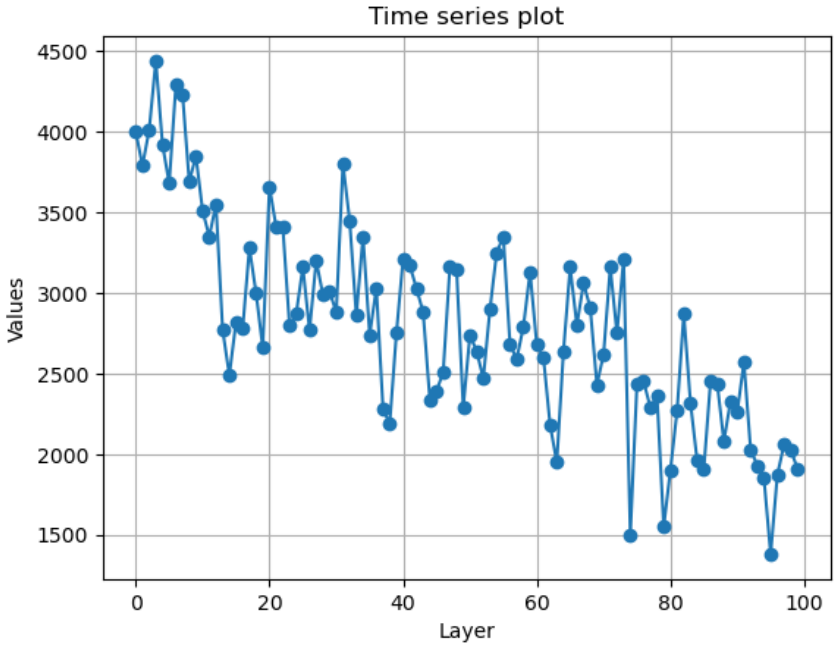
What is the predictor to be possibly eliminated on a second step of the backward elimination algorithm?

a) X_1
b) X_2
c) X_4
d) We cannot tell from the above output only.

Exercise 1 solution

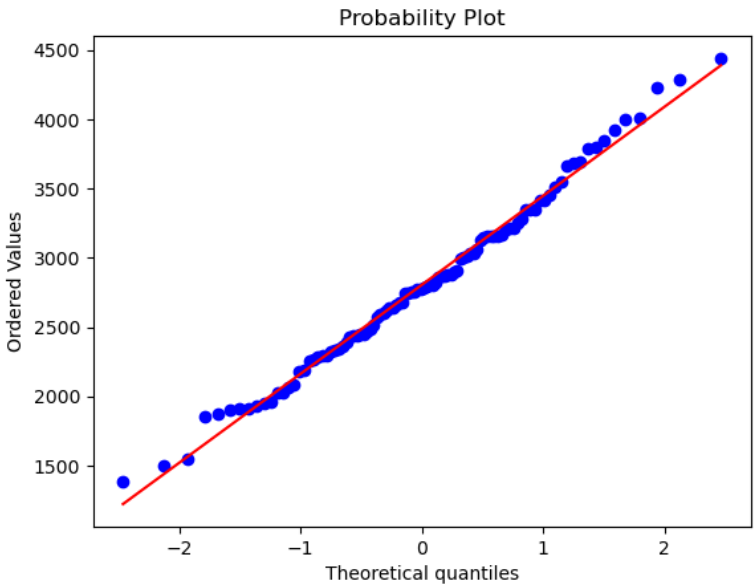
a)

Since the time order of the measurements is known, we can plot the time series as follows:



The pattern is clearly not random as it appears that the temperature is decreasing as the number of layers increases.

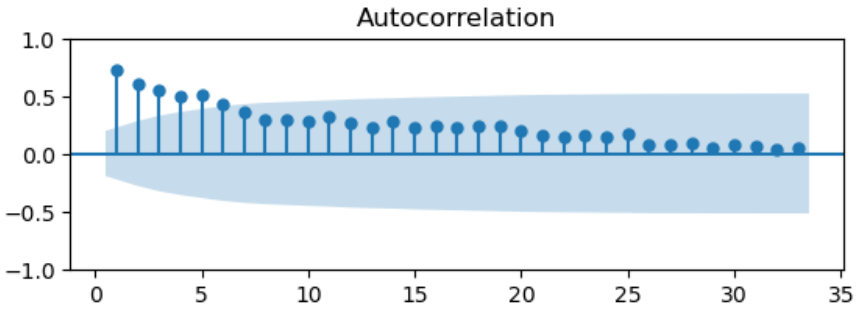
The normality assumption is met (Shapiro-Wilk's test p-value = 0.663).

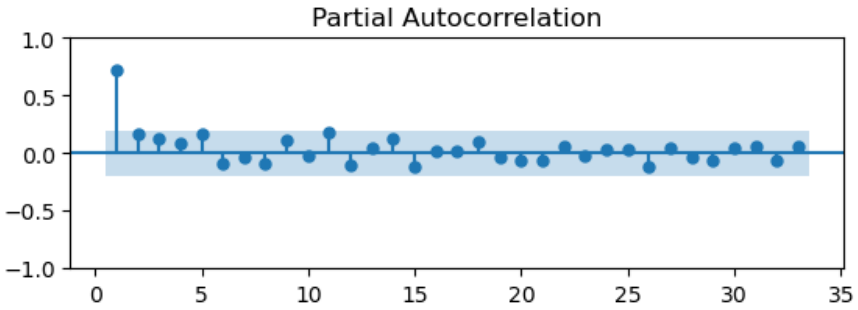


Since we know the time order of the data we can check the randomness and autocorrelation.

The randomness assumption is NOT met (Runs test p-value = 0.000).

The sample ACF shows the typical pattern of a non-stationary time series (linear decay).

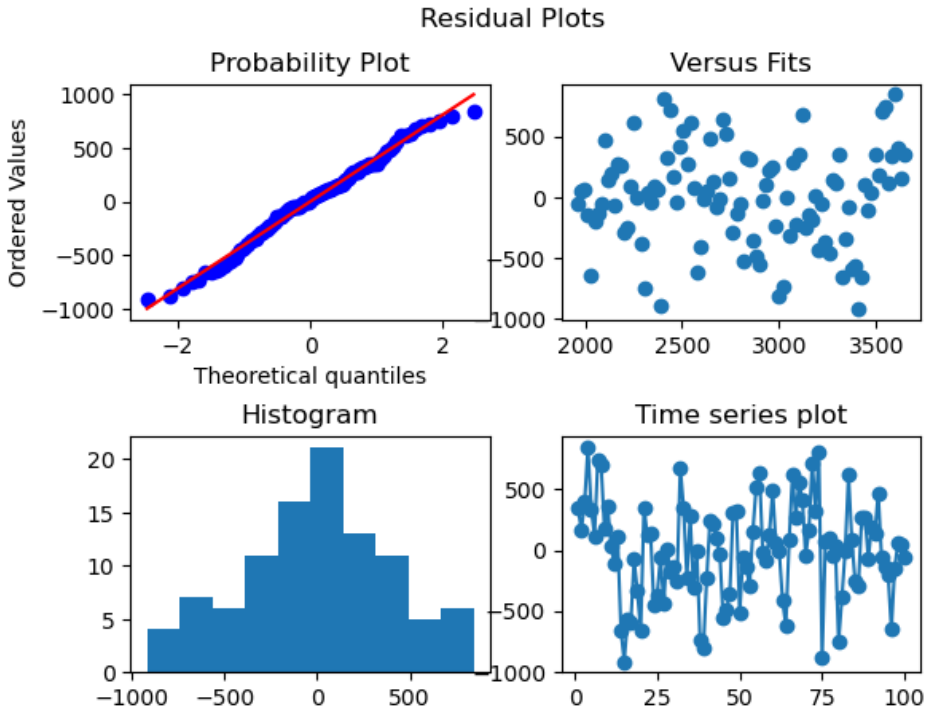




We can try to fit a trend model using the layer number as regressor.

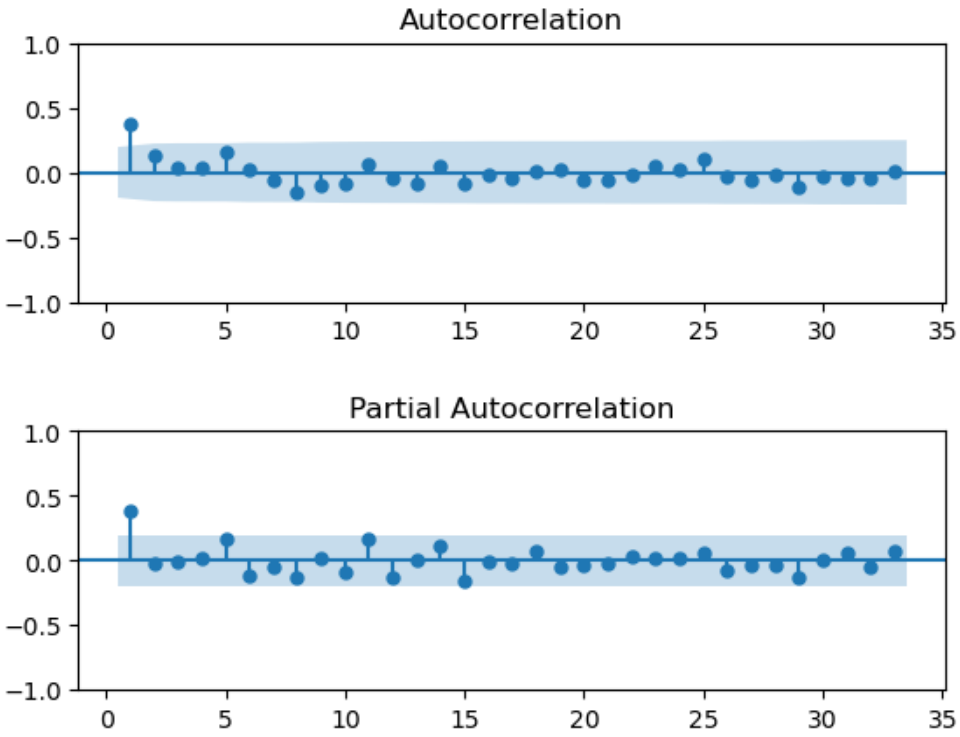
```
REGRESSION EQUATION
-------------------
val =  + 3663.885 const -16.998 layer

COEFFICIENTS
------------
 Term      Coef  SE Coef  T-Value    P-Value
const 3663.8848  81.2907  45.0714 2.5144e-67
layer  -16.9977   1.3975 -12.1628 2.7090e-21

MODEL SUMMARY
-------------
       S   R-sq  R-sq(adj)
403.4091 0.6015     0.5975

ANALYSIS OF VARIANCE
--------------------
    Source   DF     Adj SS     Adj MS   F-Value    P-Value
Regression  1.0 2.4074e+07 2.4074e+07  147.9331 2.7090e-21
     const  1.0 3.3059e+08 3.3059e+08 2031.4282 2.5144e-67
     layer  1.0 2.4074e+07 2.4074e+07  147.9331 2.7090e-21
     Error 98.0 1.5948e+07 1.6274e+05       NaN        NaN
     Total 99.0 4.0023e+07        NaN       NaN        NaN
```

Let’s check the assumptions on the residuals.



The normality assumption is met (Shapiro-Wilk's test p-value = 0.449).

The randomness assumption is NOT met (Runs test p-value = 0.003).

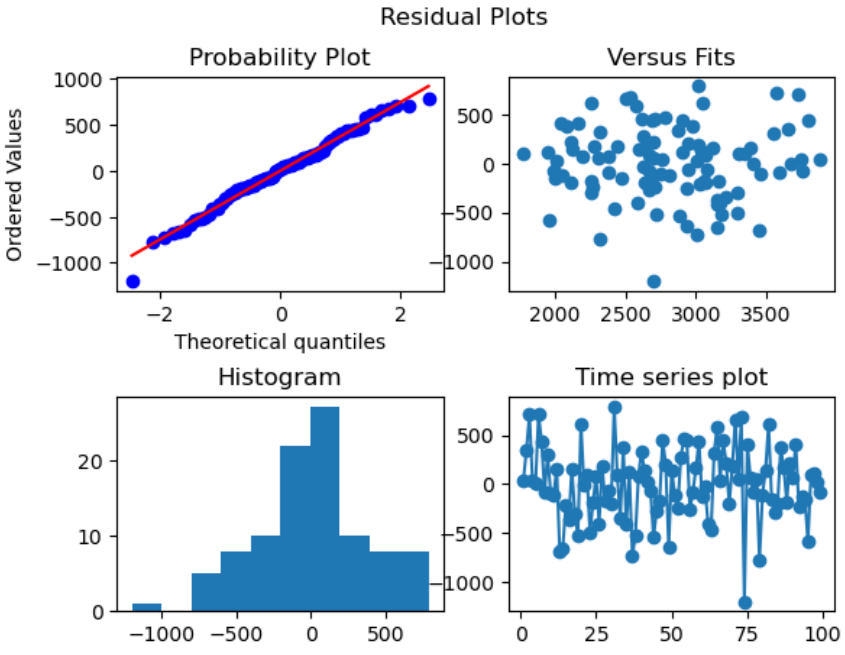


From the sample ACF and PACF it seems that the time series of the residuals is now stationary but still autocorrelated. The ACF seems to follow a "geometric" decay while only the first lag of the PACF is high. These patterns suggest to use an AR(1) model for the data.

Let's add an autoregressive term ("lag1") to the model.

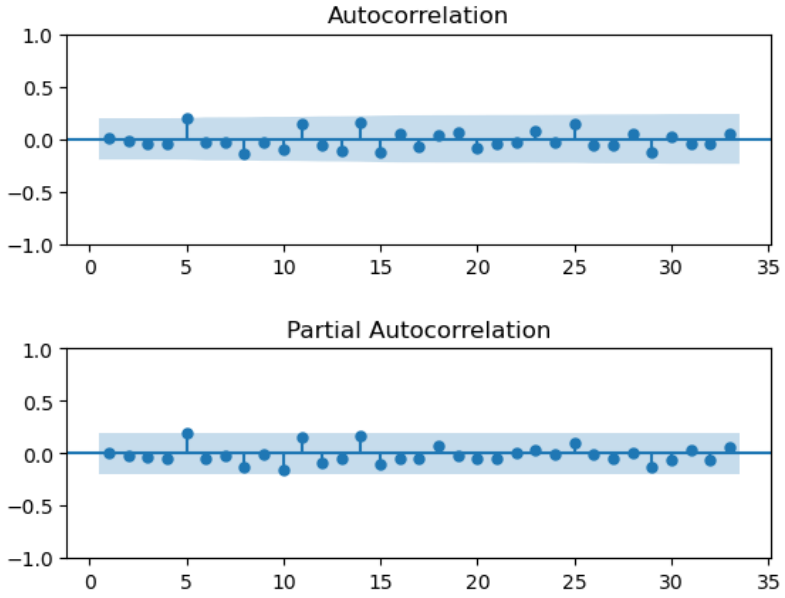
```
REGRESSION EQUATION
-------------------
val =  + 2255.631 const -10.355 layer + 0.379 lag1

COEFFICIENTS
------------
 Term      Coef  SE Coef  T-Value    P-Value
const 2255.6311 354.4765   6.3633 6.7234e-09
layer  -10.3548   2.0706  -5.0009 2.5719e-06
 lag1    0.3787   0.0940   4.0280 1.1254e-04

MODEL SUMMARY
-------------
     S   R-sq  R-sq(adj)
375.44 0.6493      0.642

ANALYSIS OF VARIANCE
--------------------
    Source   DF     Adj SS     Adj MS  F-Value    P-Value
Regression  2.0 2.5050e+07 1.2525e+07  88.8578 1.4409e-22
     const  1.0 5.7075e+06 5.7075e+06  40.4912 6.7234e-09
     layer  1.0 3.5251e+06 3.5251e+06  25.0086 2.5719e-06
      lag1  1.0 2.2869e+06 2.2869e+06  16.2244 1.1254e-04
     Error 96.0 1.3532e+07 1.4096e+05      NaN        NaN
     Total 98.0 3.8582e+07        NaN      NaN        NaN
```

Let's check the assumptions on the residuals.



The normality assumption is met (Shapiro-Wilk's test p-value = 0.514).

The randomness assumption is met (Runs test p-value = 0.631).



All the terms are significant, including the new autoregressive term. The R-sq(adj) has also improved from the first model.

All assumptions are met. The model is adequate.

b)

The oxygen concentration at the end of the next layer can be computed from the fitted regression equation:

```
val =  + 2255.631 const -10.355 layer + 0.379 lag1

const = 1
layer = 101
lag1 = 1910

val = 1933.1333
```

c)

The special cause control chart is suitable for this situation. We can design an I-MR control chart on the residuals of the last model using the following formulas to compute the control limits.

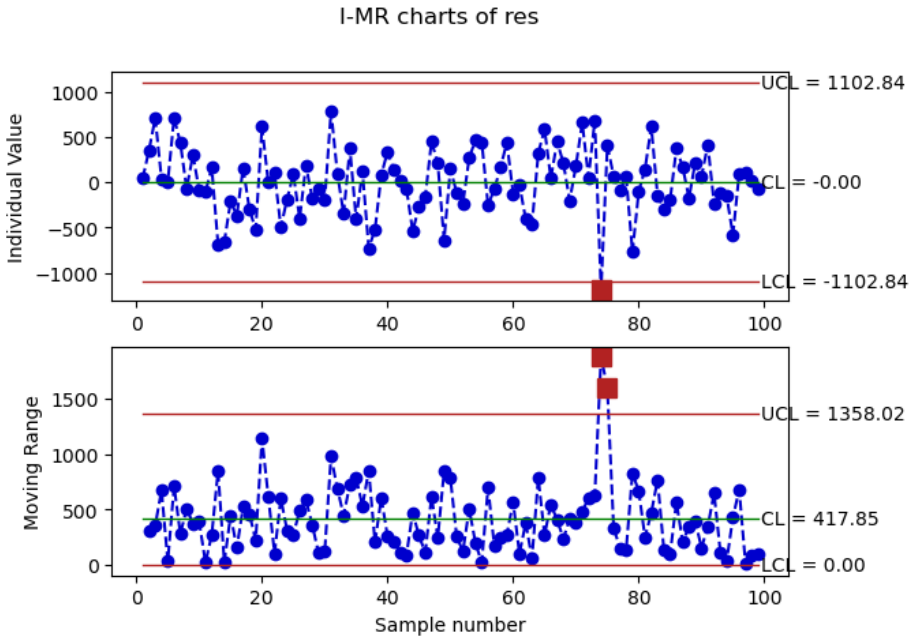
I chart:

- $UCL = \bar{x} + 3\left(\frac{\bar{MR}}{d_2}\right)$
- $CL = \bar{x}$
- $LCL = \bar{x} - 3\left(\frac{\bar{MR}}{d_2}\right)$

MR chart:

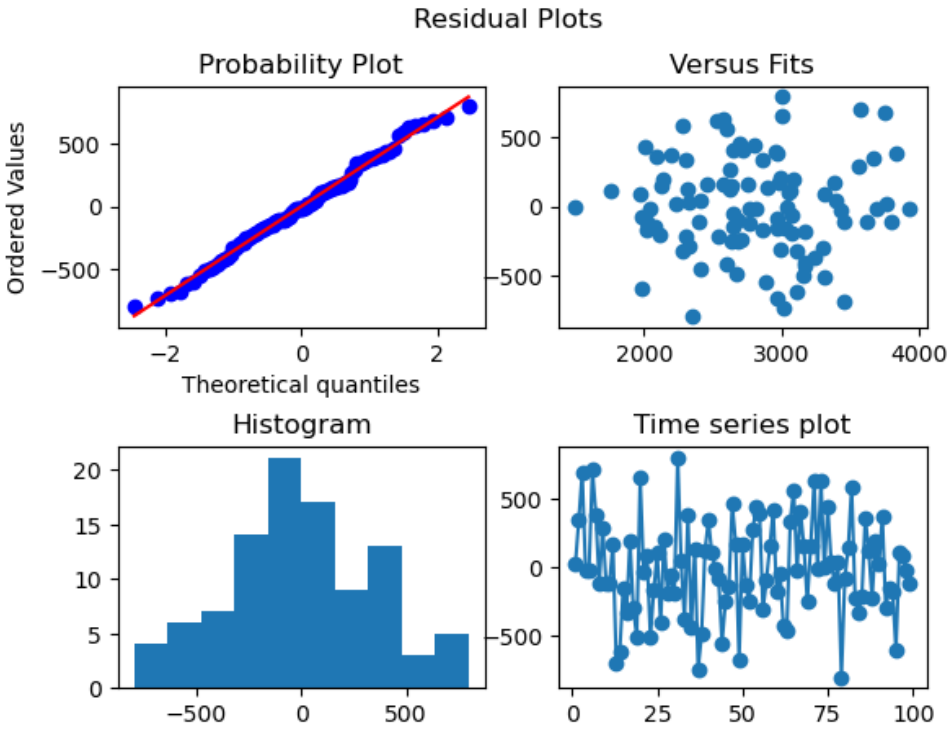
- $UCL = D_4\bar{MR}$
- $CL = \bar{MR}$
- $LCL = 0$

The type I error α must be equal to 0.0029, therefore the K we need to use to compute the values of the factors d2 and D4 is equal to 2.978.



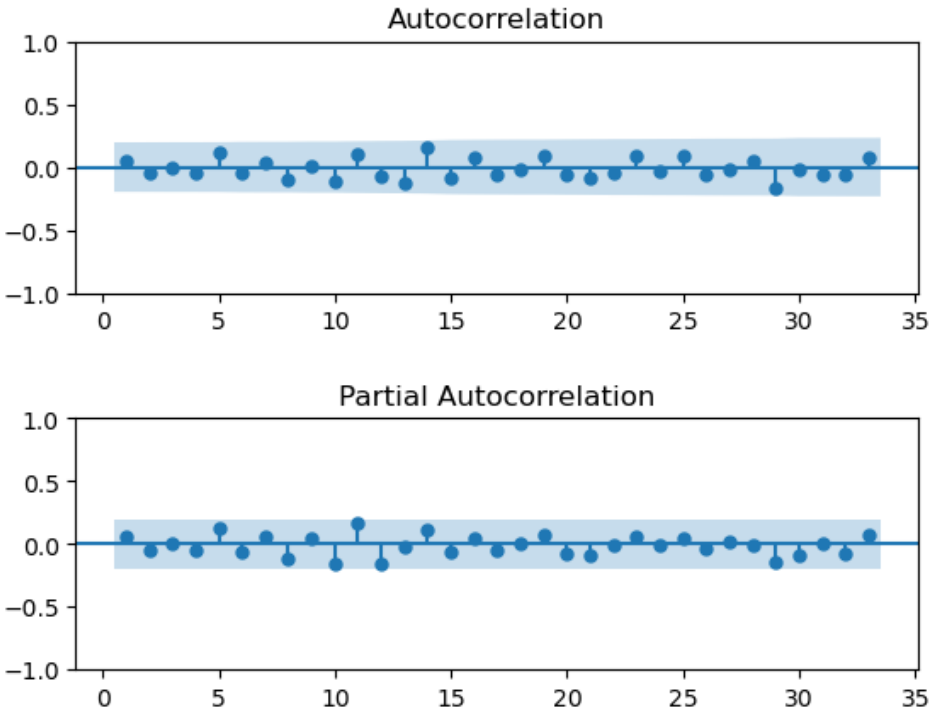
At index 74 there is a violation of the control limits in the I cc. This violation seems to affect the MR cc as well. Since we are assuming that an assignable cause was found, we need to remove the OOC points from the data and re-estimate the control limits of the CC. To do so, we need to re-fit the model using a dummy variable as regressor (= 0 for all layers except for the OOC layer).

```
REGRESSION EQUATION
-------------------
val =  + 2014.440 const -8.896 layer + 0.443 lag1 -1267.733 dummy

COEFFICIENTS
------------
 Term       Coef  SE Coef  T-Value    P-Value
const  2014.4399 342.9623   5.8736 6.2616e-08
layer    -8.8960   2.0063  -4.4340 2.4842e-05
 lag1     0.4425   0.0910   4.8652 4.5320e-06
dummy -1267.7329 366.3758  -3.4602 8.1015e-04

MODEL SUMMARY
-------------
       S   R-sq  R-sq(adj)
355.6633 0.6885     0.6787

ANALYSIS OF VARIANCE
--------------------
    Source   DF     Adj SS     Adj MS  F-Value    P-Value
Regression  3.0 2.6564e+07 8.8548e+06  70.0006 5.6583e-24
     const  1.0 4.3641e+06 4.3641e+06  34.4997 6.2616e-08
     layer  1.0 2.4869e+06 2.4869e+06  19.6600 2.4842e-05
      lag1  1.0 2.9941e+06 2.9941e+06  23.6698 4.5320e-06
     dummy  1.0 1.5145e+06 1.5145e+06  11.9730 8.1015e-04
     Error 95.0 1.2017e+07 1.2650e+05      NaN        NaN
     Total 98.0 3.8582e+07        NaN      NaN        NaN
```



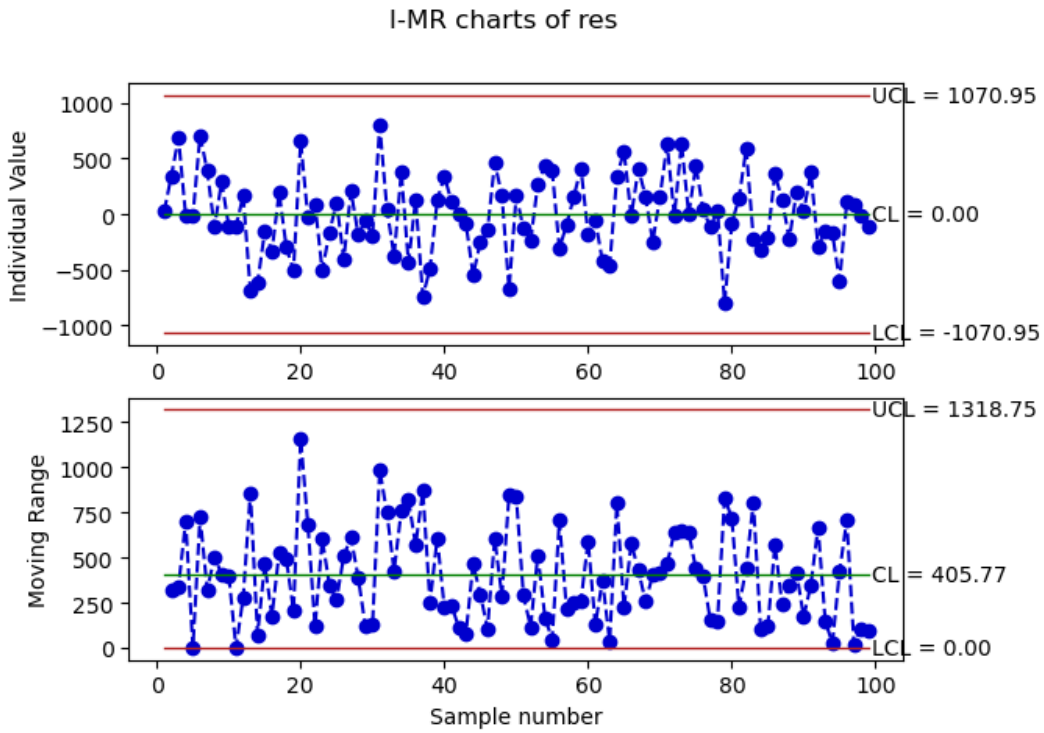
The normality assumption is met (Shapiro-Wilk's test p-value = 0.729).

The randomness assumption is met (Runs test p-value = 0.742).



The model is adequate.

Now design the I-MR cc using the new residuals.



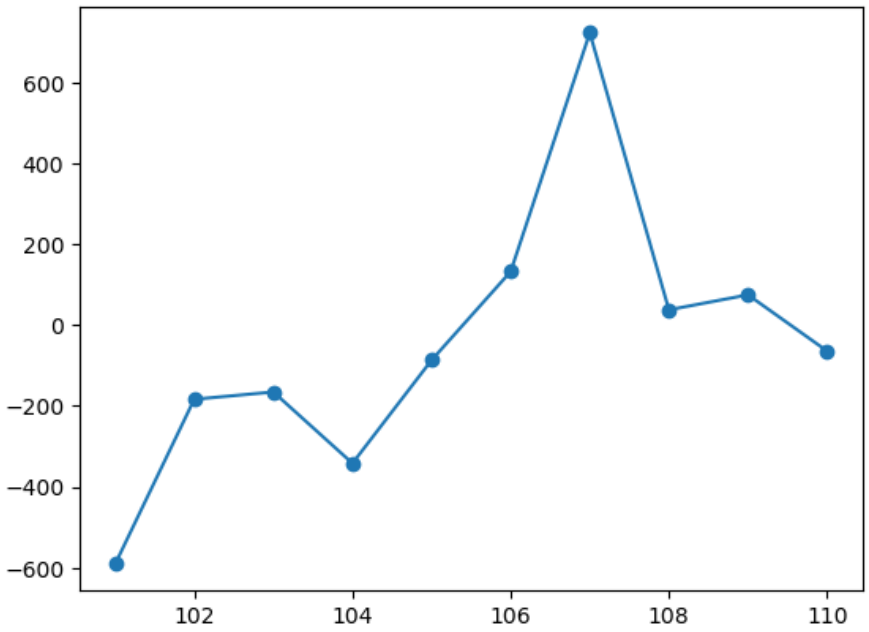
No out-of-control points are detected. Phase 1 is completed.

d)

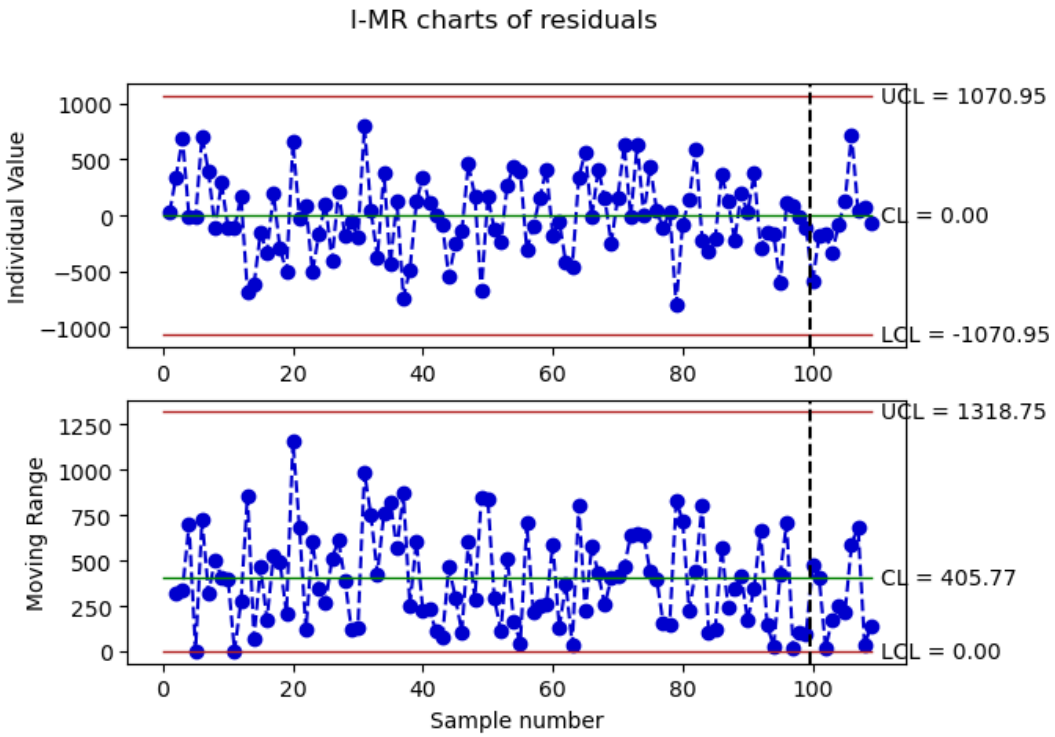
To determine whether the new data are in-control or not, the last model fitted in point 2 shall be used.

```
val =  + 0.201 const -0.001 layer + 0.443 lag1 -0.127 dummy
```

The resulting residuals for the layers from 101 to 110 are:



By plotting the new residuals on the previously design special cause control chart, we get:



The new observations are in control.

Exercise 2 solution

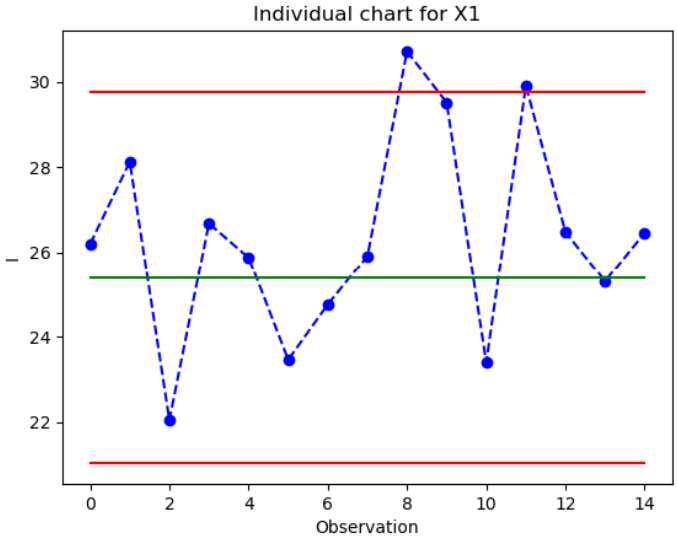
a)

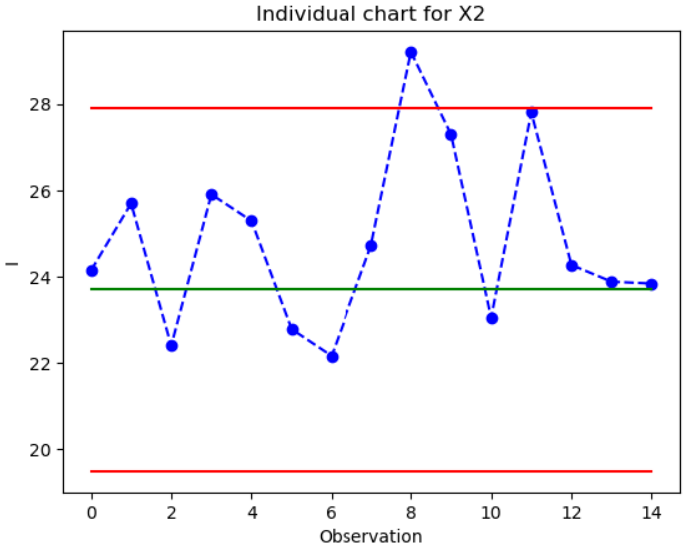
Using the Bonferroni's correction, each individual control chart is designed with a Type I error $\alpha' = \alpha/3$, where $\alpha = 0.01$, and hence K = 2.935.

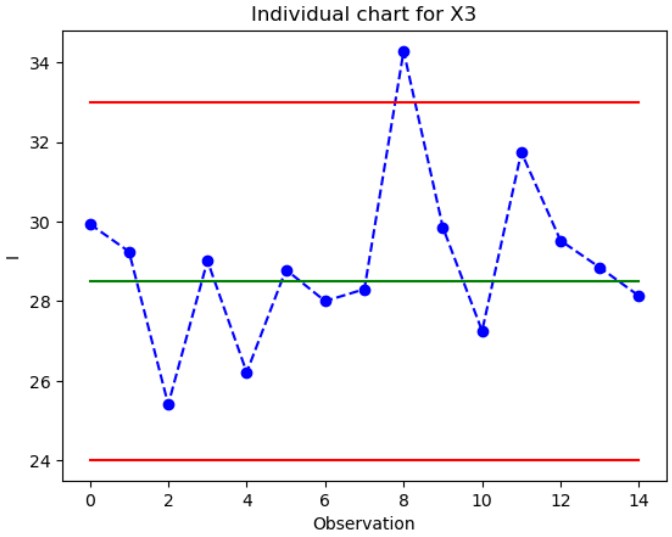
Being known the process parameters, the control limits of the three individual control charts are:

x_1	$UCL = \mu_1 + K\sigma_1 = 29.75$ $LCL = \mu_1 - K\sigma_1 = 21.05$
x_2	$UCL = \mu_2 + K\sigma_2 = 27.91$ $LCL = \mu_2 - K\sigma_2 = 19.48$
x_3	$UCL = \mu_3 + K\sigma_3 = 32.99$ $LCL = \mu_3 - K\sigma_3 = 24.01$

Phase 2 control charts are the following:







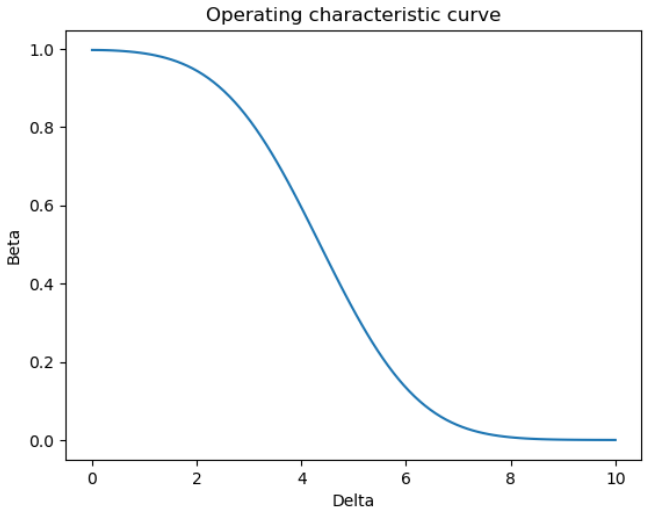
Observation n. 8 is out of control in all the three control charts. The control chart for x1 also signals an out of control for observation n. 11.

b)

The Type II error β_1 for the control chart con x_1 is:

$$\beta_1 = \Phi\left(K - \frac{\Delta\mu_1}{\sigma_1}\right) - \Phi\left(-K - \frac{\Delta\mu_1}{\sigma_1}\right)$$

The operating characteristic curve is:



The Type II error for $\Delta\mu_1 = 3$ is $\beta = 0.814$.

c)

Under in-control conditions, the scores of the first principal component are:

$$\mathbf{z}_1 = [z_{11}, z_{21}, \dots, z_{n1}]^T = (\mathbf{X} - \boldsymbol{\mu})\mathbf{u}_1, \text{ such that } \mathbf{z}_1 \sim N(\mu_{PC1}, \sigma^2_{PC1})$$

Where:

$$\mu_{PC1} = 0$$

$$\sigma^2_{PC1} = \lambda_1 = 5.402$$

With $\alpha = 0.01$ we have K = 2.576

Thus, the control chart to monitor the first principal component is:

$$UCL = \mu_{PC1} + K\sigma_{PC1} = 5.987$$

$$CL = \mu_{PC1} = 0$$

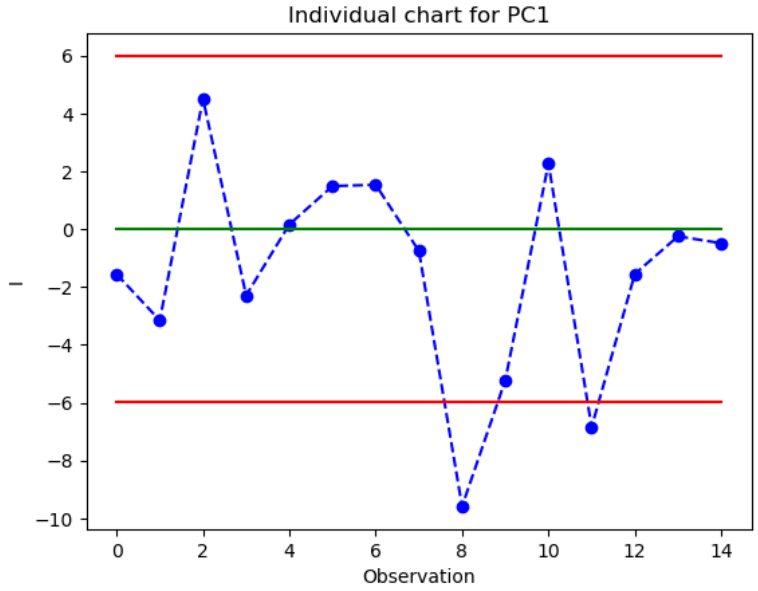
$$LCL = \mu_{PC1} - K\sigma_{PC1} = -5.987$$

To apply this control chart to the new data, they shall be projected along the direction spanned by the first principal component.

Being $\boldsymbol{u}_1 = [-0.589, -0.561, -0.582]^T$, the new data projections are:

$$\boldsymbol{z}_1 = -0.589(\boldsymbol{x}_1 - \mu_1) - 0.561(\boldsymbol{x}_2 - \mu_2) - 0.582(\boldsymbol{x}_3 - \mu_3)$$

The individual control chart on the first principal component is:



The control chart confirms the results obtained in point a), i.e., observations n. 8 and 11 are signaled as out of control.

d)

In out-of-control conditions we have $\boldsymbol{\mu}^* = [\mu_1 + \Delta\mu_1, \mu_2, \mu_3]^T$, and hence the out-of-control mean of the principal component is:

$$\mu^*_{PC1} = \mathrm{E}[u_{11}(X - \mu_1) + u_{12}(X - \mu_2) + u_{13}(X - \mu_3)] = u_{11}\Delta\mu_1$$

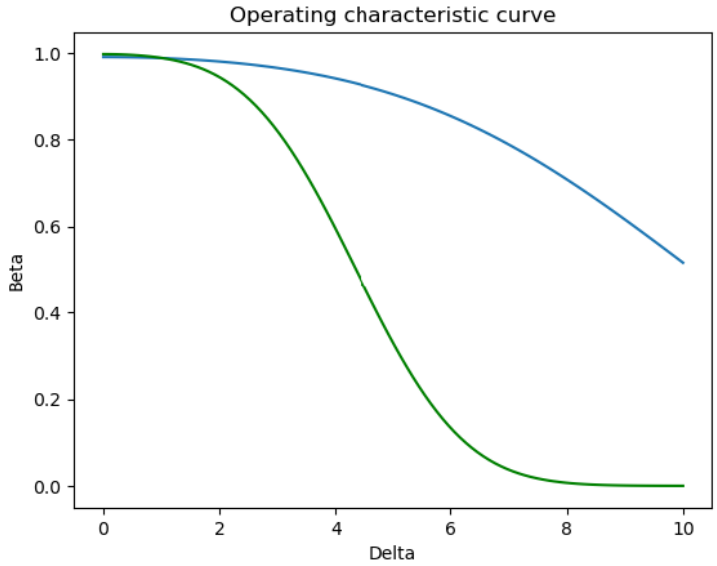
The Type II error can be computed as follows:

$$\beta_1 = \Phi\left(\frac{UCL - u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right) - \Phi\left(\frac{LCL - u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right) =$$

$$= \Phi\left(\frac{K\sqrt{\lambda} - u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right) - \Phi\left(\frac{-K\sqrt{\lambda} - u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right) =$$

$$= \Phi\left(K - \frac{u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right) - \Phi\left(-K - \frac{u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right)$$

The resulting operating characteristic curve is the one in blue in the next figure, where the green curve is the operating characteristic curve computed in point b):



The Type II error for $\Delta\mu_1 = 3$ is $\beta_1 = 0.964$

In case the shift affects the mean of one single variable, the control chart on the first PC is less effective than applying univariate control charts on the original variables. The reason is that the first PC is a basically a mean of the three variables (remind: $\boldsymbol{u}_1 = [-0.589, -0.561, -0.582]^T$), and hence the deviation on one signal variable is mitigated by the weight in the linear combination.

Exercise 3 solution

1)

Answer: d

Explanation: In the case where Alpha-to-Enter > Alpha-to-Remove there is a possibility the algorithm to fall in an infinite loop. For example, assume that we fix Alpha-to-Enter = 0.05 and Alpha-to-Remove = 0.01 and at some iteration of the stepwise approach in the forward step we have the smallest p-value = 0.03 for a variable X_k. Then as p-value(X_k)=0.03 < 0.05=Alpha-to-Enter, the algorithm will include X_k in the model. At the next step (backward elimination) though, we will have p-value(X_k)=0.03 > 0.01=Alpha-to-Remove and thus the variable will be moved out of the model and so we will get into an infinite loop.

2)

i): Answer: b

Explanation: At the first step of the backward elimination, we will remove X_2, as it has the highest p-value(X_2)=0.241, which exceeds the threshold Alpha-to-Remove=0.05.

ii): Answer: d
After the first backward elimination, we need to rerun the regression model with the remaining four predictors {X_1, X_3, X_4 and X_5} and all the p-values will be updated, and it is not possible to know in advance which will be the larger and whether this will still exceed the Alpha-to-Remove=0.05 threshold. Thus, we cannot tell.

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 1 point 4) and exercise 3 question 1).**

Exercise 1 (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 2 (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 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. $r(X, Y)$?

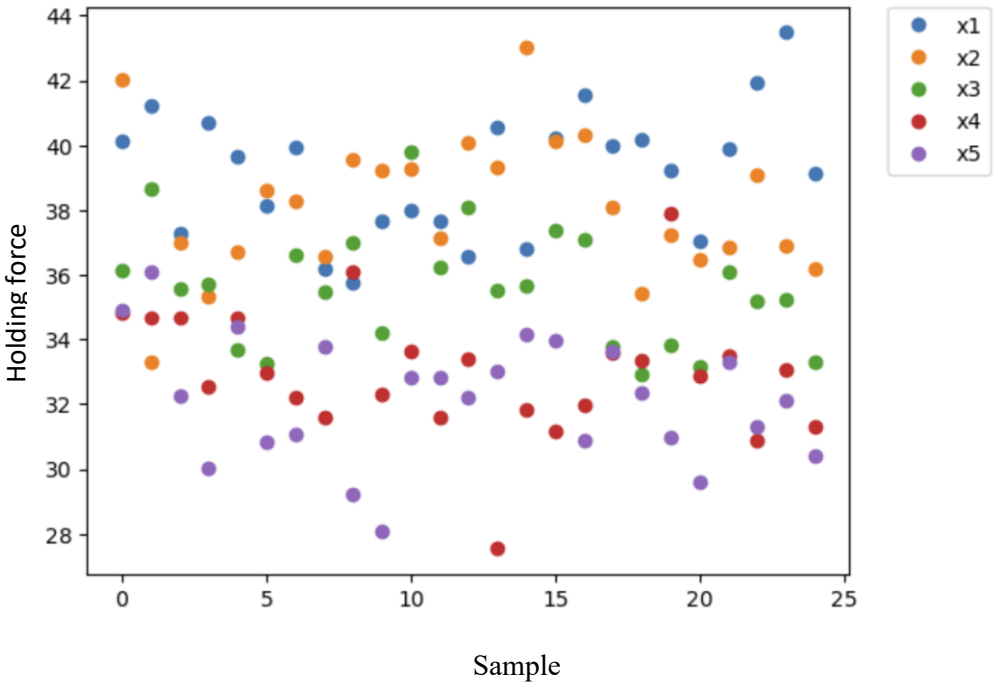
a) $\mathrm{r(X, Y)} = +0.9$
b) $\mathrm{r(X, Y)} = -0.9$
c) $\mathrm{r(X, Y)} = 0$
d) None of the above

la b) perchè R2 nella simple lin regress è il quadrato del coefficiente di correlazione di pearson. è b) e non a) perchè il modello ha catturato correlazione negeativa fra X e Y

Exercise 1 solution

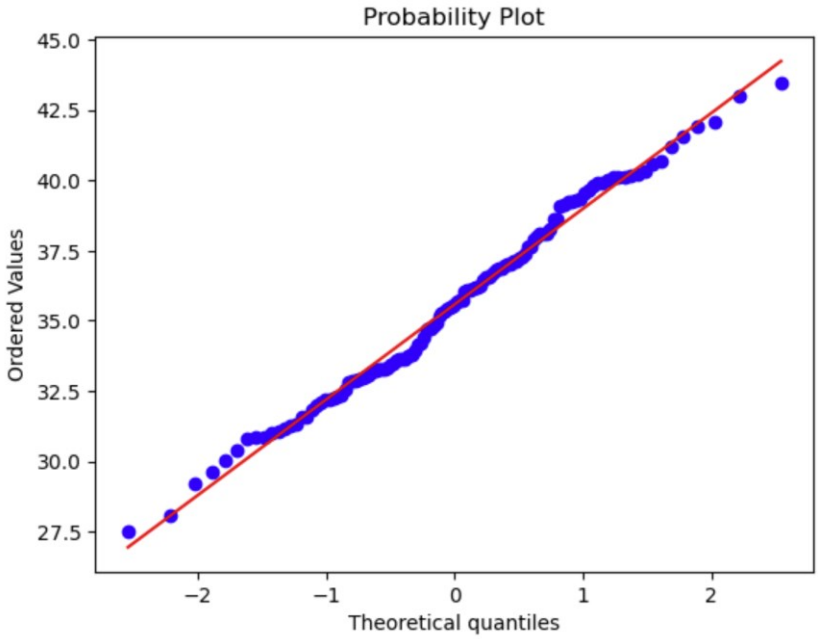
1)

Since the time order of available measurements is known only between samples, and not within samples, we can plot the data as follows, where different colors were used for individual observations in the samples :



There seems to be no systematic pattern over time (between samples), whereas a systematic pattern may be present within the sample, as moving from observation x1 to x5 a decreasing mean may be present. This is a potential violation of randomness assumption one shall take care of.

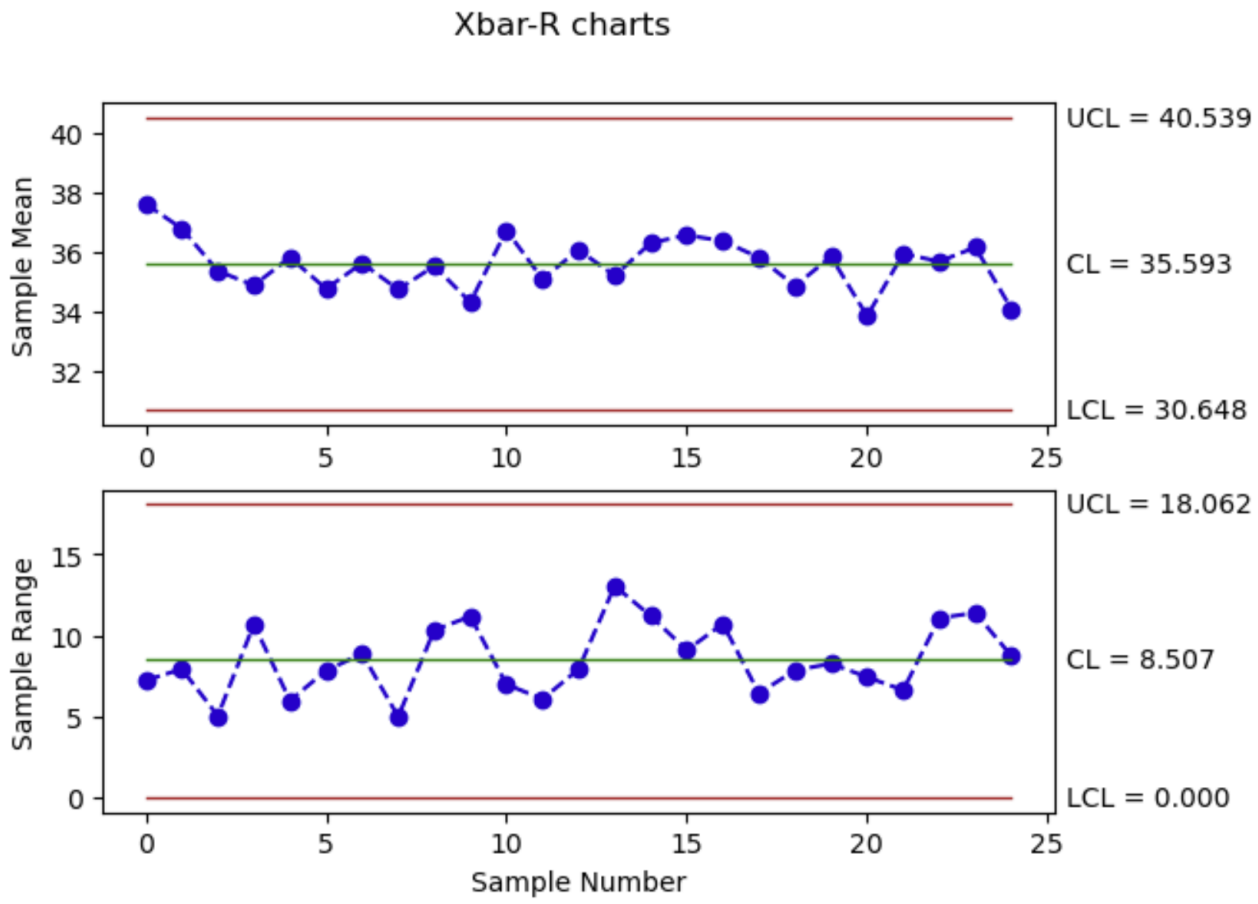
The normality assumption is met (Shapiro-Wilk's test p-value = 0.351).



Without knowing the time order within the sample it is not possible to make additional tests to check the randomness of the data. Let's assume data to be normal and independent, design the Xbar-R control chart and discuss the results.

Being $ARL_0 = 350$, then $\alpha = \frac{1}{350} = 0.0025$ and K = 3.023.

The resulting control chart with n = 5 is:

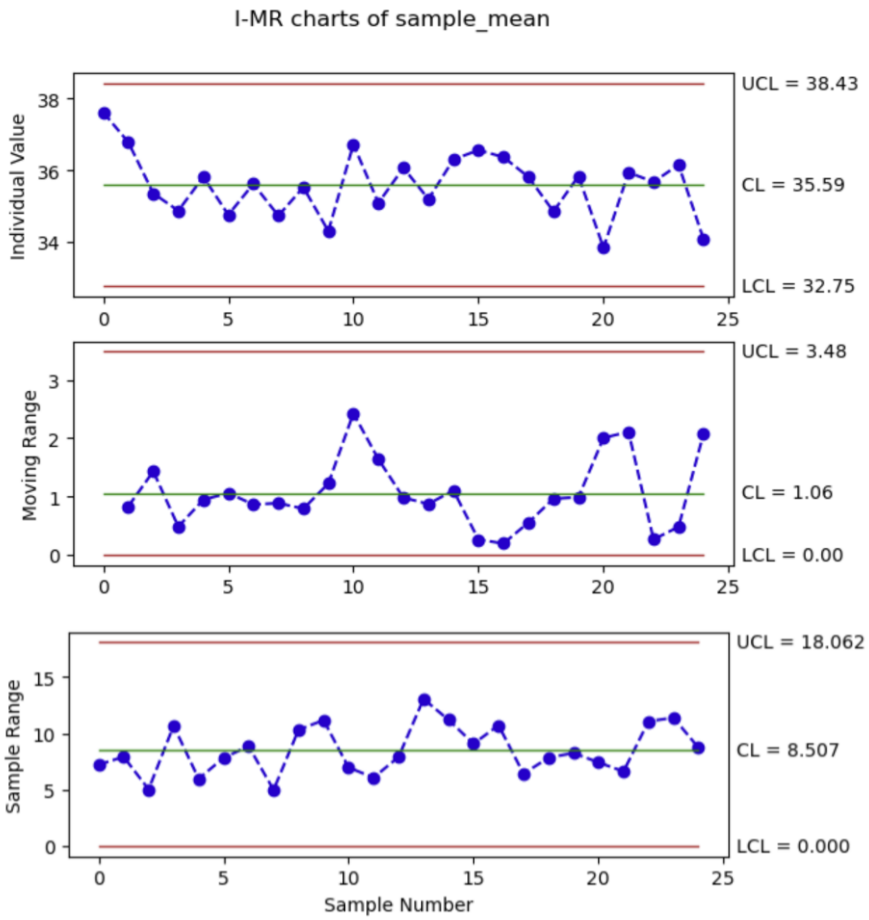


Hugging is evident in the X-bar chart. This may be the consequence of a violation of randomness assumptions *within* the sample. Thus, the X-bar – R control chart is not an appropriate statistical monitoring method for these data.

2)

A more appropriate approach would consist of designing an I-MR-R control chart, to monitor the within and between sample variability.

By using the same ARL_0 used in point 1), the I-MR-R control chart is the following:

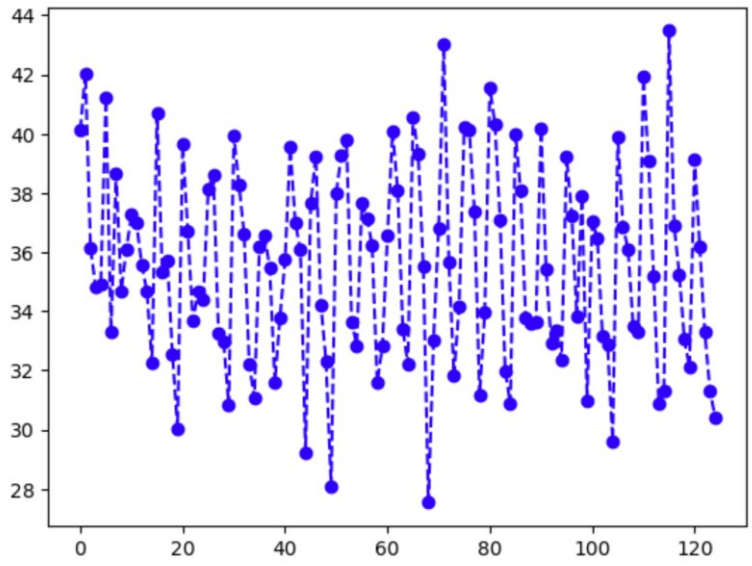


No violation of the control limit is present. The control chart design phase is over.

3)

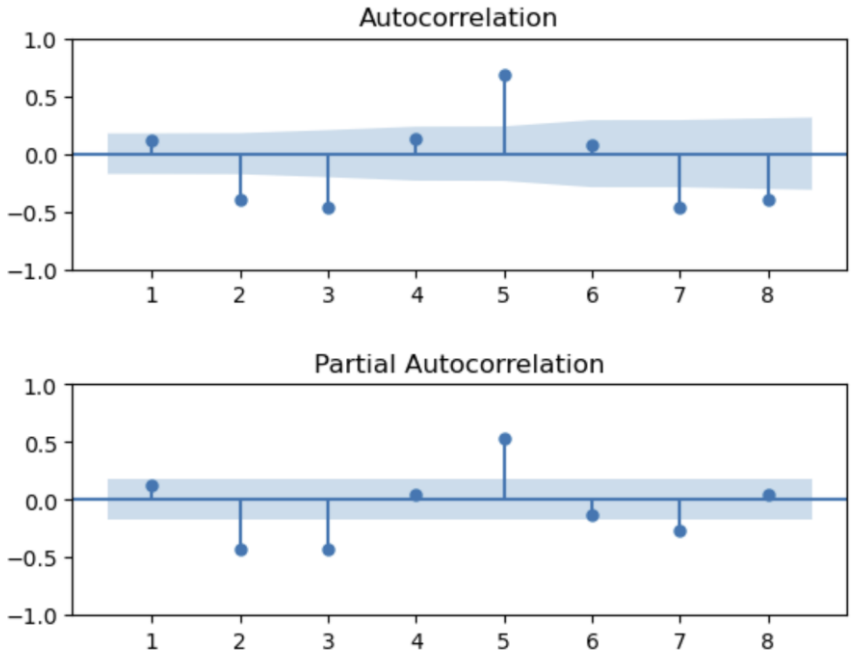
An even more effective control charting scheme would consist of modelling the systematic source of non-random variability within the sample. Being known the time order of individual measurements within the samples, it is possible to make additional analysis and tests.

The time series pattern of the individual measurements is the following:

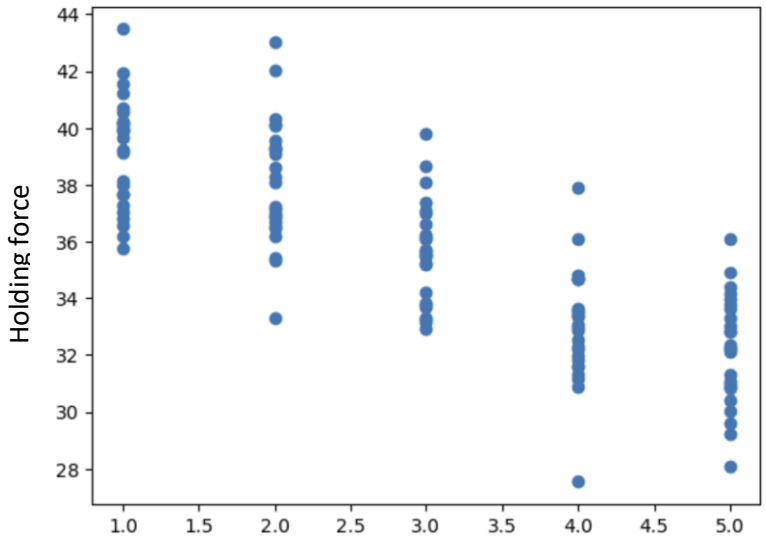


The p-value of the runs-test is p-val = 0.178.

Regardless of the runs-test result, a systematic and non-random pattern in the data is evident, as confirmed by the sample ACF and PACF functions.



A positive autocorrelation is present at lag 5, corresponding to the sample size n = 5. To further investigate this dependence, it is possible to plot the individual measurements for each actuator with respect to the time-order of within-sample measurements:



The figure shows that there is a decreasing trend of the holding force values moving from the first observation in the sample to the last one.

One way to model this pattern is to use as a regressor the time order within the sample. This can be done by using a “within sample trend” regressor, referred to as *t*, defined as follows:

	index	variable	value	t
0	0	0	40.12903...	1
1	1	0	42.04866...	2
2	2	0	36.15604...	3
3	3	0	34.82940...	4
4	4	0	34.92002...	5
5	5	1	41.20058...	1
6	6	1	33.28095...	2
7	7	1	38.64877...	3
8	8	1	34.69688...	4
9	9	1	36.10008...	5
10	10	2	37.27697...	1
11	11	2	36.98005...	2
12	12	2	35.57365...	3
13	13	2	34.69850...	4
14	14	2	32.24100...	5
15	15	3	40.69518...	1
16	16	3	35.35674...	2
17	17	3	35.73104...	3
18	18	3	32.54150...	4
19	19	3	30.02019...	5

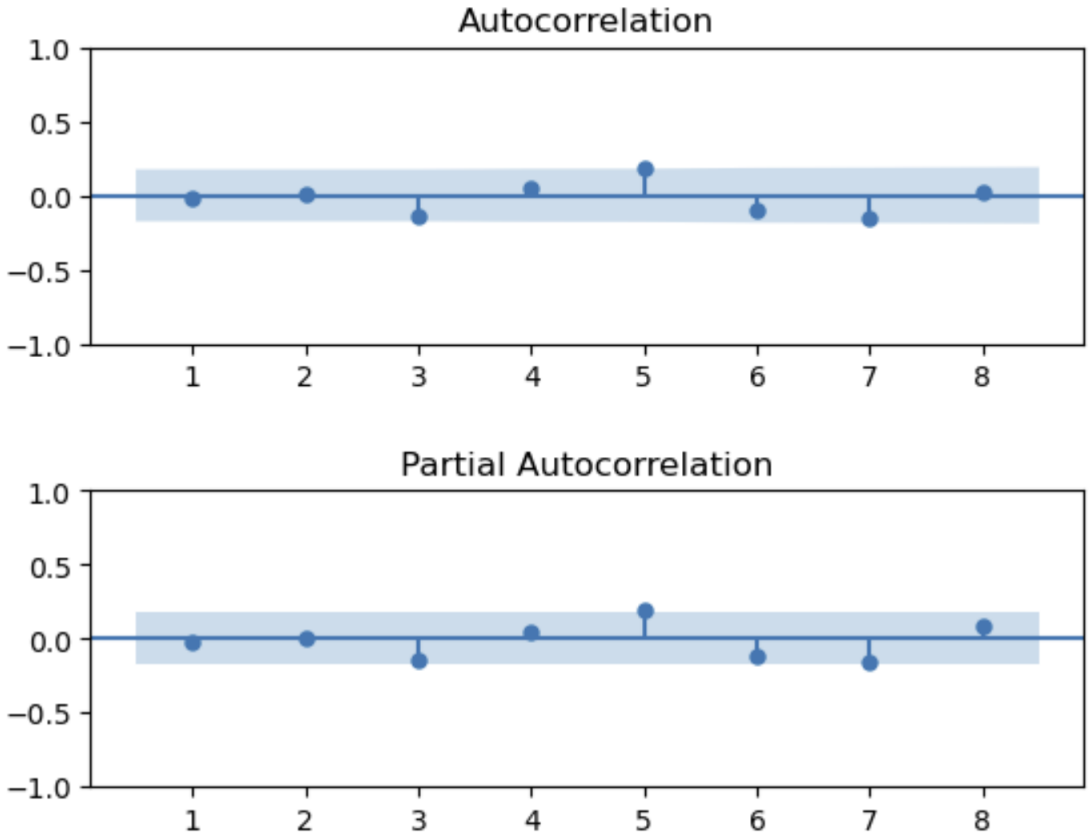
Let’s fit a linear model of the holding force against the variable t.

```
REGRESSION EQUATION
-------------------
Holding force =  41.328 - 1.911 t

COEFFICIENTS
------------
 Term    Coef  SE Coef  T-Value     P-Value
const 41.3275   0.4224  97.8480 1.6187e-118
t     -1.9114   0.1273 -15.0094  1.4097e-29

MODEL SUMMARY
-------------
     S   R-sq  R-sq(adj)
2.0135 0.6468      0.644

ANALYSIS OF VARIANCE
--------------------
    Source    DF     Adj SS     Adj MS   F-Value     P-Value
Regression   1.0   913.3796   913.3796  225.2833  1.4097e-29
     const   1.0 38817.3565 38817.3565 9574.2273 1.6187e-118
     t       1.0   913.3796   913.3796  225.2833  1.4097e-29
     Error 123.0   498.6862     4.0544       NaN         NaN
```

```
Total 124.0  1412.0657        NaN        NaN          NaN
```

Let’s check model residuals.

Sample ACF and PACF



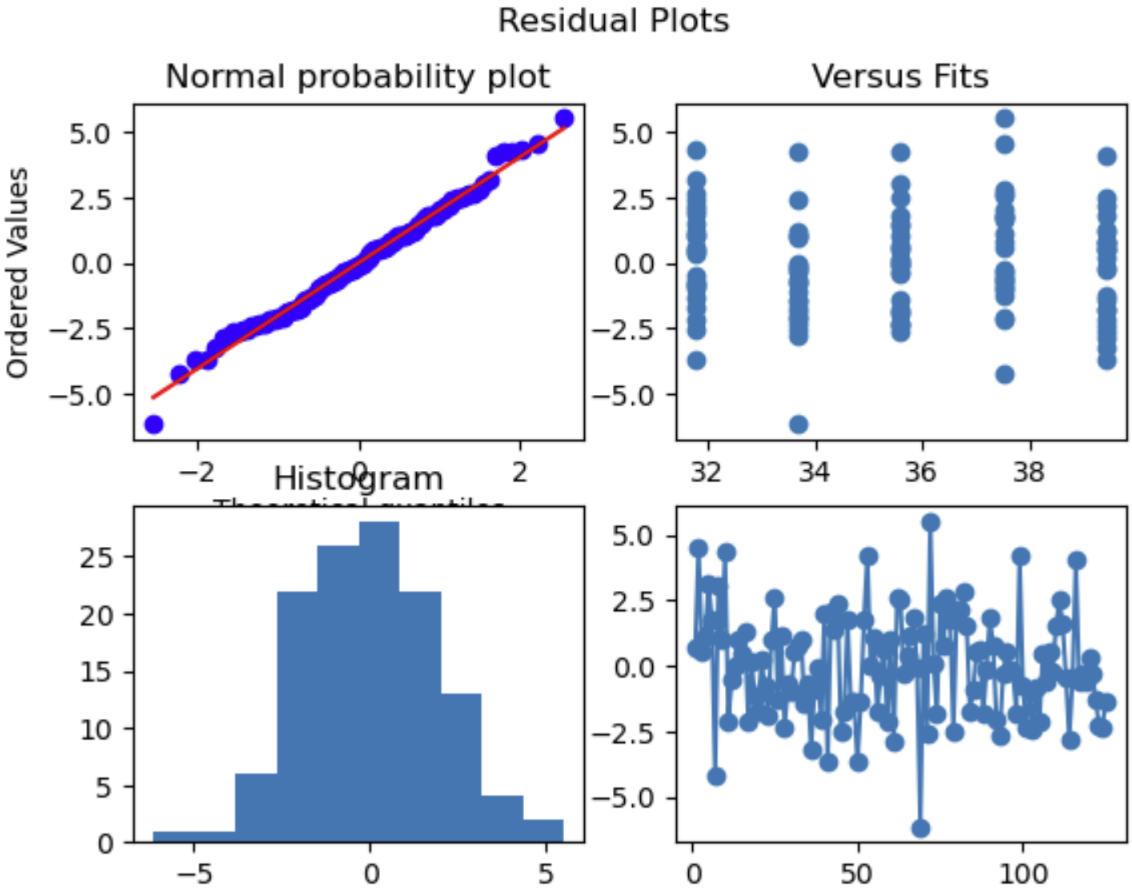
Runs-test of residuals: p-val 0.552

Bartlett test at lag 5 (with alpha = 0.05):

```
Test statistic rk = 0.185666
Rejection region starts at 0.391993
```

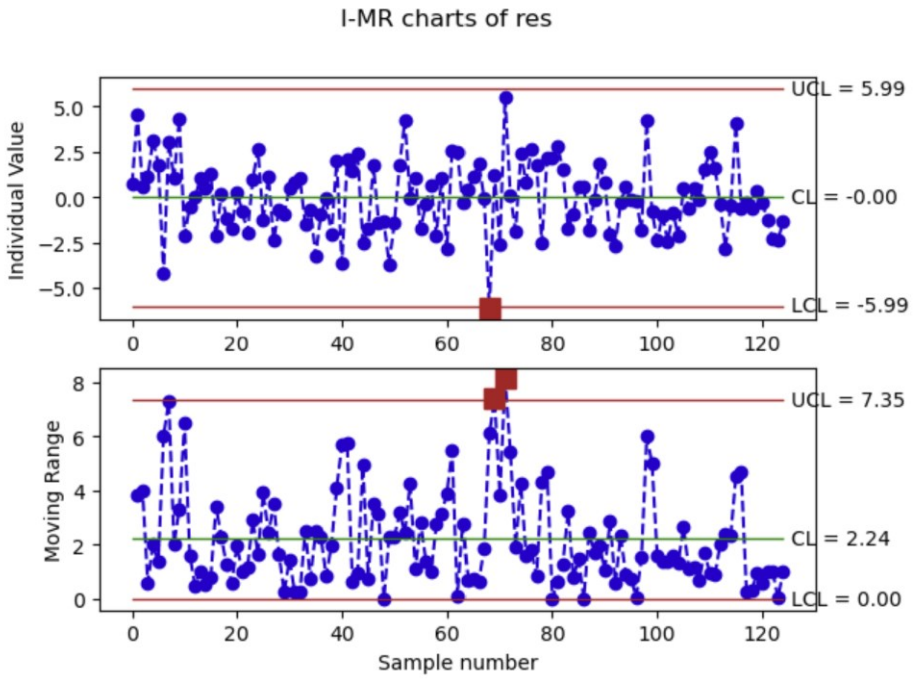
According to the runs-test and Bartlett test, residuals are random.

Shapiro-Wilk’s test on the normality of residuals: p-val = 0.757



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

The special cause control chart with $ARL_0 = 400$ for the model residuals is the following:



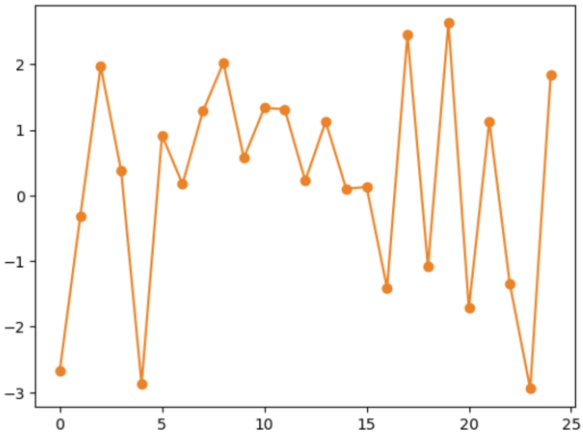
Two violations of the limits are present, but according to the exercise text, no assignable cause is found for them. Therefore, they can be labelled as false alarms. The control chart design is over.

4)

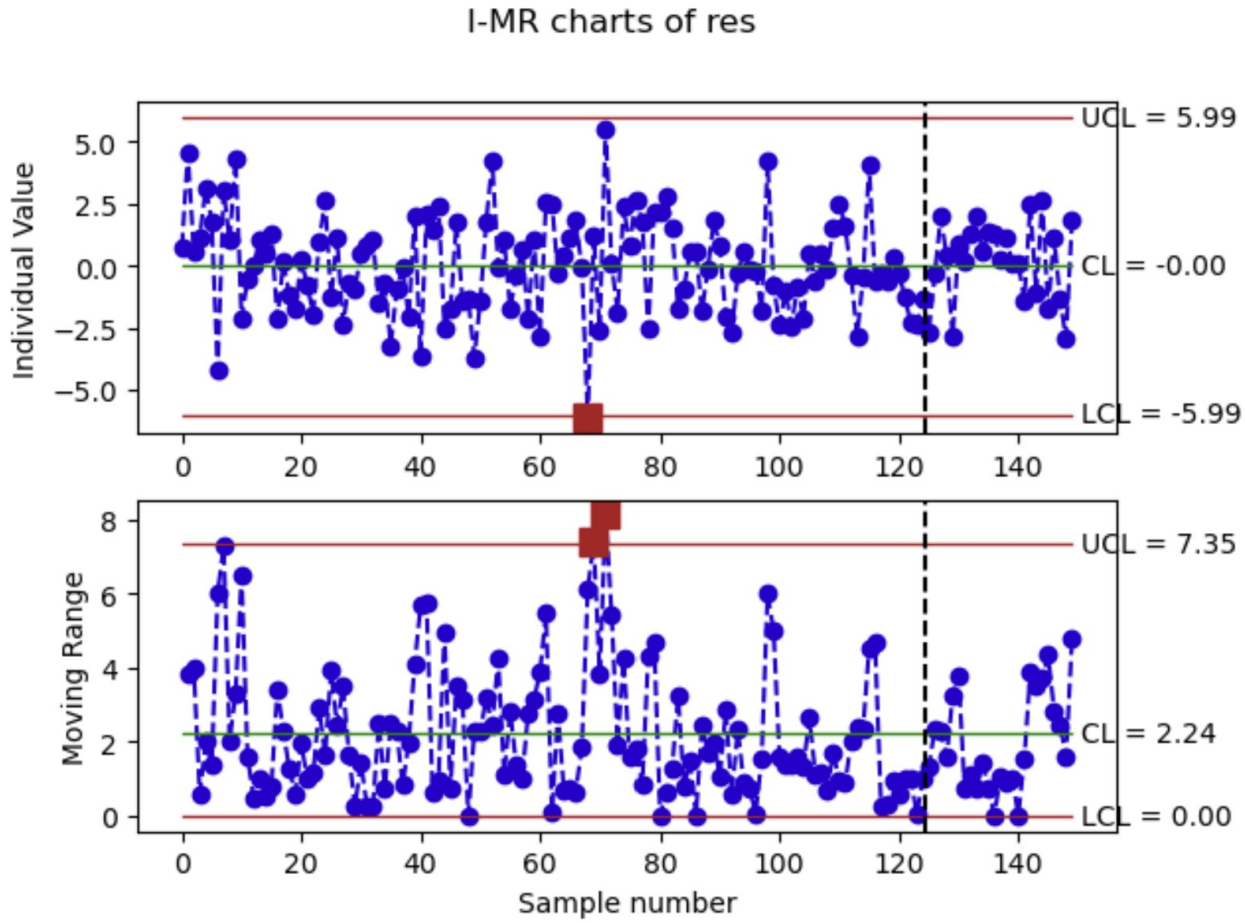
To determine whether new data are in-control or not, the same model used in point 3 shall be fit to them.

```
Holding force =  41.328 - 1.911 t
```

The resulting residuals for the new 5 samples are:



By plotting the new residuals on the previously design special cause control chart, we get:

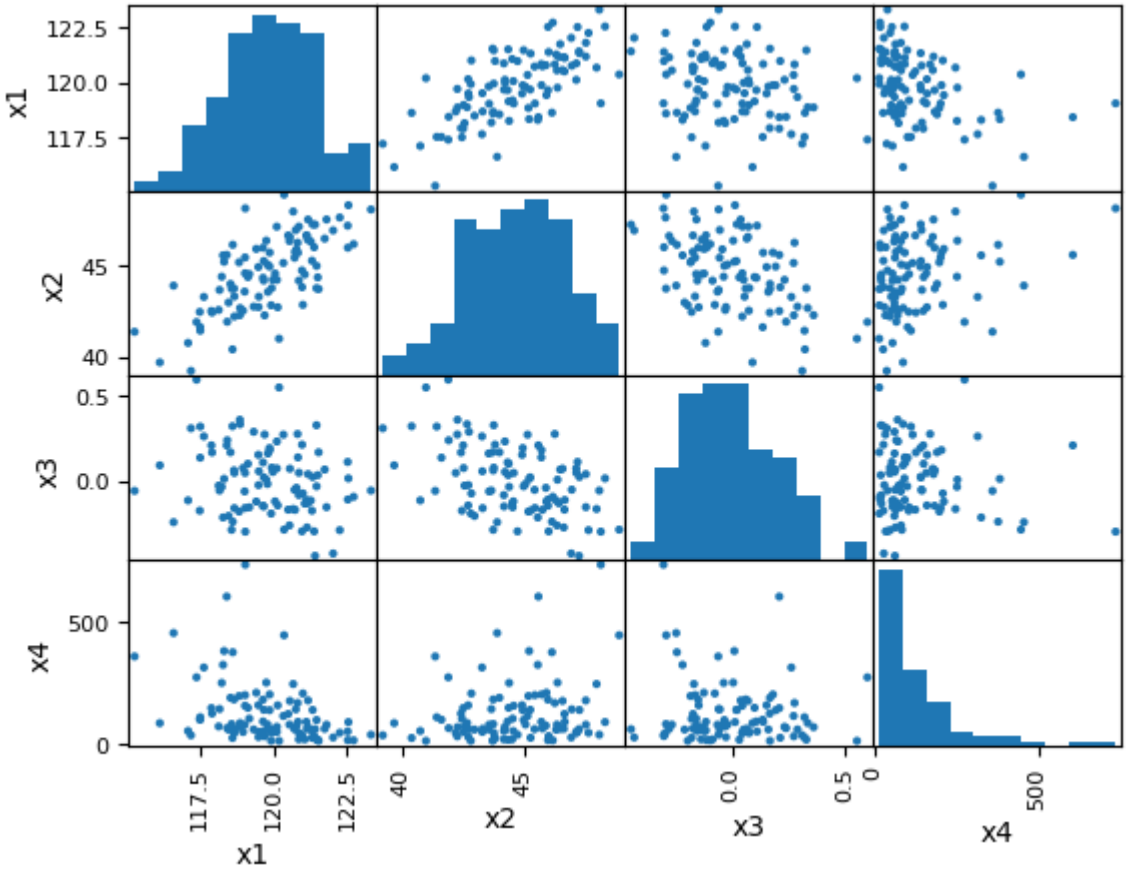


The new observations are in control.

Exercise 2 solutions

1)

Inspect the data and estimate the mean and the variance-covariance matrix.



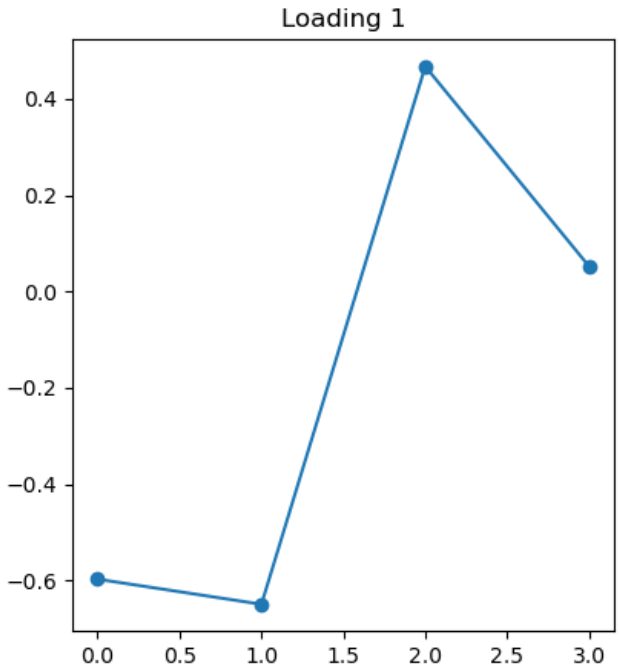

```
Sample Mean:
x1    119.78458
x2     44.62588
x3      0.00111
x4    127.80437
dtype: float64

Sample Variance-Covariance Matrix:
           x1         x2        x3            x4
x1   2.411731   2.201534 -0.074078    -75.914402
x2   2.201534   4.235564 -0.203526     54.559614
x3  -0.074078  -0.203526  0.041152     -2.478084
x4 -75.914402  54.559614 -2.478084  15482.351080
```

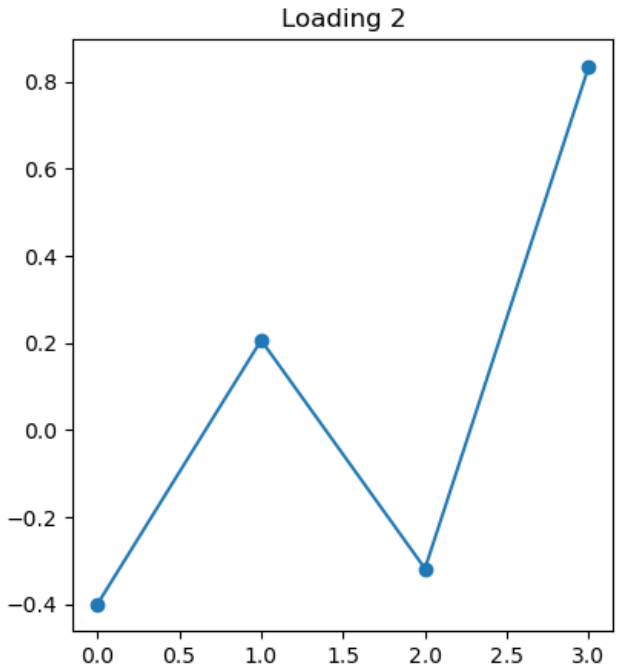
Since there is a large difference in the scale and variance of the individual variables, we shall apply PCA to the standardized data (which is equivalent to apply PCA using the correlation matrix of the original data).

```
Explained variance ratio

 [0.49179434 0.3195449  0.16620453 0.02245623]

Cumulative explained variance ratio
 [0.49179434 0.81133924 0.97754377 1]
```

We need to retain the first 2 PCs to capture at least 80% of the total variance.

```
Eigenvalues

 [1.96717736 1.2781796 ]

Eigenvectors
 [[-0.59742388 -0.64956733  0.46736347  0.05213804]
 [-0.39983512  0.20609723 -0.31777229  0.83467154]]
```

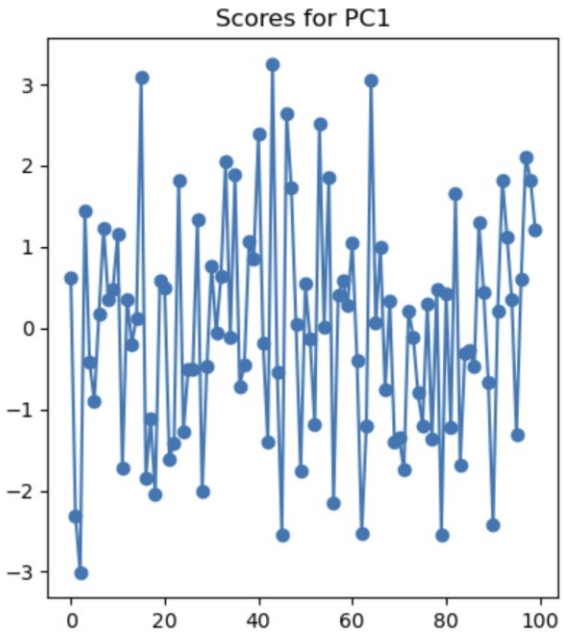


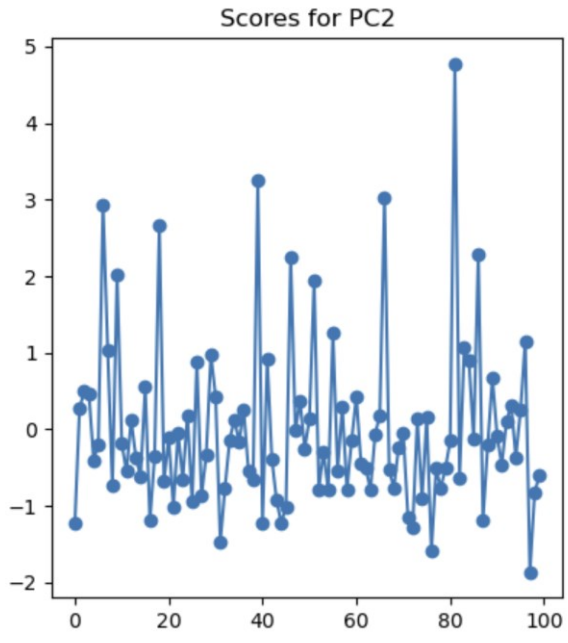


2)

We can design two univariate control charts for the first two PCs (uncorrelated) or one multivariate control chart, but first compute the scores and check the normality and independence assumptions.

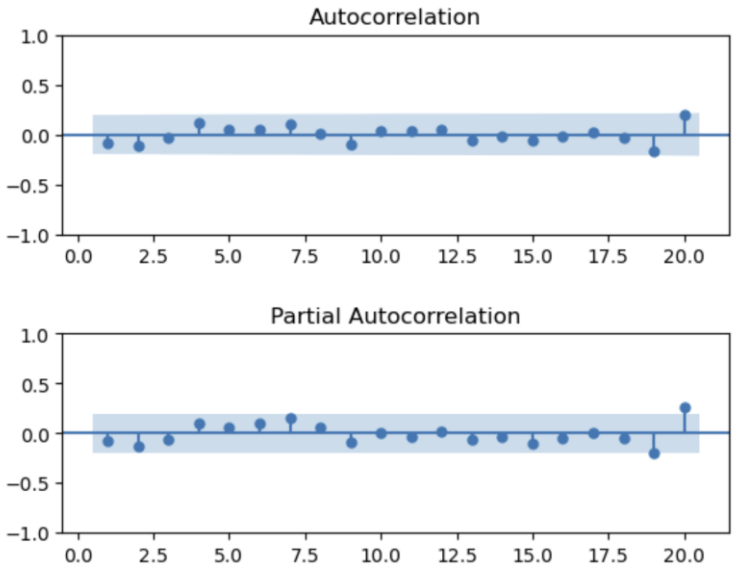
Randomness is met as shown below (although the p-value of the runs-test is borderline for PC2):





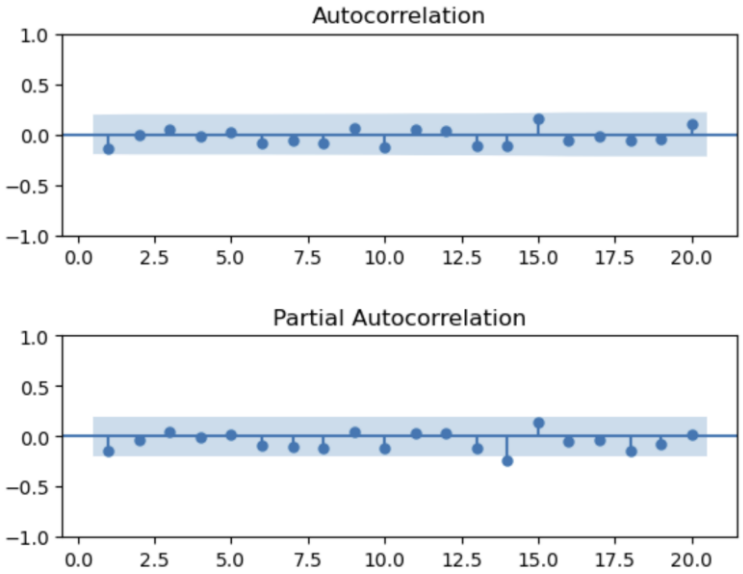
PC1: Runs test p-value = 0.675

PC1: SACF and SPACF:

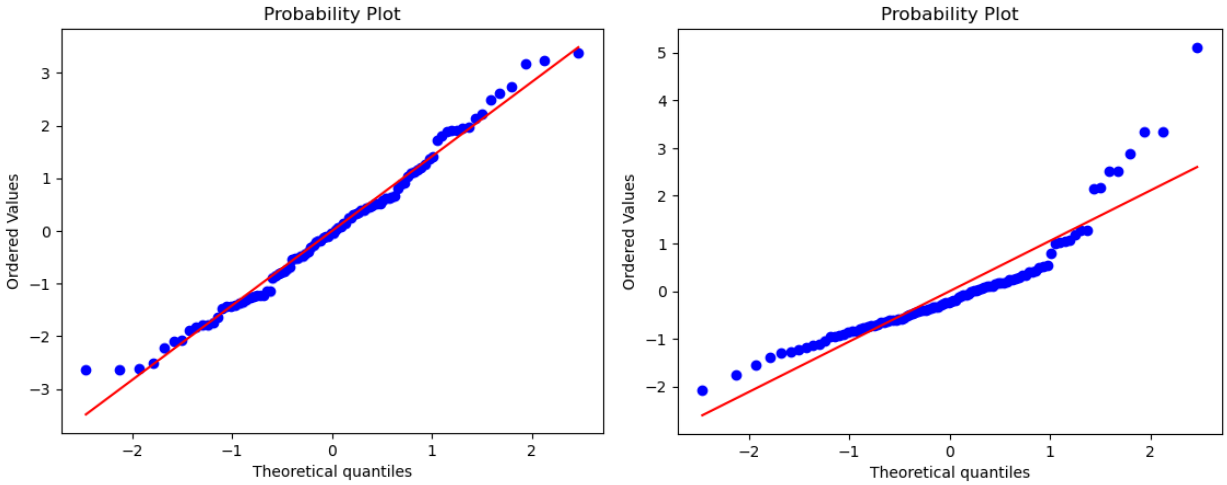


PC1: Runs test p-value = 0.043

PC1: SACF and SPACF:

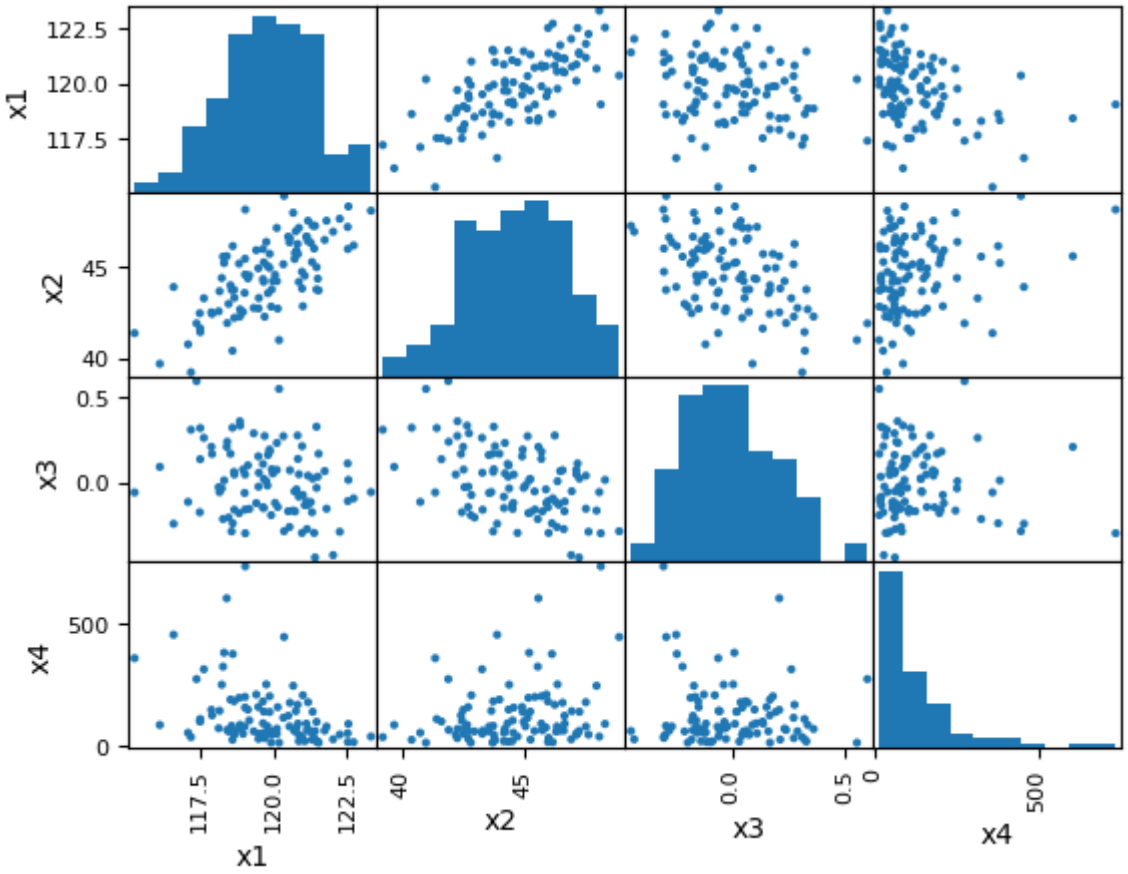


However, normality is violated for PC2:



The p-values of the Shapiro-Wilk's test are p-val = 0.244 and p-val = 0.000 for PC1 and PC2 respectively.

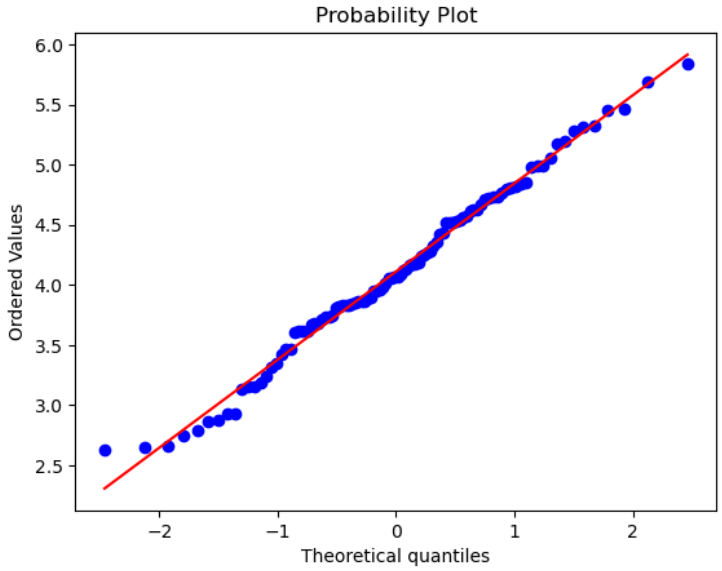
Let's have a look at the original dataset to check if there is some variable that may be responsible for the non-normality of the second PC.



We can see that the distribution of x4 is very skewed, and the weight associated to x4 in the second PC is very high. This is the reason why the second PC is not normal.

The test for normality on x4 gives a p-value = 0.000.

Apply Box-Cox to x4 to try to recover normality. After Box-Cox transformation (lambda ≈ 0.0), this is the distribution of x4 (SW p-value = 0.518).

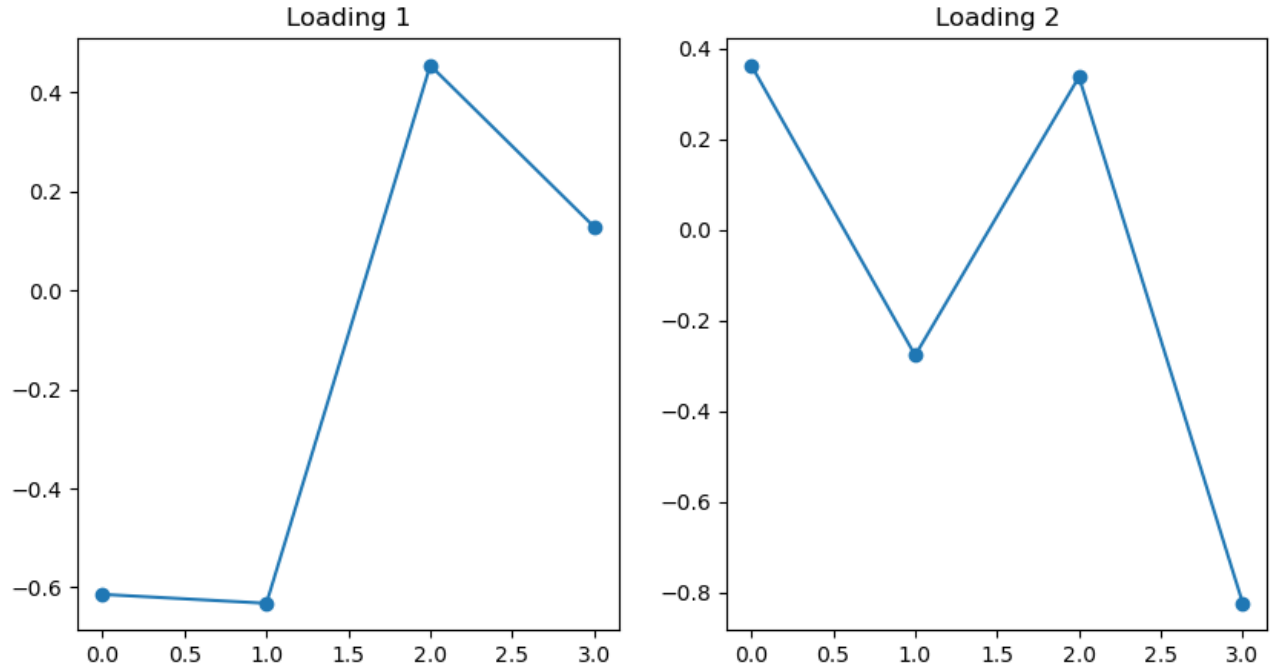


Let's standardize the new data (with the transformed variable) and re-estimate the PCA.

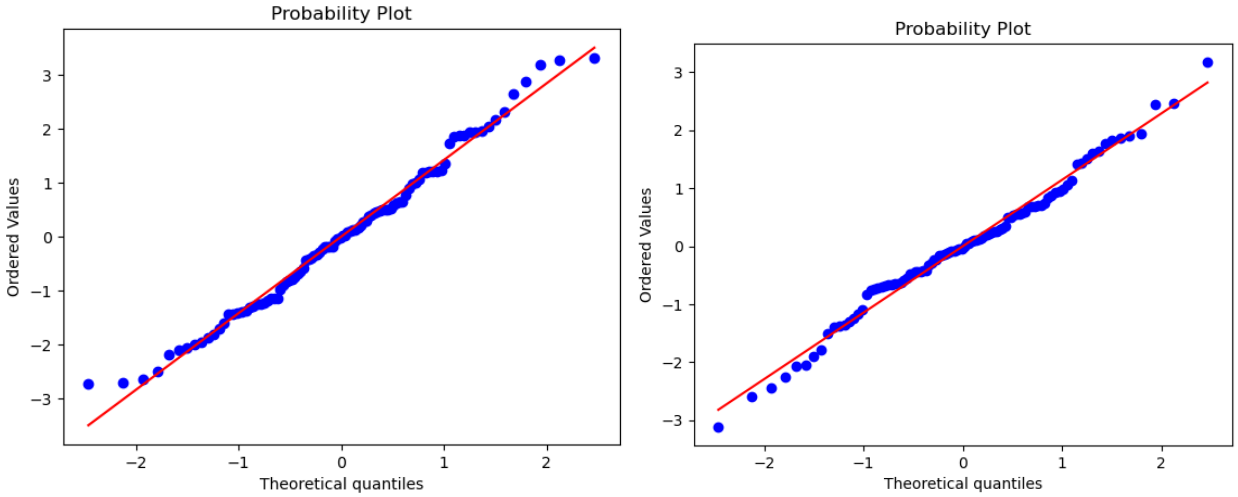
```
Explained variance ratio
 [0.494686   0.32310045 0.16783834 0.01437521]

Cumulative explained variance ratio
 [0.494686   0.81778645 0.98562479 1.        ]
```

We still need to keep the first 2 PCs to retain at least 80% of the variability.

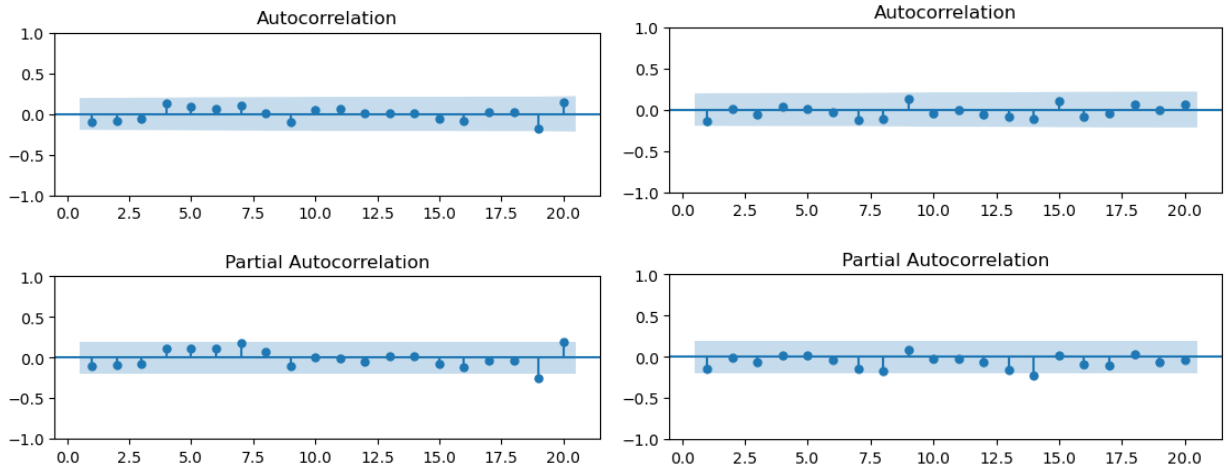
```
Eigenvalues
 [1.97874401 1.2924018 ]

Eigenvectors
 [[-0.61411239 -0.63211932  0.45467491  0.12869292]
 [ 0.36195886 -0.2773057   0.33655629 -0.82390363]]
```



Now check again the normality and the randomness of the new PC1 and PC2.



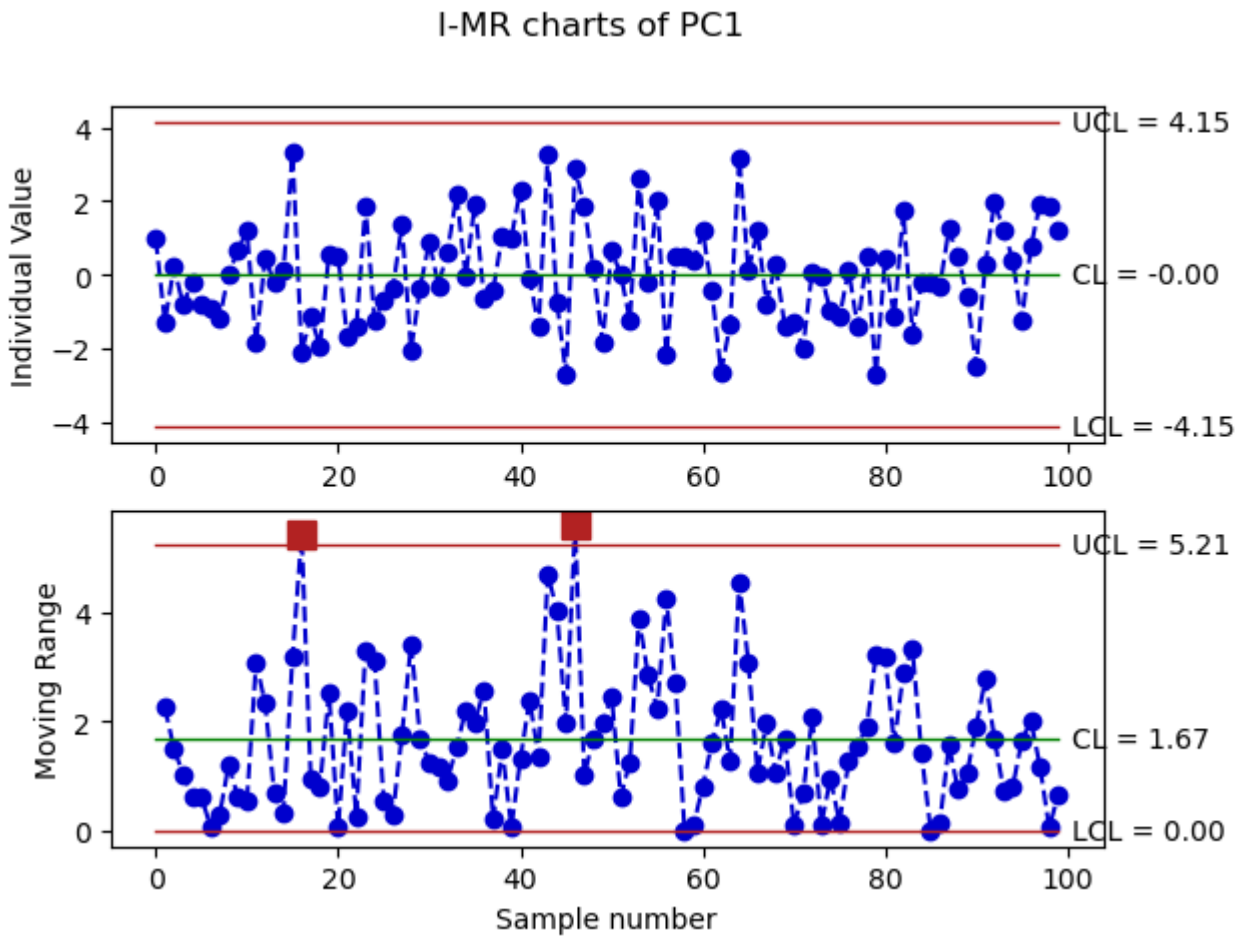
No violation of normality hypothesis (SW p-value 0.298 and 0.543).

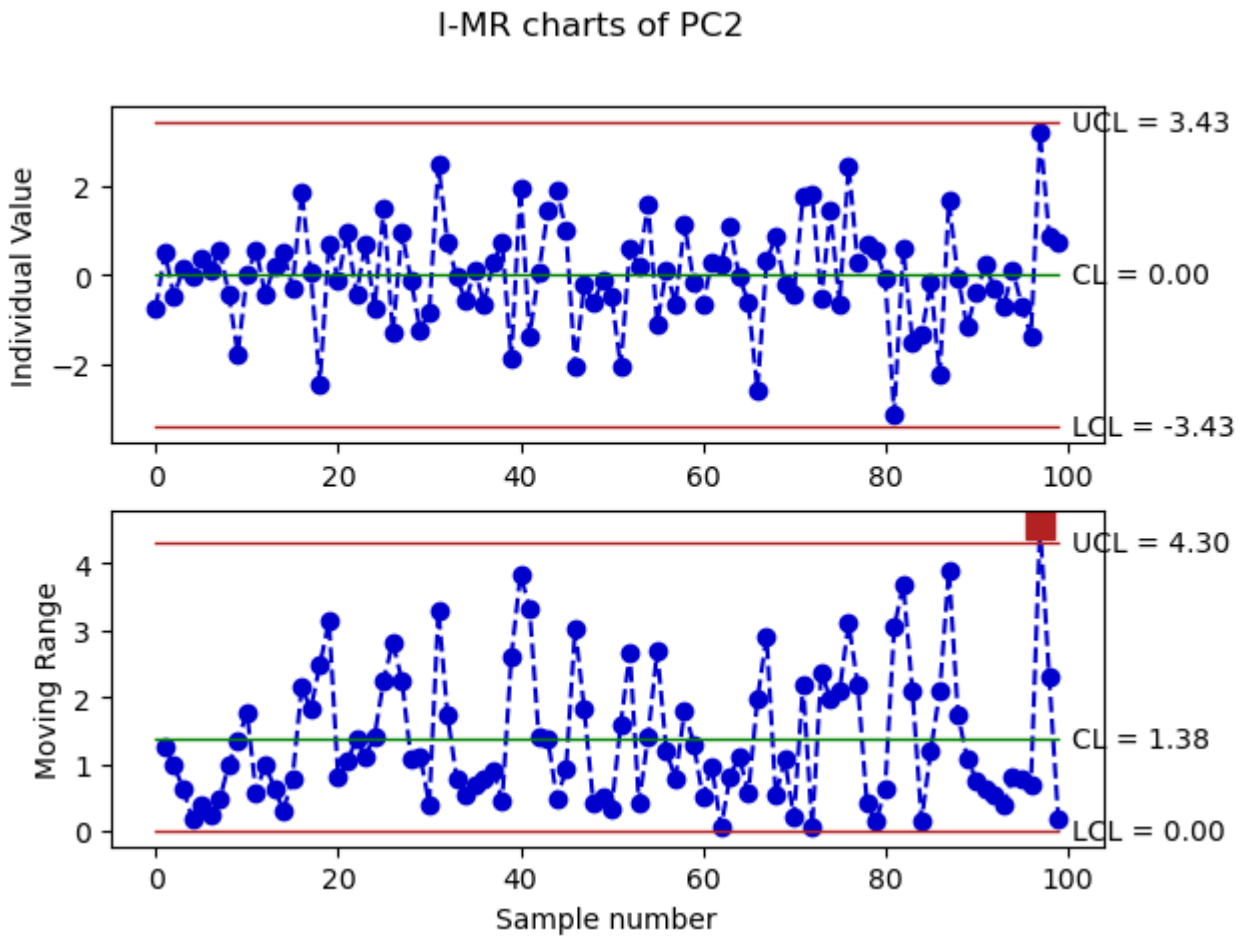


No violation of randomness (runs test p-value 0.421 0.158) or autocorrelation in the time series.

We can now design the two I-MR control charts on the two variables, but first we need to determine the value of K. Being $\alpha_{fam} = 0.01$ the family-wise type I error, and being the two PCs independent, we shall use the following correction: $\alpha = 1 - (1 - \alpha_{fam})^{1/2}$. Thus K = 2.806.

The control charts are the following:


I-MR charts of PC1
Individual Value
UCL = 4.15
CL = -0.00
LCL = -4.15
Moving Range
UCL = 5.21
CL = 1.67
LCL = 0.00
Sample number

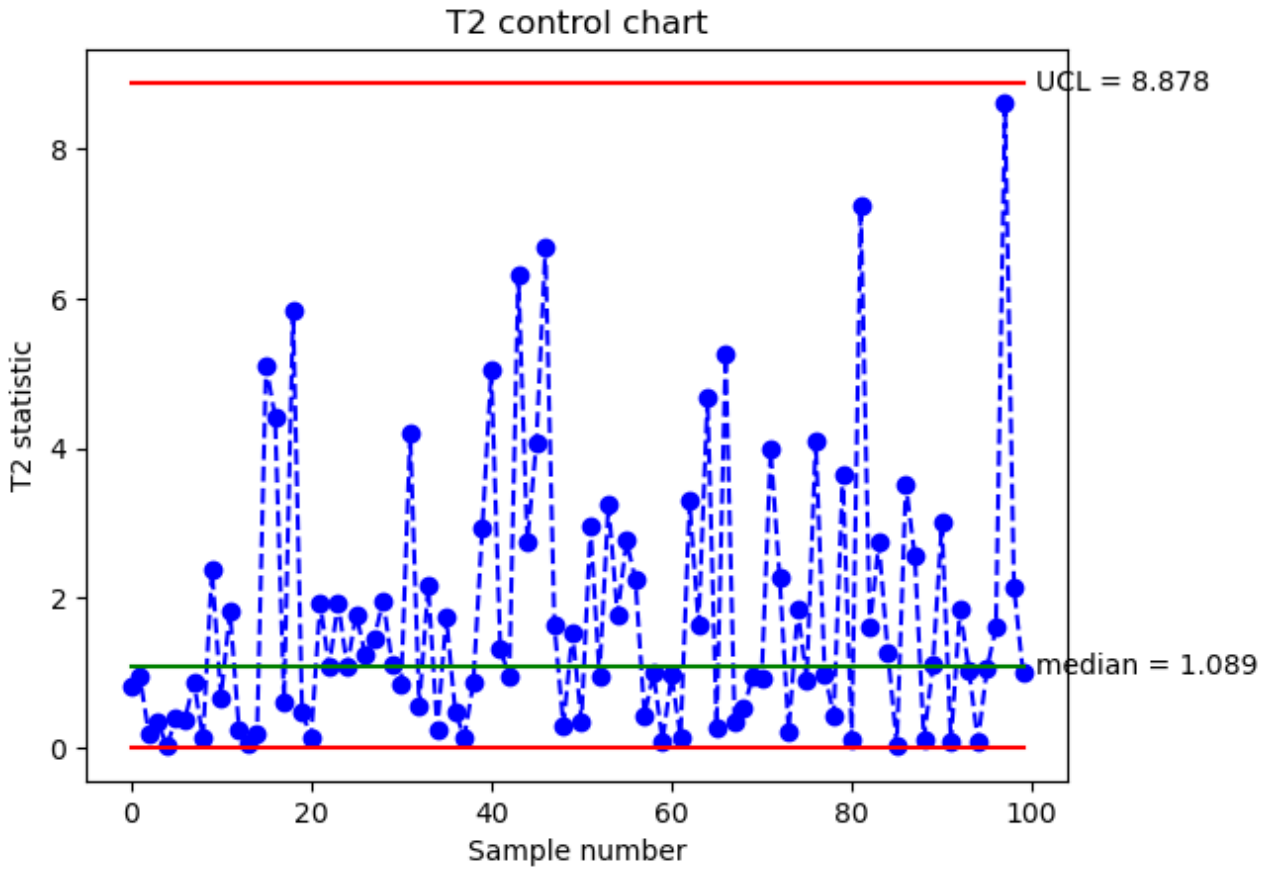


There are violations in the MR charts of both PC1 and PC2. Assuming there is NOT an assignable cause we shall keep them. If the violation is only in the MR chart, we can design the MR charts with prob. limits to check if the violation is due to the non-normality of the MR statistic.

Alternatively to designing 2 I-MR CC, we can design one multivariate (T2) control chart after estimating the mean and the variance-covariance of the scores using the short range estimator S2:

```
The short range estimator is:

          PC1       PC2
PC1  2.161556  0.005111
PC2  0.005111  1.465010
```

In this case, since we are designing only one chart, we don't need to apply any correction to alpha which is set at 1%.



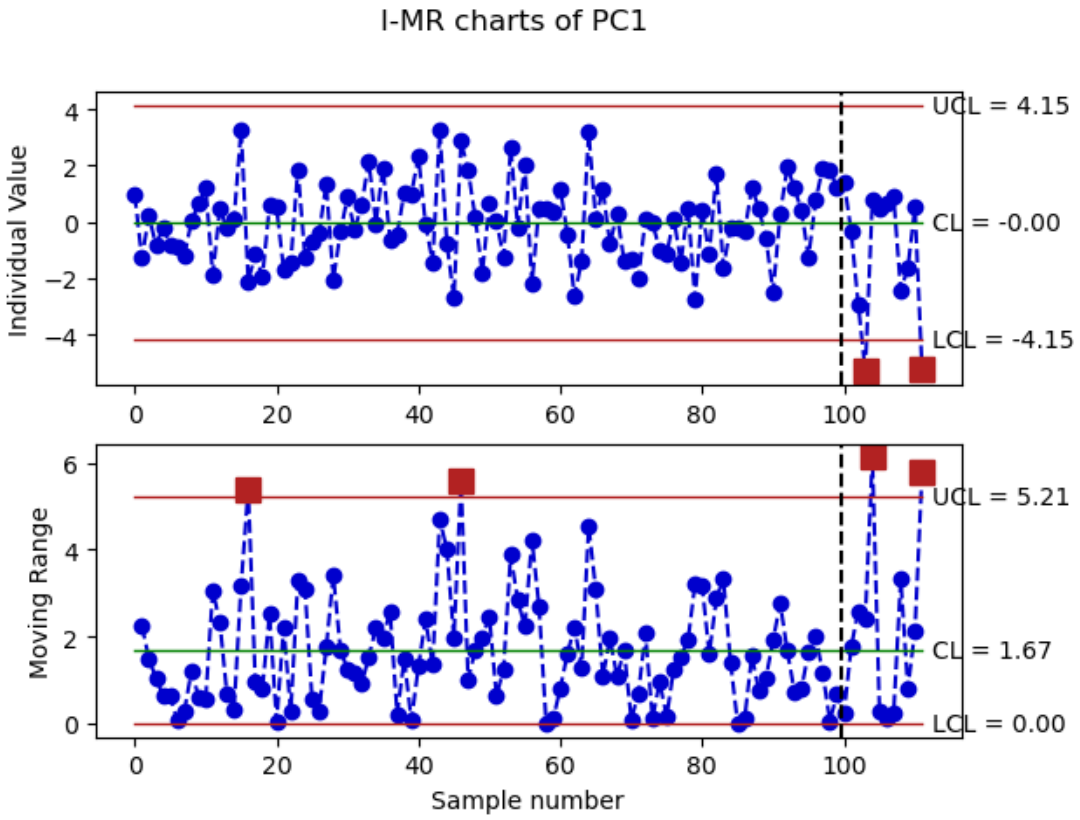
The T2 control chart does not indicate any OOC.

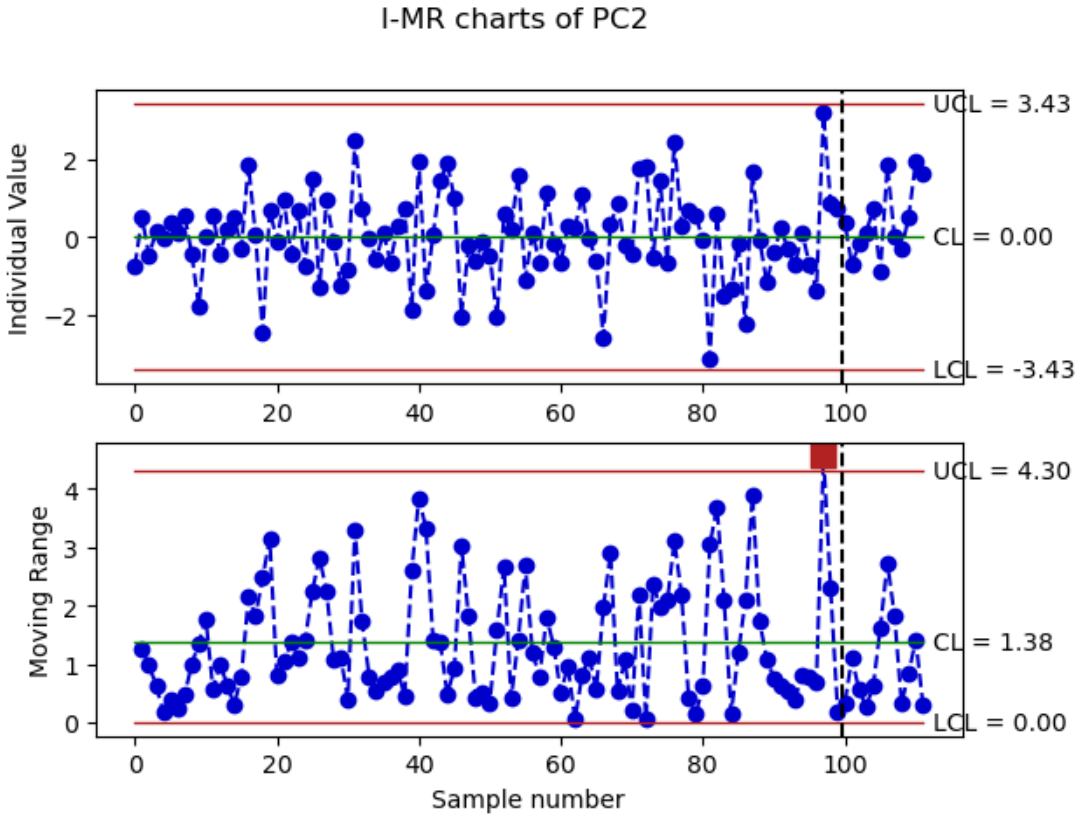
3)

We need to apply the same transformations to the new dataset, i.e.:

1. Box-Cox on x4 using the previously estimated lambda.
2. Standardize using the previously estimated sample mean and standard deviation.
3. Transform the new data in the PC space spanned by PCs identified in Phase 1.

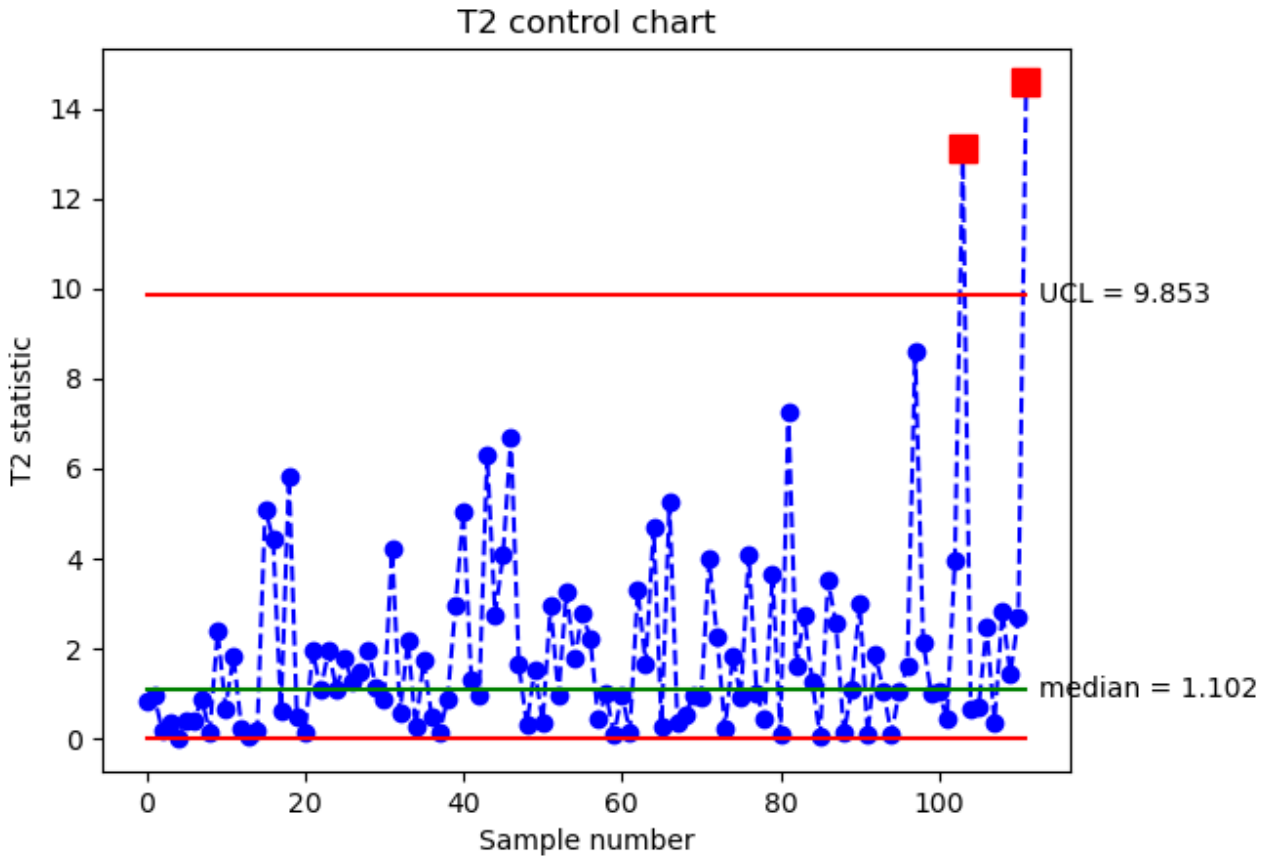
To test the new observations we can use the two univariate control charts or the T2 cc.






```
Out of control points for the I chart on PC1: [103 111]
Out of control points for the MR chart on PC1: [ 16  46 104 111]

Out of control points for the I chart on PC2: []
Out of control points for the MR chart on PC2: [97]
```

Or use the T2 control chart.




```
Out of control points for the T2 chart: [103 111]
```

Exercise 3 solution

Question 1

Answer: a

Explanation: In the simple linear regression for the T-test statistic (T) of the slope, we have that $T^2 = F$ (where F is the overall F-test). Thus $T = \pm\sqrt{F} = \pm\sqrt{16.81} = \pm 4.1$

For the T-test we also have:

$$T = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$$

and from the fitted model $\hat{\beta}_1 = -1.33$, therefore the T-test statistic will have a negative sign and so we have $T = -4.1$.

Question 2

Answer: b

Explanation: In the simple linear regression we have that $R^2 = \big(r(X,Y)\big)^2$. Thus $r(X,Y) = \pm\sqrt{R^2} = \pm\sqrt{0.81} = \pm 0.9$.

Furthermore, we know that:

$$r(X,Y) = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \quad \text{and} \quad b_1 = \frac{S_{XY}}{S_{XX}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}\sqrt{\frac{S_{YY}}{S_{XX}}} = r(X,Y)\sqrt{\frac{S_{YY}}{S_{XX}}}$$

thus, the estimated slope and the correlation coefficient will have the same sign. Since in this problem $b_1 = -2.84$, the correlation will be negative and therefore: $r(X,Y) = -0.9$ is the correct answer.

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

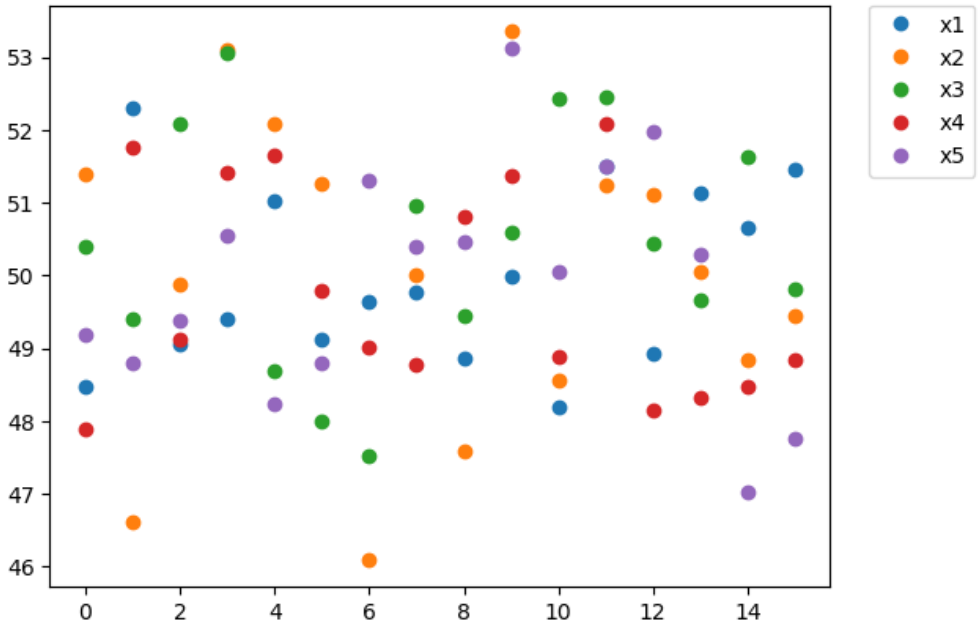
In a data set with five variables (X_1, X_2, …, X_5), where correlation $|r(X_i, X_j)| < 1$, for $i \neq j$, we applied PCA. Which of the following can represent the ordered proportion of variance explained by the first four components?

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

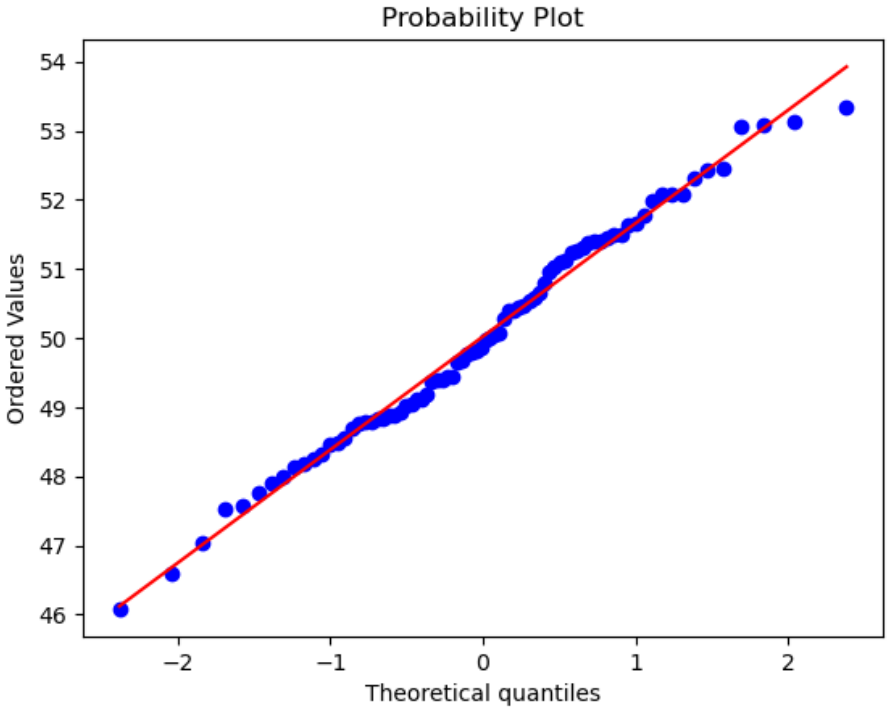
1)

Import the temperature data and plot them.



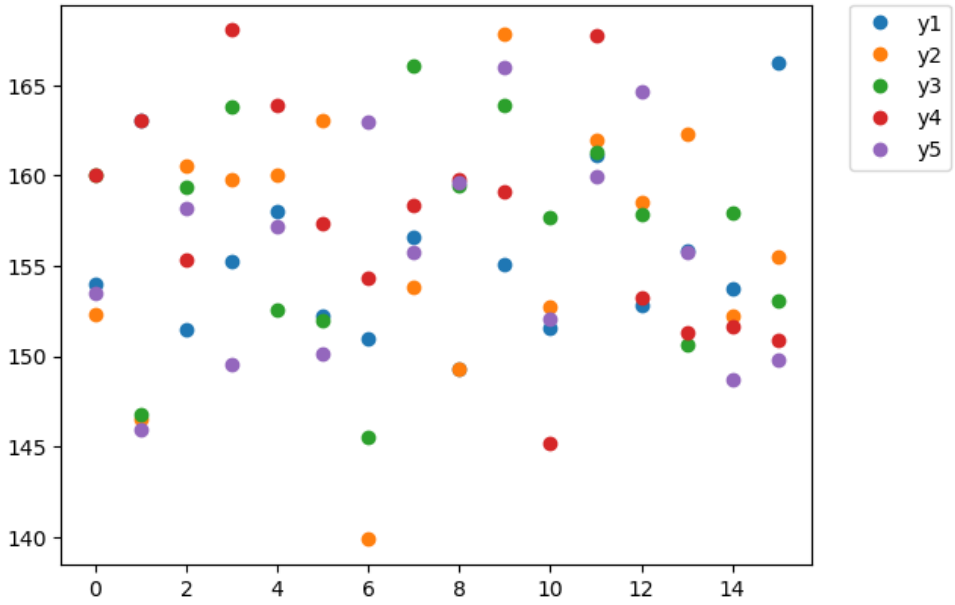
No information is given out about the time order within each sample and we need to rely on a qualitative assessment to evaluate randomness. No strange patterns appear in the data and the process seems stationary. I have no reason to reject the randomness hypothesis.

Let's check the normality.



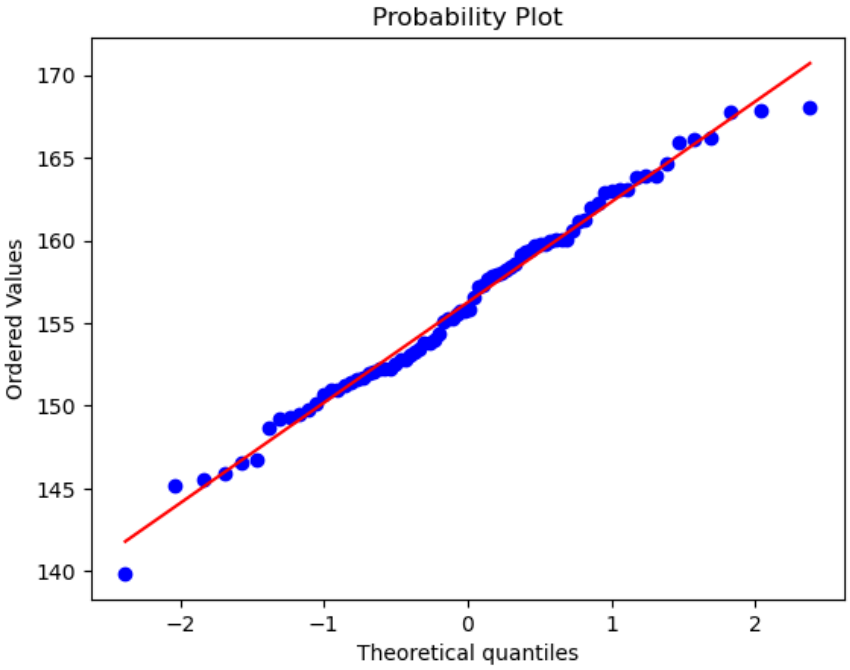
We cannot reject the normality hypothesis (SW p-value = 0.611).

Let's import the pressure data as well.



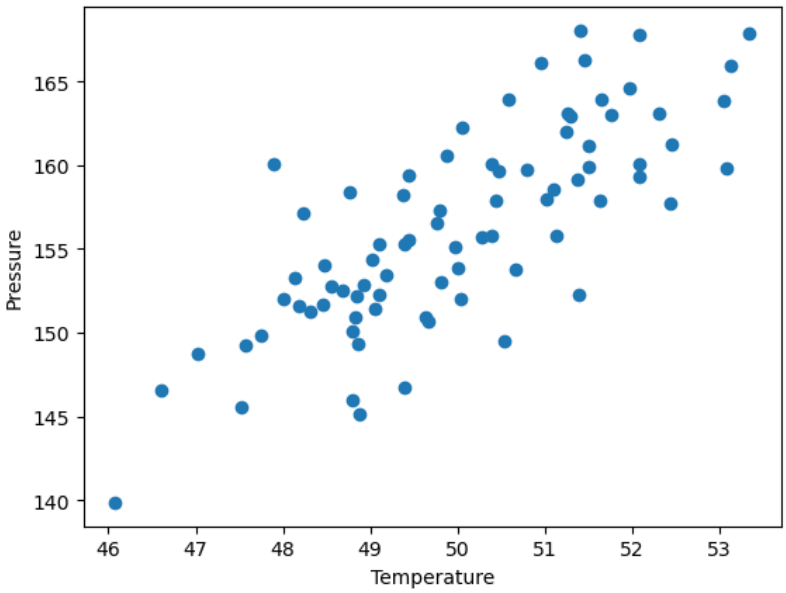
No trend in the data and observations appear to be random.

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.



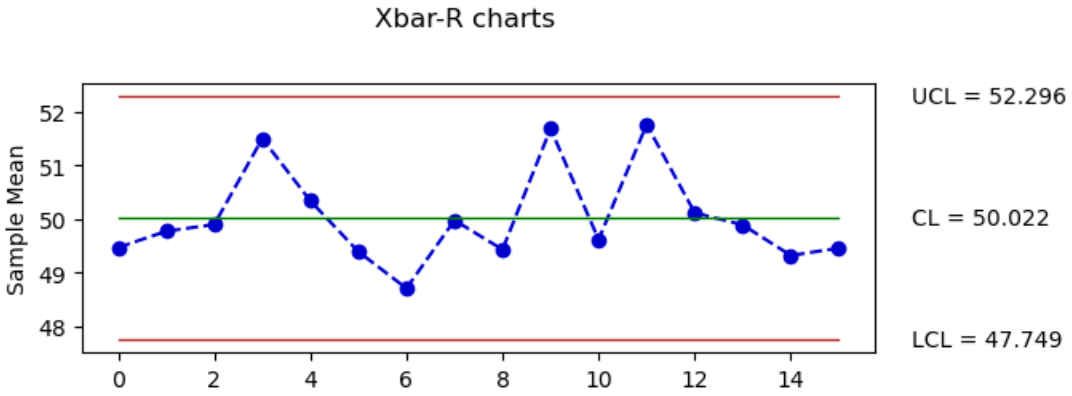
MECH ENG STUDENTS:

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.

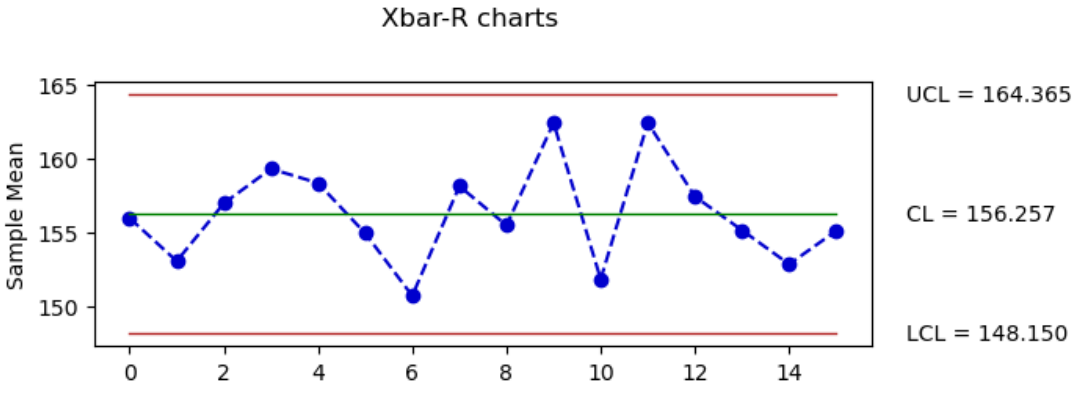
alpha = alpha_fam/2 = 0.001

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.

non importa se le osservazioni nei sample provengono da differenti locations, una volta che abbiamo assunto la NID per ognuna e la MNID facciamo un T2 control chart nel sium

First, we compute the grand mean (Xbarbar):

```
temp      50.022375
press    156.257500
```

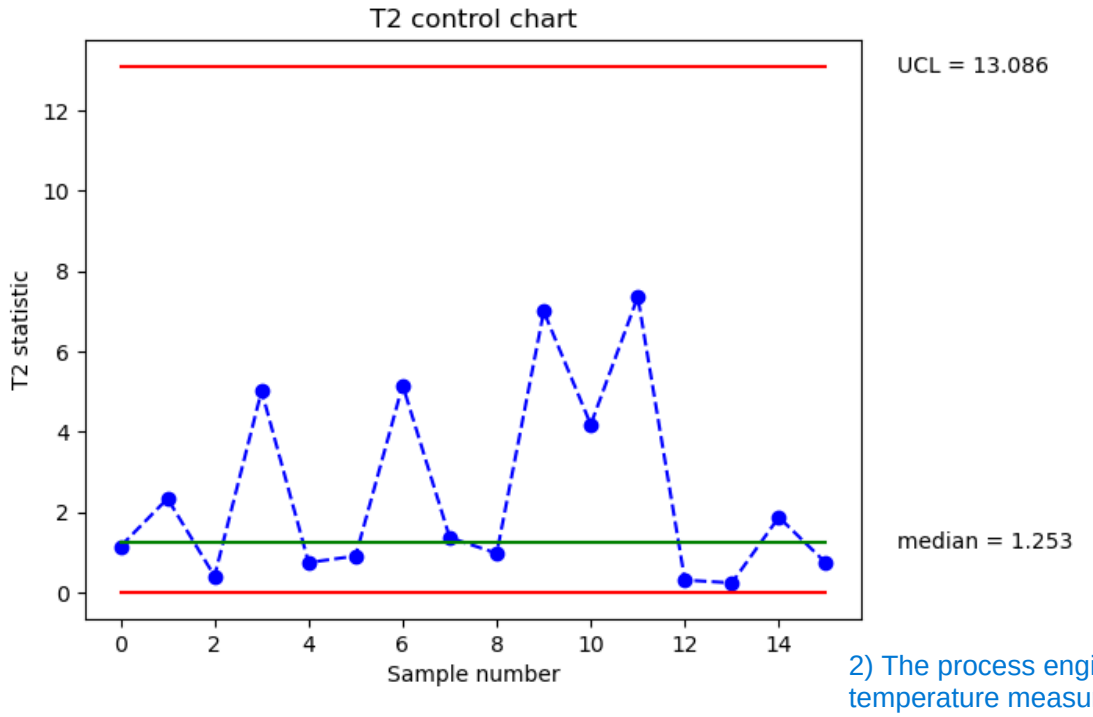
Then we compute the variance-covariance matrix (S):

```
            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) 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).

2)

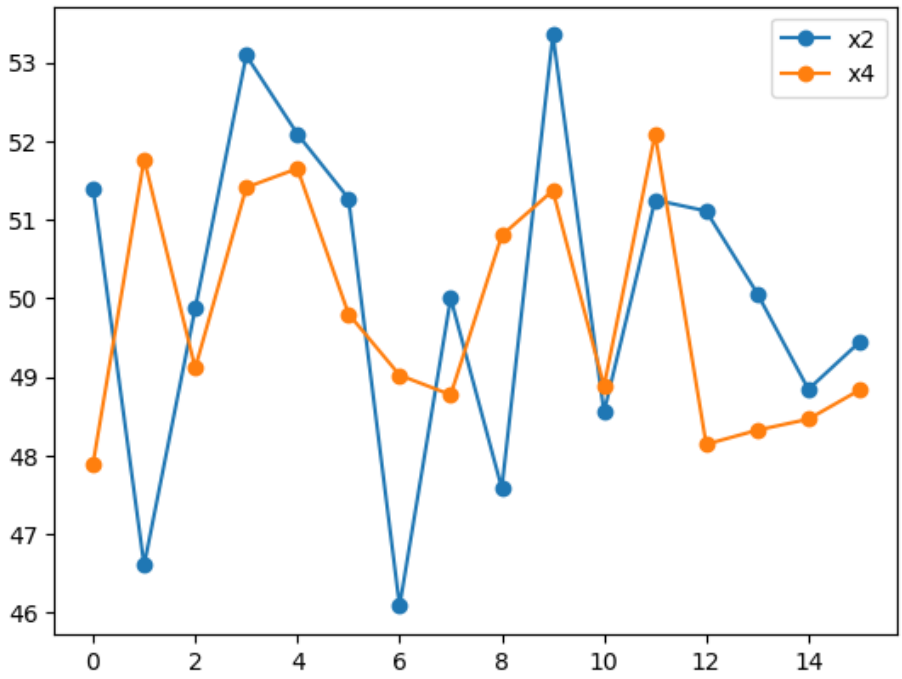
We can perform a paired t-test.

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

...pairwisely because they are referring to the same "piece"

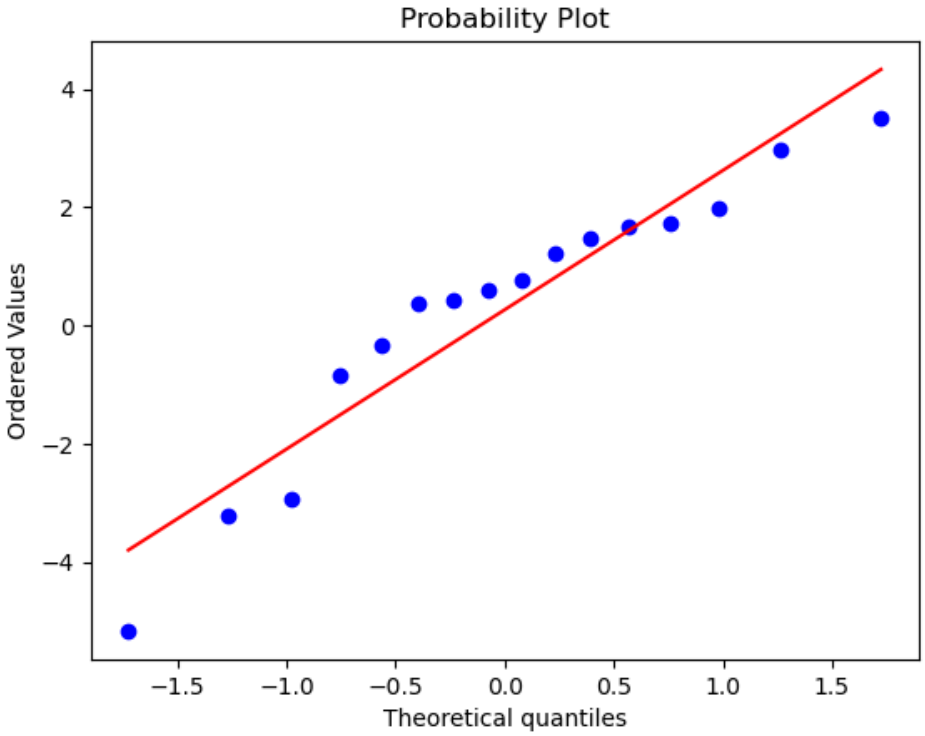
↳ T_2 and T_4 cannot be considered indipendent for some reason even if we have assumed T_1, T_2, T_3, T_4 and T_5 indipendent before

in the DOT PLOT we have "concluded" RANDOMNESS basically



Process seems stationary with no significant deviations.

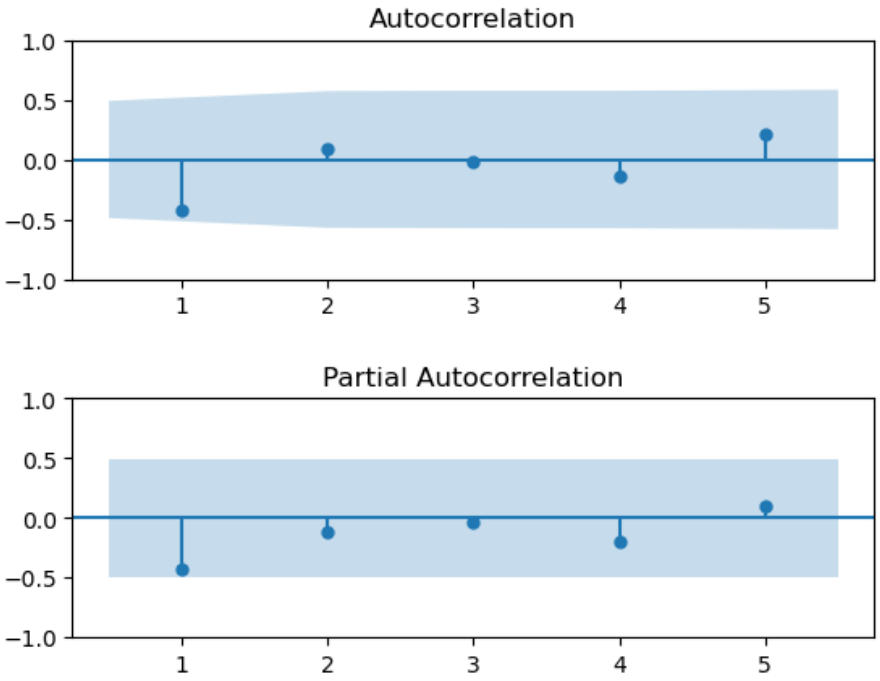
Now we can compute the difference between the paired data (d = x2 – x4) and check for normality and randomness of the difference variable.



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
```

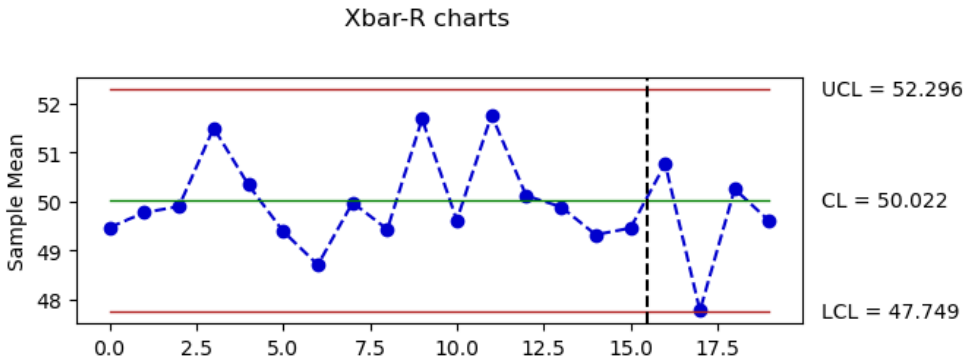
Based on the result we got from the paired t-test, there is no statistical evidence to reject H0.

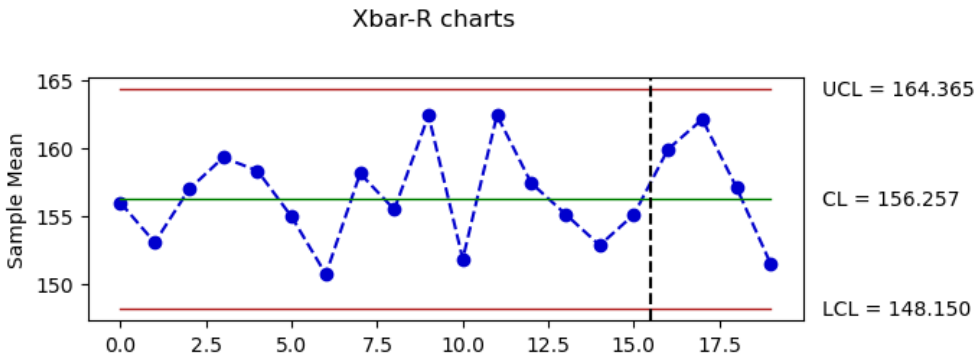
3)

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.

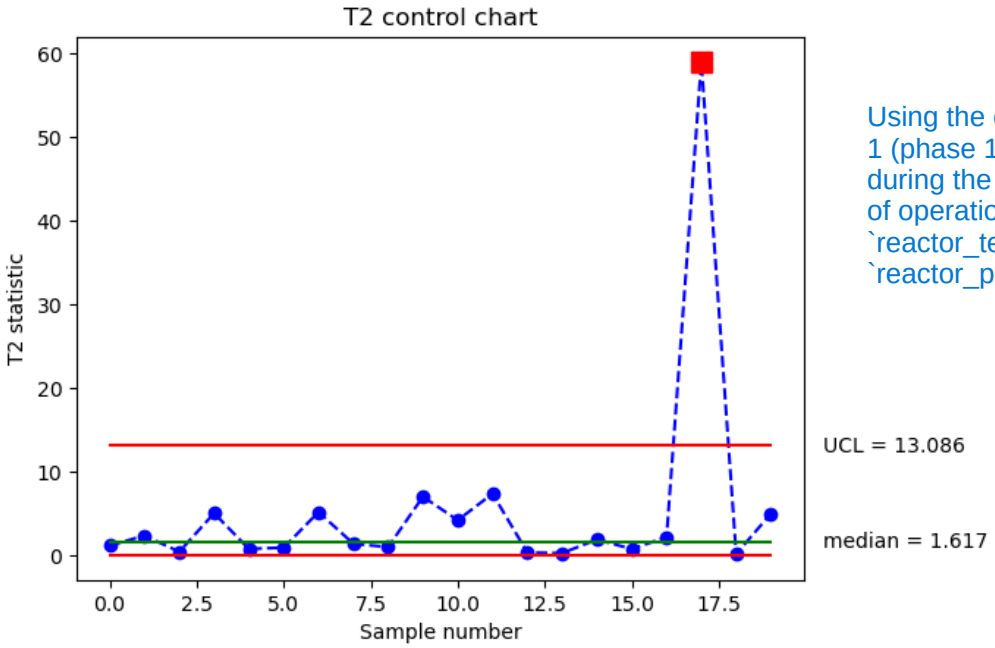




All the new observations appear to be in-control.

OTHER STUDENTS

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



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.

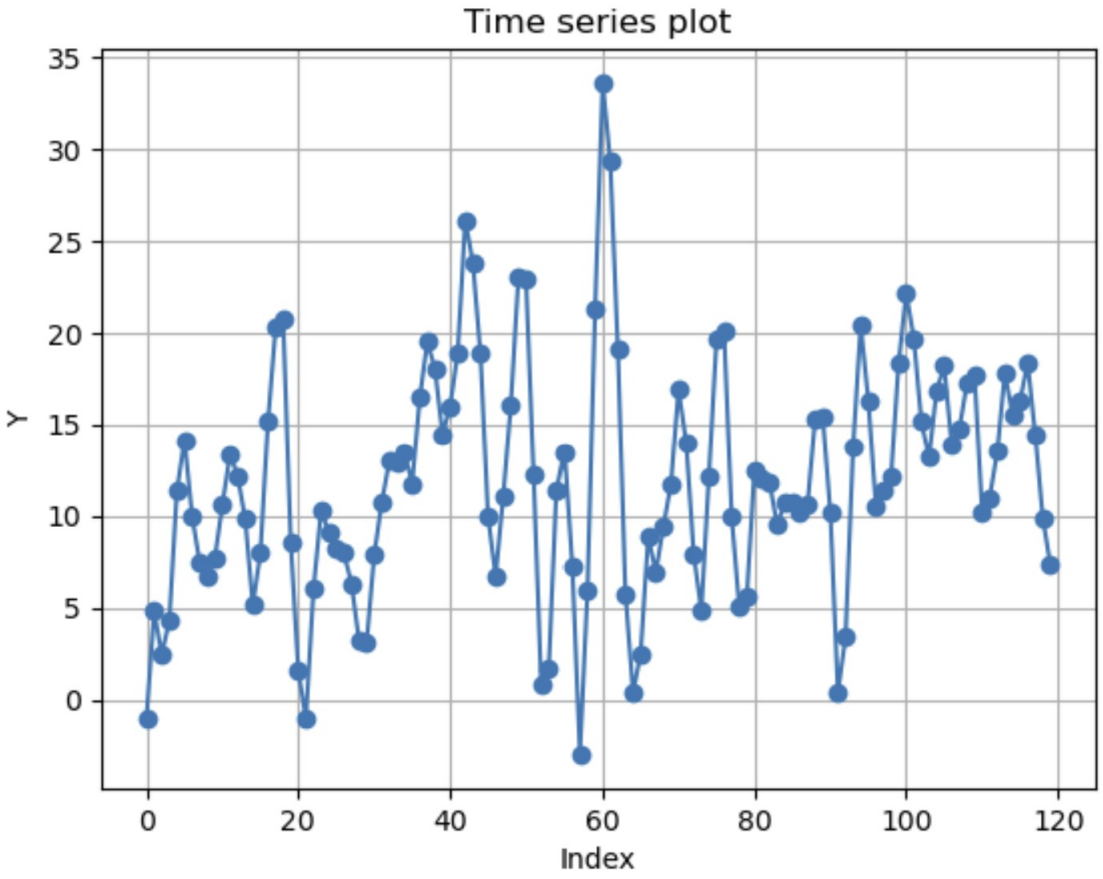
The second sample of phase 2 data is out-of-control. This could only be captured thanks to the multi-variate approach which takes into account the correlation structure between the two random variables.

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.

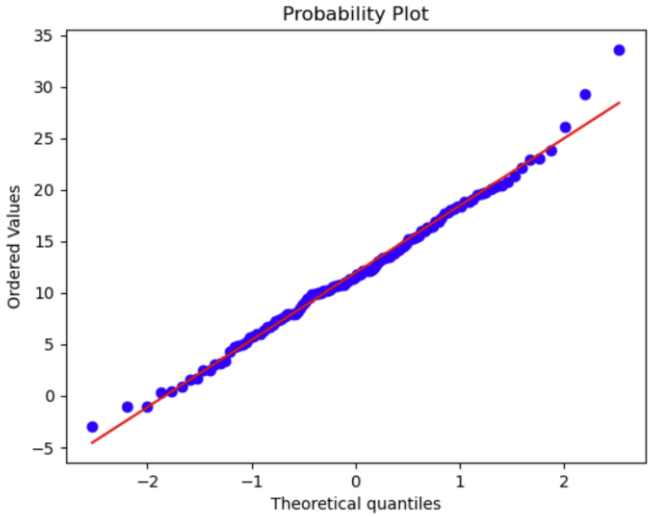
Exercise 2 solution

1)

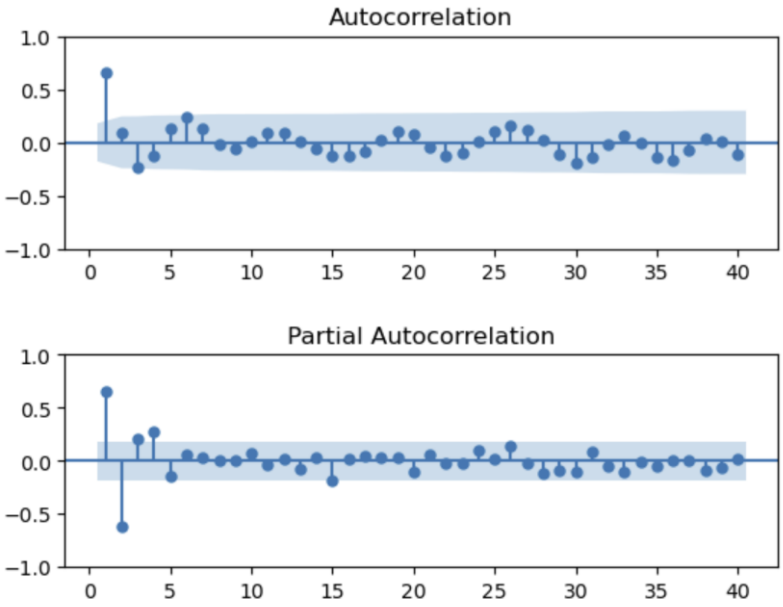
Data snooping: the time series of the performance indicator is the following:



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

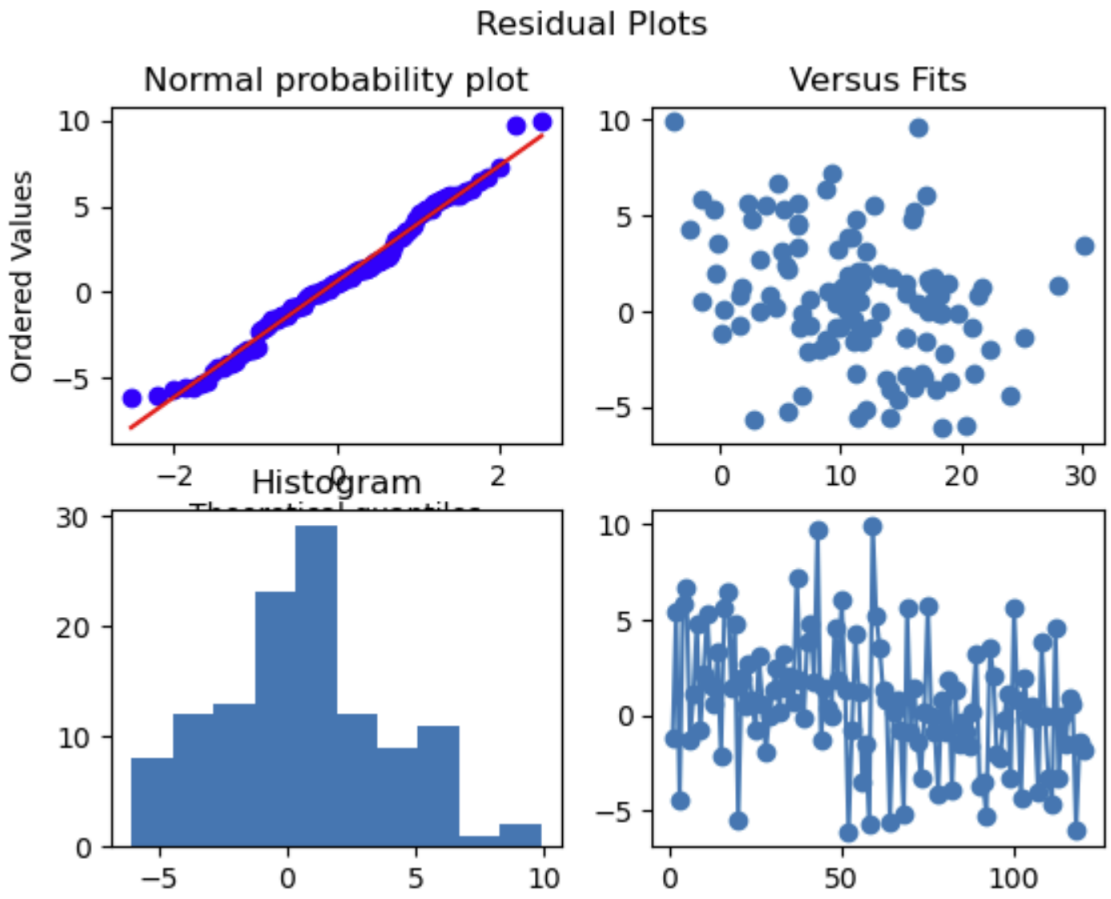


LBQ test with L = 10:

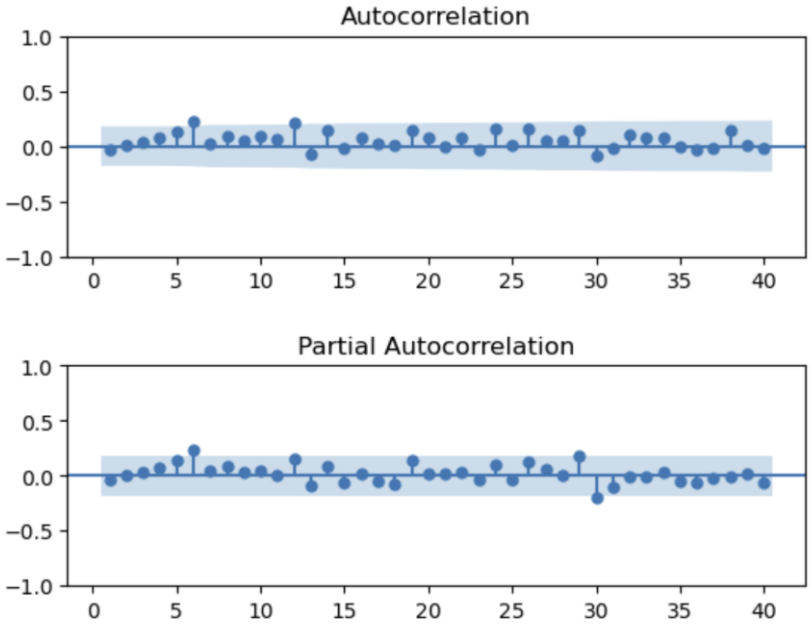
- Q0_LBQ = 75.172
- p-value = 0.000

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



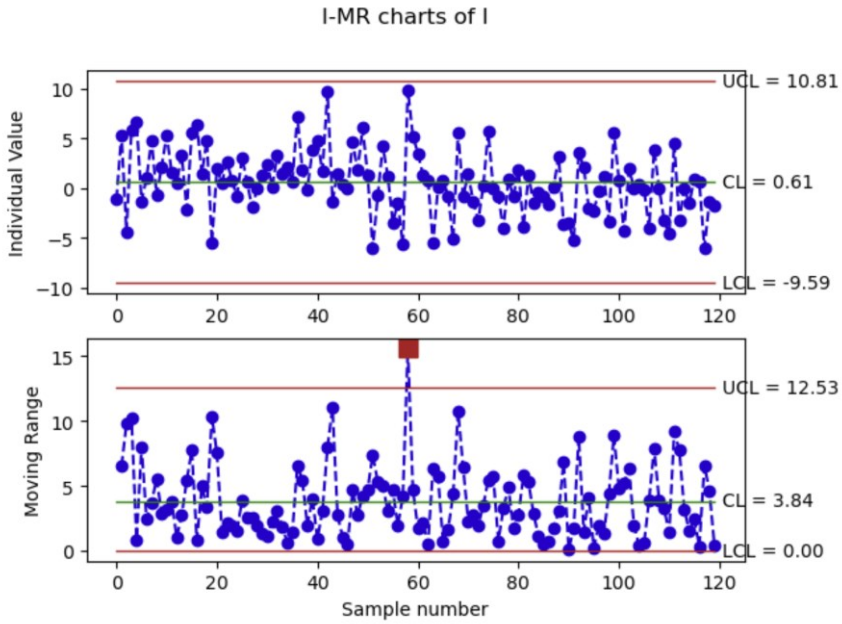
LQB test with L = 10:

- Q0_LBQ = 12.235
- p-value = 0.269

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

2) 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.

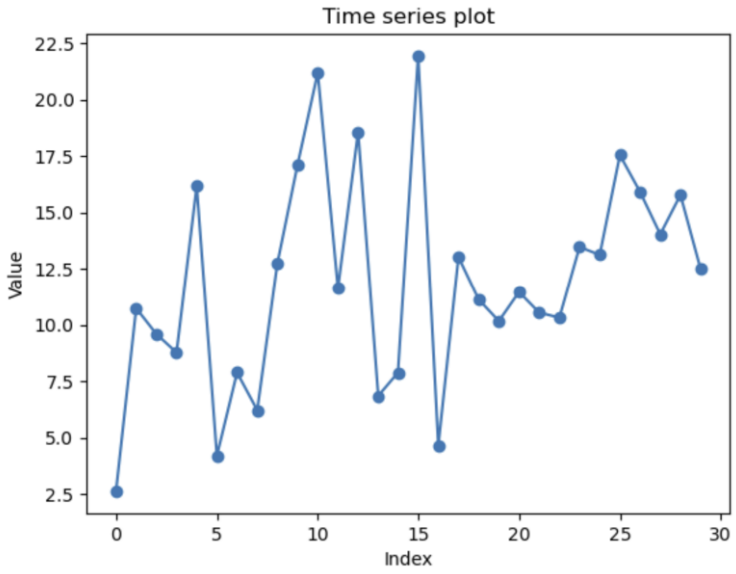
The special cause control chart applied to the model residuals with K = 3 is the following:



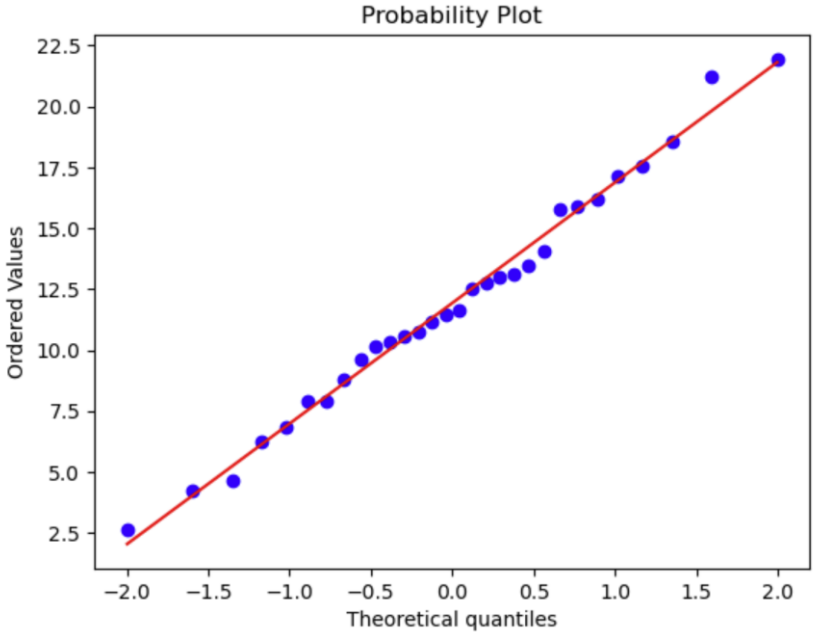
There is only one point (observation 58) that violates the upper control limit in the MR chart. Since no assignable cause is present, no further action is needed and the control chart design is over. Note that this violation might be also caused by an intrinsic violation of the normality assumption in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

3) 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).

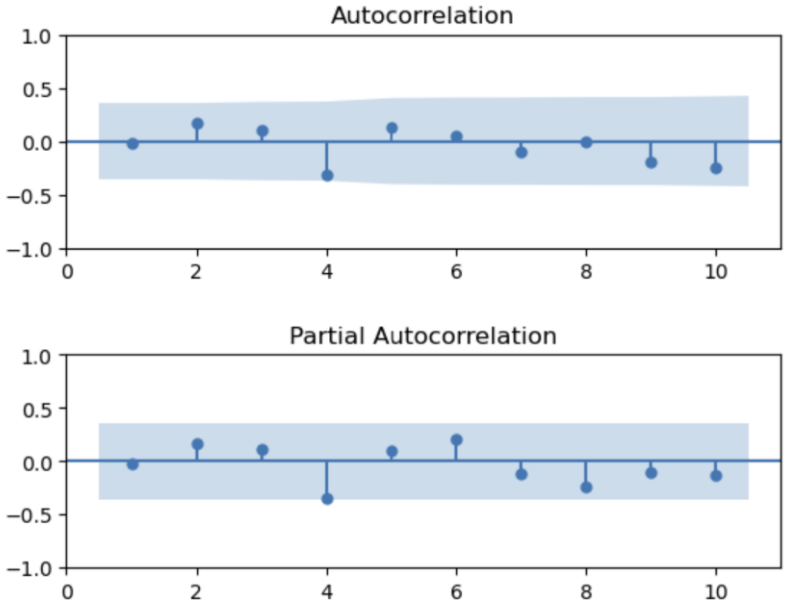
By applying a batching operation with batch size = 4, we get the following time series:



After batching data are normal and independent:



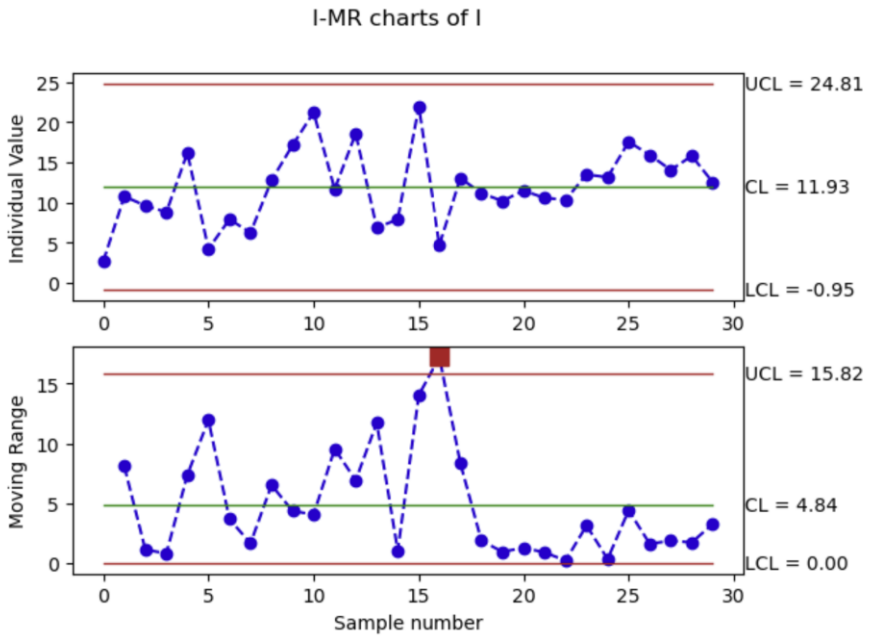
Shapiro-Wilk test p-value = 0.966



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.

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

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., $|r(X_i, X_j)| < 1$, 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).

QUALITY DATA ANALYSIS

09/06/2023

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
- **Multichance students should skip: point b) in Exercise 1, point a) in Exercise 2**

Exercise 1 (15 points)

The concentration of a contaminant (measured in ppm) in the production of synthetic rubber is monitored over time. '230609_ex1.csv' contains the measurements collected in 50 consecutive samples.

a) Being known that a negative value is the result of a temporary miscalibration of the measuring device, fit a suitable model to these data;
b) Based on the result of point a), estimate the 95% prediction interval for the contaminant concentration in the next sample.
c) Based on the result of point a), design an appropriate control chart for these data with $ARL_0 = 250$.
d) From historical data, it is known that the most appropriate model for this process yielded a standard deviation of residuals equal to $\sigma_\varepsilon = 2.5$. Determine, with a statistical test, if the model fitted at point a) is such that the standard deviation of residuals is greater than this value (report also the p-value of the test). Discuss the result.

Exercise 2 (15 points)

A company produces aluminum laminates. The quality control department has recently introduced a statistical monitoring tool to keep under control the planarity of the laminates. It consists of an $\bar{X}$ control chart designed such that the number of samples before a false alarm is equal to 250.

a) Estimate and draw the curves of ARL_1 as a function of the mean shift δ expressed in standard deviation units with a sample size $n = 4$ and $n = 8$, respectively (show the two curves for $\delta \in [0\ 2]$ and report the ARL_1 values for $\delta = 1$ and $\delta = 2$).
b) Estimate and draw the curves of ARL_1 as a function of the sample size n for two values of the shift, $\delta = 1$ and $\delta = 2$, where δ is expressed in standard deviation units (show the two curves for $n \in [2\ 20]$ and report the ARL_1 values for $n = 3$ and $n = 6$).
c) The head of the quality control department is interested in selecting an optimal sample size n to minimize the lack of quality costs in the presence of a mean shift equal to $\delta = 2$ standard deviation units. Knowing that samples are gathered every 4 hours, the cost of planarity measurements for each laminate is $C_1 = 2$ €and an extra cost equal to $C_2 = 15$ € is due for each hour spent in the out-of-control state, determine the optimal sample size that minimizes the overall expected costs (assume the cost of the process in its in-control state as a reference baseline). Discuss the results.

Exercise 3 (3 points)

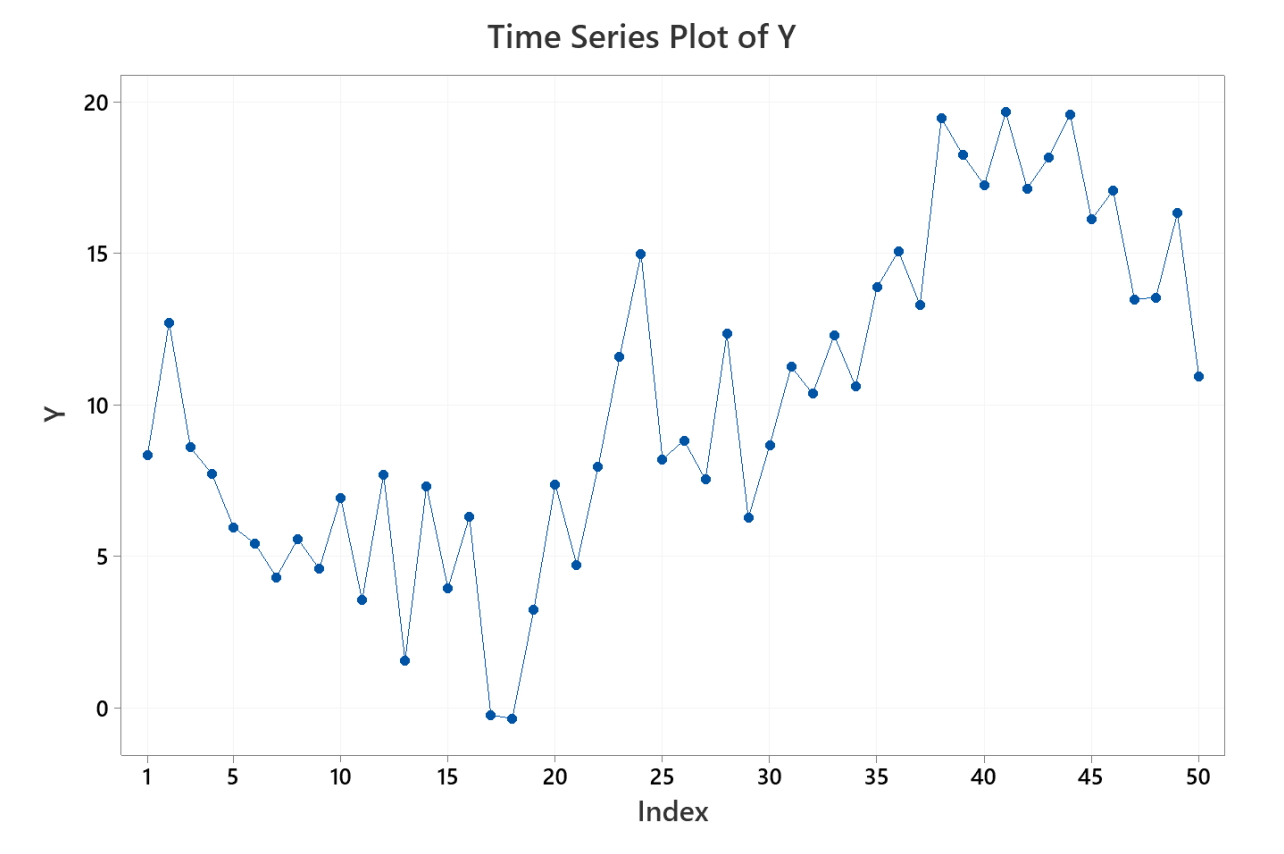
A quality characteristic X_t follows a stationary AR(1) model $X_t = \xi + \phi_1 X_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ with positive autocorrelation coefficient and known σ_ε^2. Let $E(X_t) = \mu$ and $V(X_t) = \sigma^2$. Compute the expressions of ξ and ϕ_1 as functions of μ, σ^2 and σ_ε^2.

Solutions

Exercise 1

a)

Time series plot of the temperature series:



It is present a meandering pattern. Negative values were observed in sample 17 and 18.

Runs test: null hypothesis is not accepted:

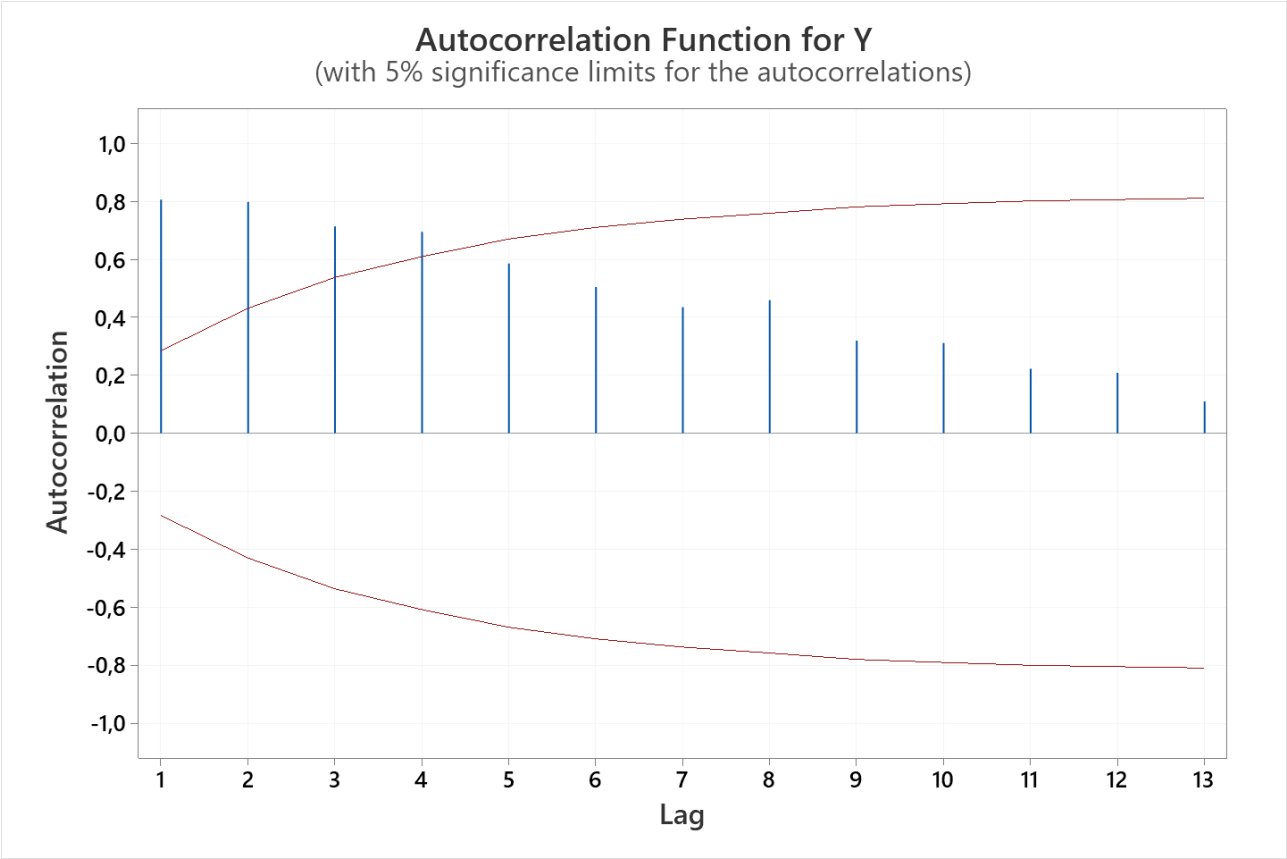
Test

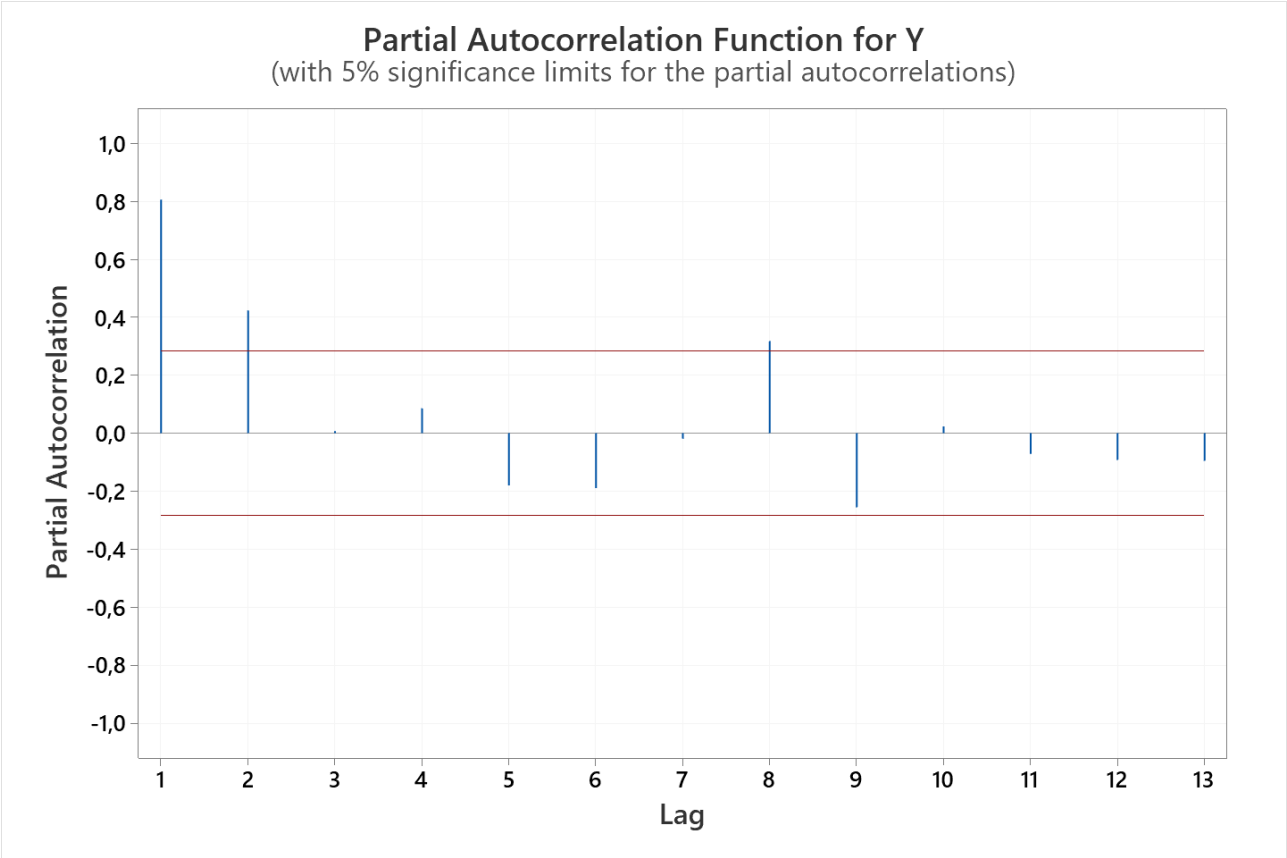
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs

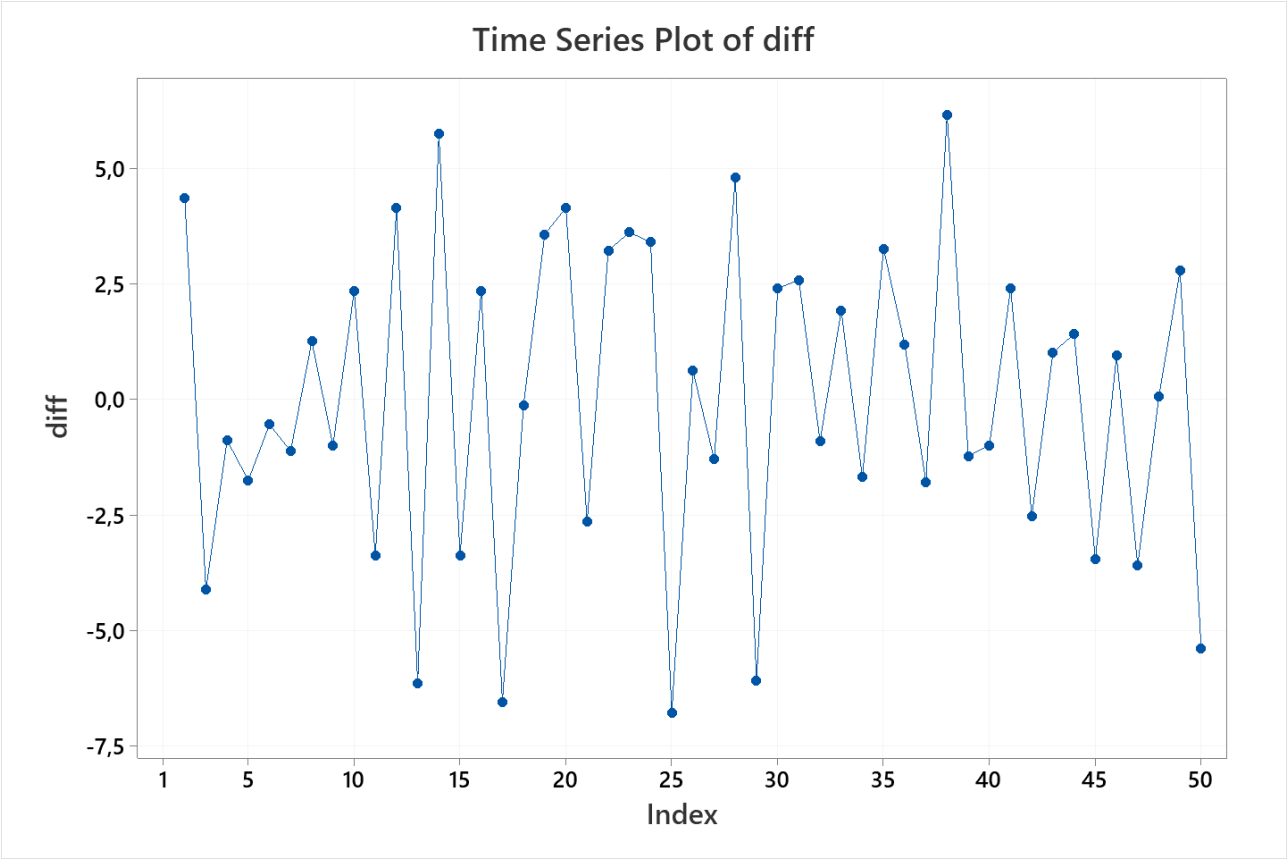
Observed	Expected	P-Value
8	25,96	0,000

Sample autocorrelation and partial autocorrelation functions:

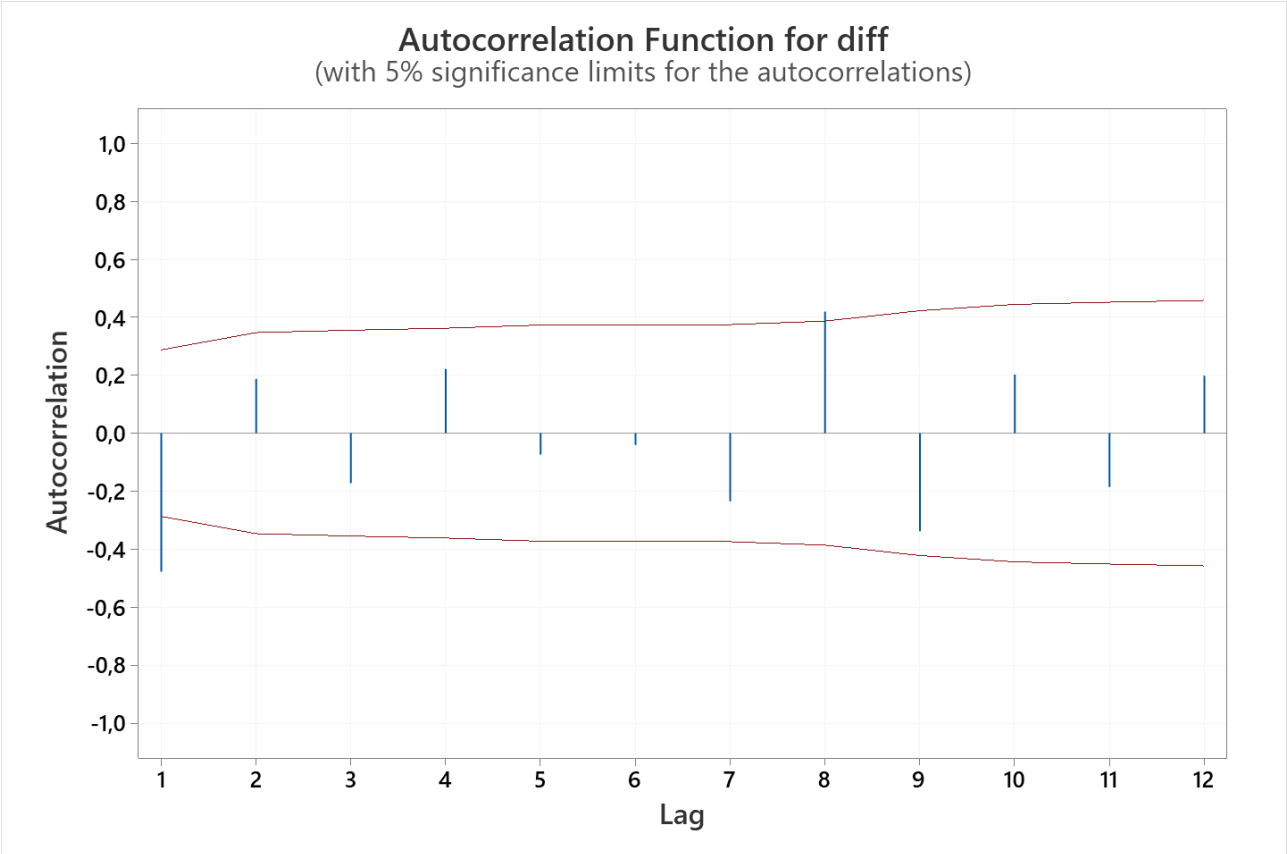


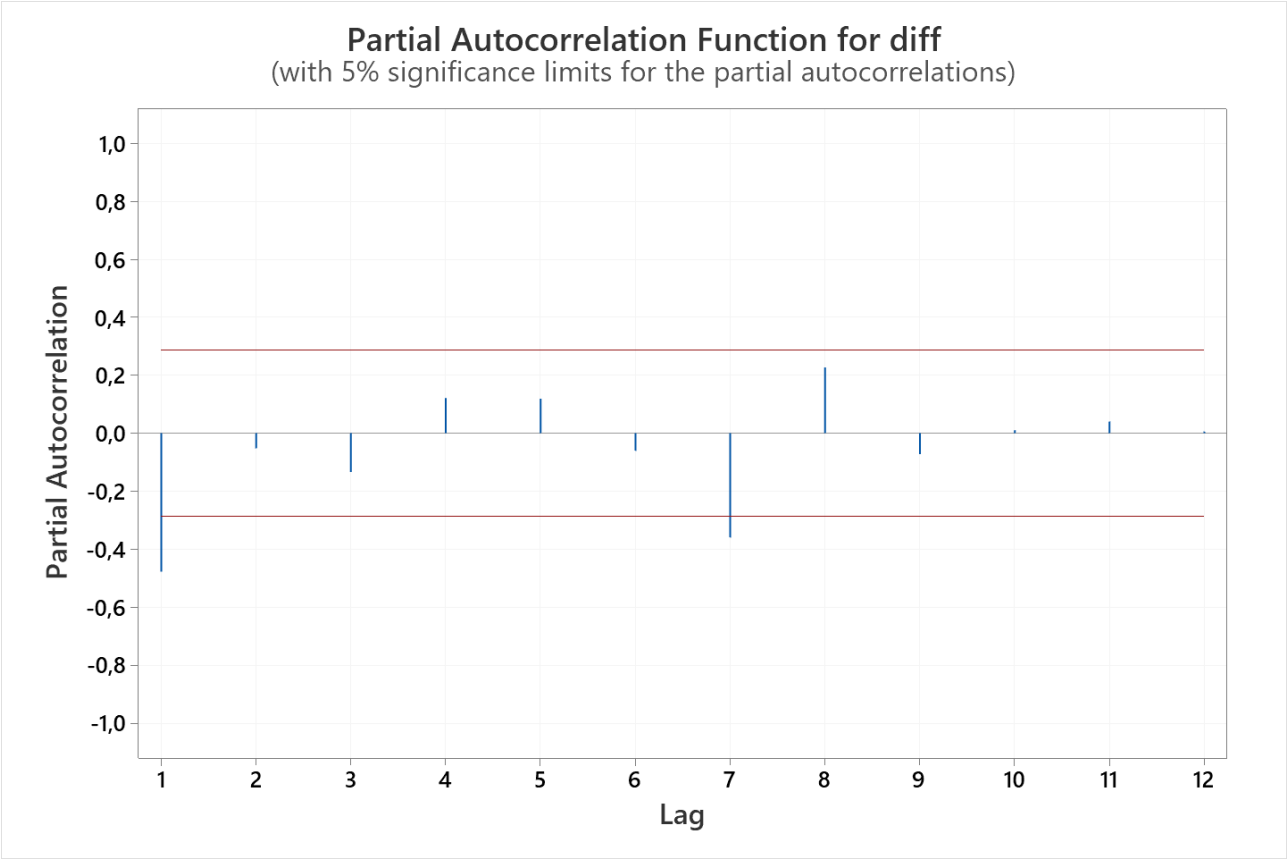


A slow decay of the SACF is present, which suggests a non-stationarity of the process. By differencing the timeseries we get:



The SACF and SPACF of the data after the differencing operation are the following:





A suitable model for the temperature time series is therefore an ARIMA(1,1,0). However, we should keep in mind that two negative values are present, caused by a temporary miscalibration of the sensor. Thus, a dummy variable that is equal to 1 for these two samples and 0 for all other samples can be included in the model.

WORKSHEET 1

Regression Analysis: diff versus AR1; dummy

Method

Categorical predictor coding	(1; 0)
Rows unused	2

Regression Equation

dummy	
0	diff = 0,251 - 0,546 AR1
1	diff = -4,47 - 0,546 AR1

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	0,251	0,413	(-0,581; 1,083)	0,61	0,547	
AR1	-0,546	0,125	(-0,797; -0,295)	-4,38	0,000	1,02
dummy						
1	-4,72	2,04	(-8,83; -0,62)	-2,32	0,025	1,02

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2,79333	32,91%	29,92%	387,706	25,92%	240,66	247,22

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	172,21	32,91%	172,21	86,106	11,04	0,000
AR1	1	130,36	24,91%	149,58	149,580	19,17	0,000
dummy	1	41,85	8,00%	41,85	41,854	5,36	0,025
Error	45	351,12	67,09%	351,12	7,803		
Total	47	523,33	100,00%				

The constant term is not significant, thus we may remove it:

WORKSHEET 1

Regression Analysis: diff versus AR1; dummy

Method

Categorical predictor coding	(1; 0)
Rows unused	2

Regression Equation

dummy	
0	diff = 0,0 - 0,540 AR1
1	diff = -4,46 - 0,540 AR1

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
AR1	-0,540	0,123	(-0,789; -0,292)	-4,37	0,000	1,02
dummy						
1	-4,46	1,98	(-8,44; -0,48)	-2,25	0,029	1,02

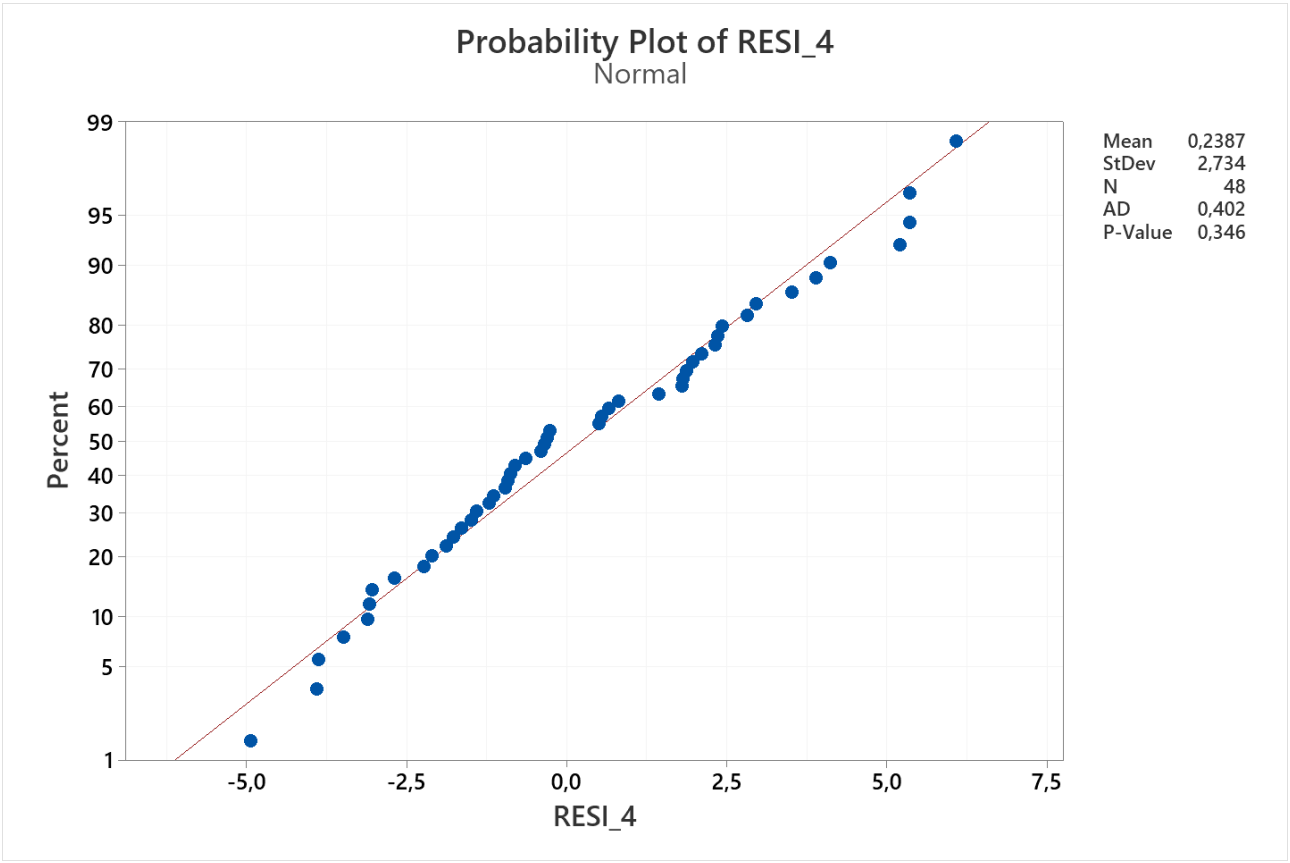
Model Summary

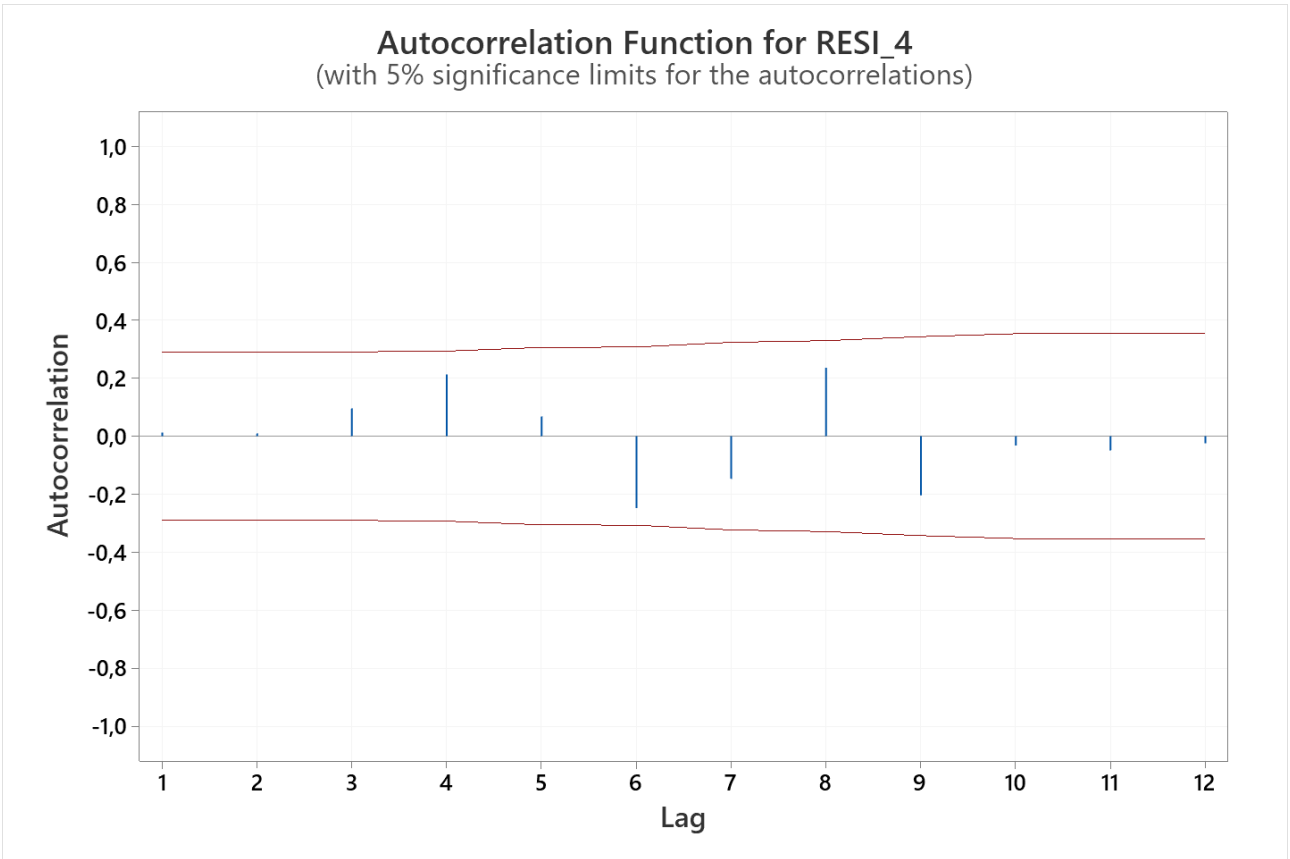
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2,77408	32,37%	29,43%	374,471	28,45%	238,67	243,74

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	169,41	32,37%	169,41	84,703	11,01	0,000
AR1	1	130,32	24,90%	147,23	147,227	19,13	0,000
dummy	1	39,09	7,47%	39,09	39,087	5,08	0,029
Error	46	353,99	67,63%	353,99	7,696		
Total	48	523,40	100,00%				

Check of residuals:


Probability Plot of RESI_4
Normal
Mean 0,2387
StDev 2,734
N 48
AD 0,402
P-Value 0,346
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-5,0
-2,5
0,0
2,5
5,0
7,5
RESI_4


Autocorrelation Function for RESI_4
(with 5% significance limits for the autocorrelations)
Autocorrelation
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,6
-0,8
-1,0
1
2
3
4
5
6
7
8
9
10
11
12
Lag

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
29	24,83	0,221

The residuals are normal and independent. The model is adequate.

b)

The 95% prediction interval for the differenced time series for observation 51 is the following:

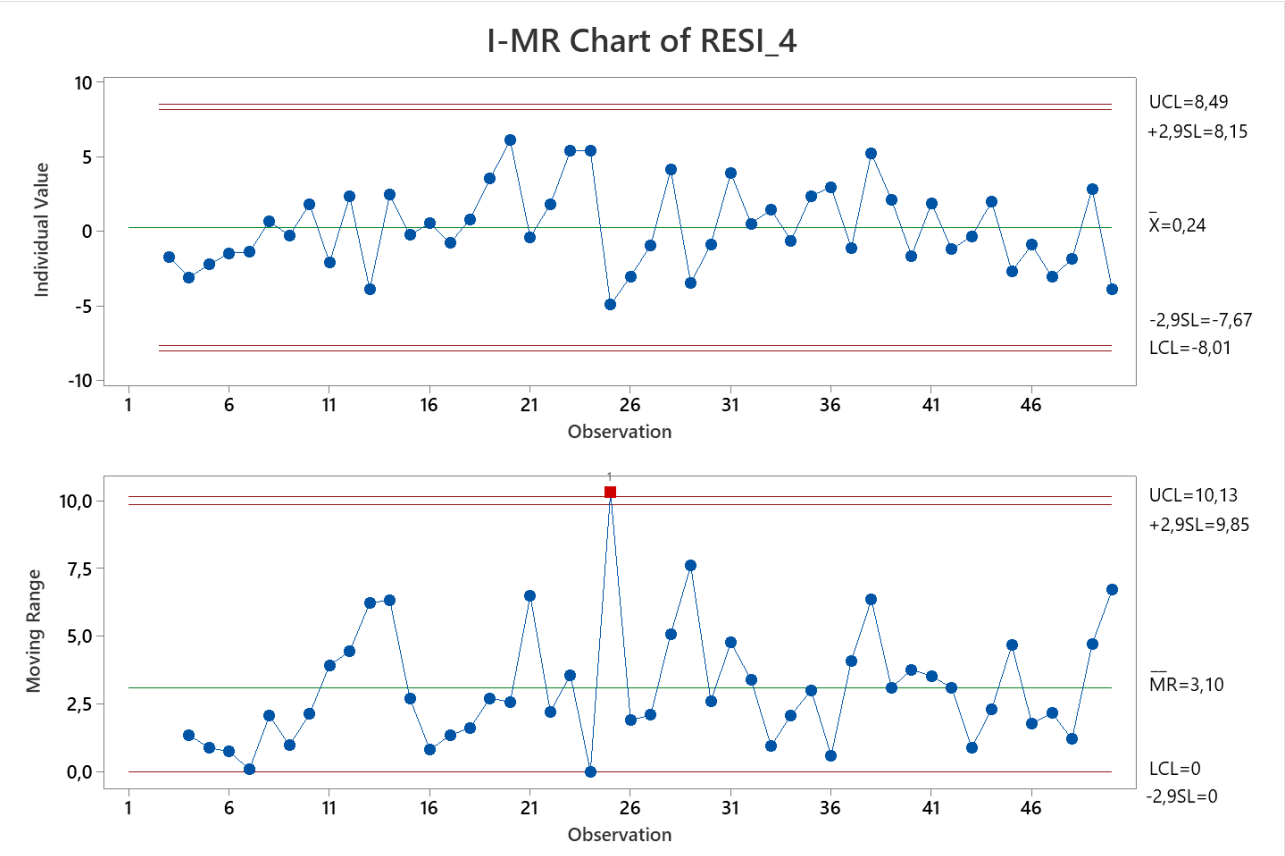
95% PI
(-2,83103; 8,65381)

This is a prediction interval on the differenced data. To obtain the prediction interval on the original data (contaminant concentration in ppm) we must sum the value of the variable at the 50th sample, i.e., $Y = 10{,}95$, thus:

$$8.119\ ppm\ \leq Y \leq 19.604\ ppm$$

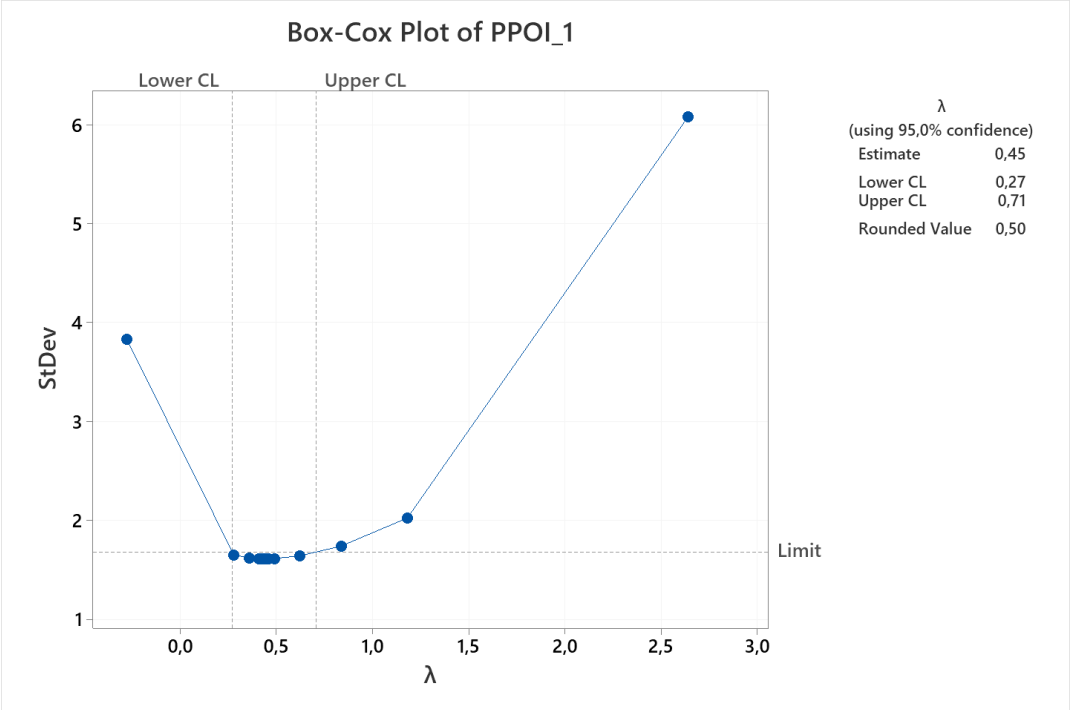
c)

The Type I error corresponding to $ARL_0 = 250$ is $\alpha = 0{,}004$, which corresponds to $k = z_{\alpha/2} = 2{,}878$. The resulting I-MR control chart for the model residuals is the following:

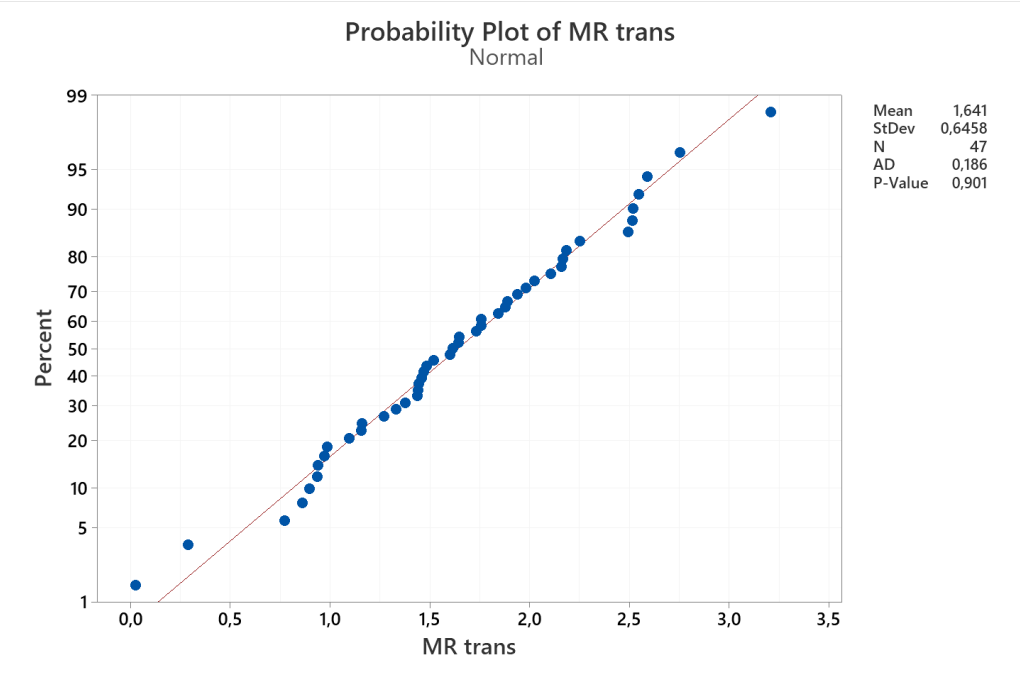


Sample 25 yields an OOC in the MR control chart. It is possible to verify if this OOC is the consequence of a violation of assumptions in the MR chart. One possible way is to transform MR data to normality and redesign the chart as follows:

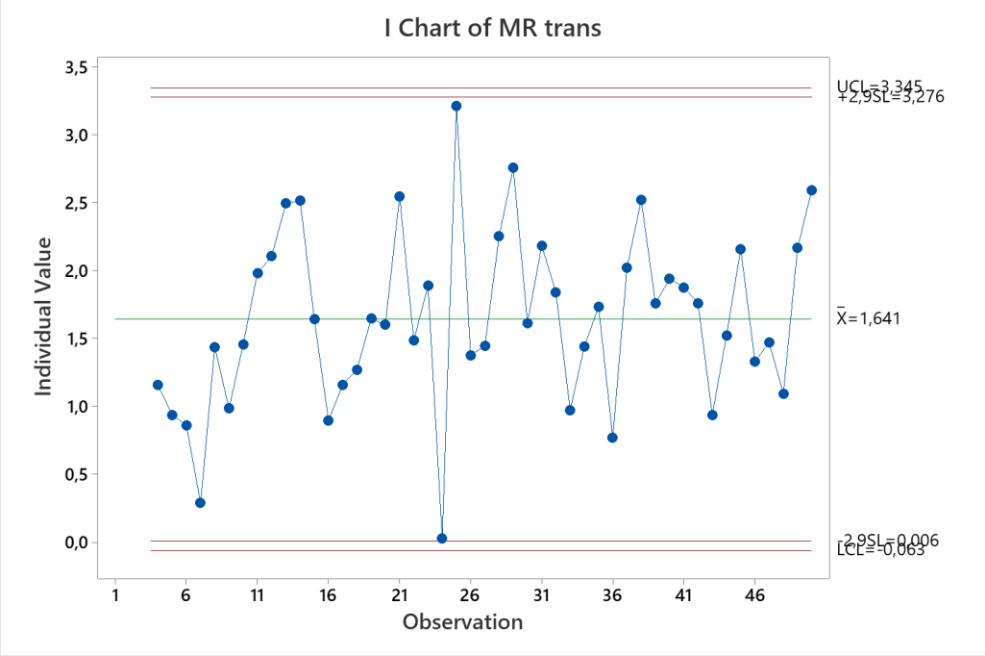
Box-Cox transformation:



Normality of MR statistic after transformation:



New MR control chart:



The OOC in the MR control chart was caused by a violation of assumptions of the chart itself.

The process is in-control.

d)

Since model residuals are normal and independent, it is possible to perform a one sample chi-squared test as follows.

By estimating the standard deviation of the model residuals as $\hat{\sigma}_\varepsilon = \sqrt{MSE} = 2.774$.

The test is such that:

$H_0: \sigma_\varepsilon = 2.5$

$H_1: \sigma_\varepsilon > 2.5$

The test statistic is $X^2 = \frac{(n-p)\hat{\sigma}_\varepsilon{}^2}{\sigma_\varepsilon{}^2} \sim X^2_{n-p}$, where $p = 2$ is the number of model terms, and $n - p = 46$.

Under H_0 we get $X^2 = 56.636$. The corresponding p-value is 0.135.

At 95% confidence, the standard deviation of residuals of the model fitted in point a) is not statistically larger than the one observed on historical data.

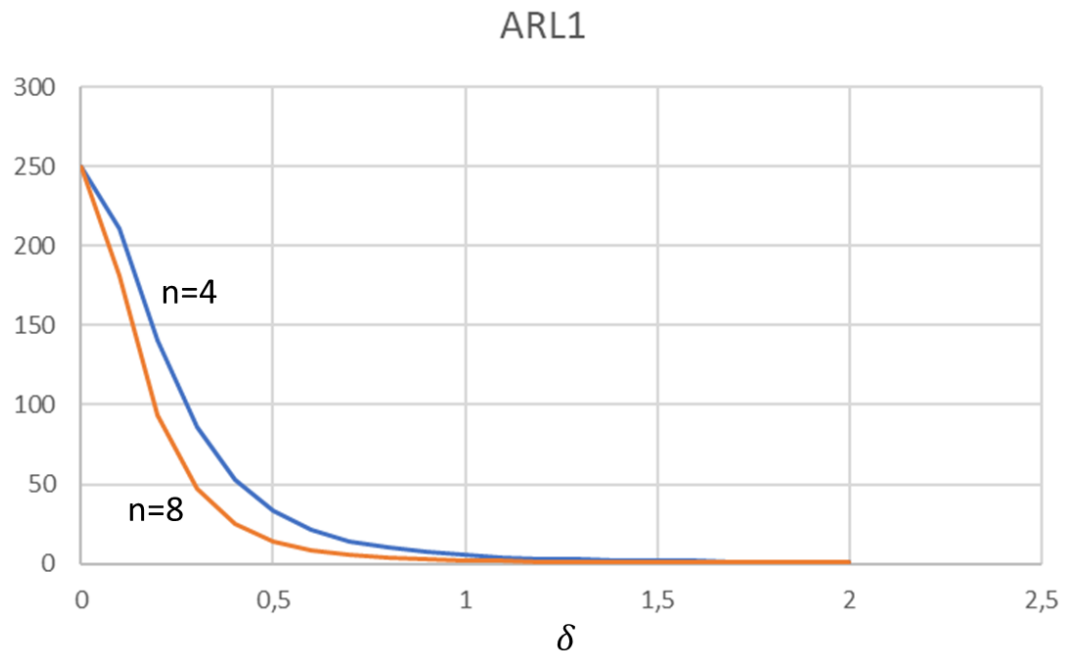
Exercise 2

The value of $K = z_{\alpha/2}$ with $\alpha = \frac{1}{250} = 0.004$ is: $K = 2.878$.

The Type II error as a function of the mean shift in standard deviation units is given by:

$\beta = \Pr(Z \leq K - \delta\sqrt{n}) - \Pr(Z \leq -K - \delta\sqrt{n})$, where $\delta = \frac{\mu_1 - \mu_0}{\sigma}$

Being, $ARL_1(\delta) = \frac{1}{1-\beta}$. The $ARL_1(\delta)$ curves for $n = 4$ and $n = 8$ are the following:



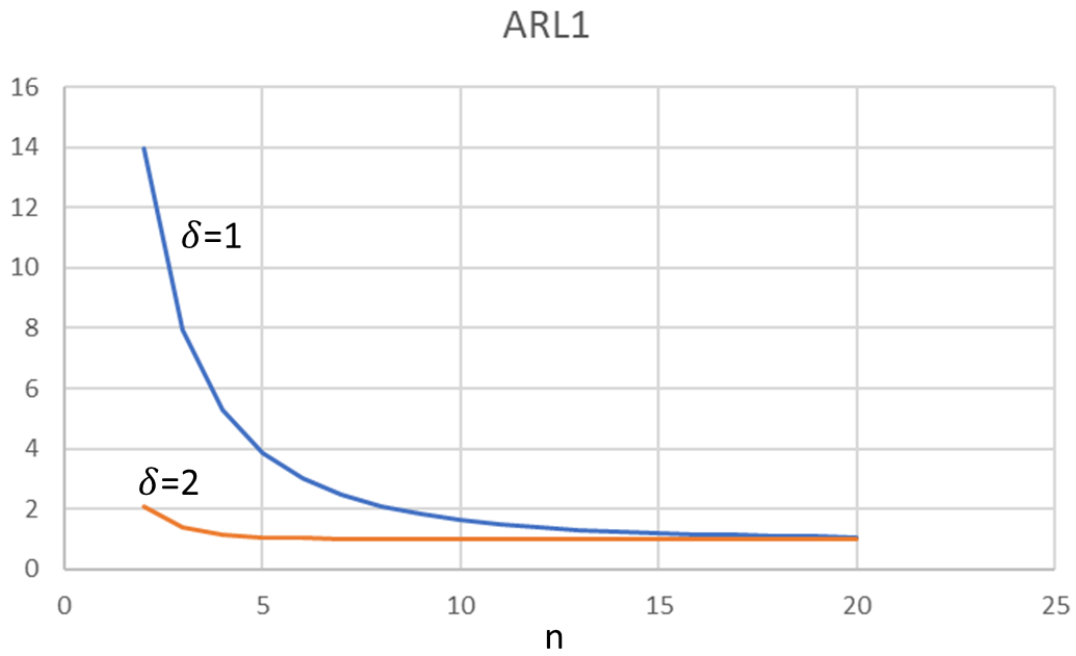
	$\delta = 1$	$\delta = 2$
ARL_1 with n=4	5.26	1.15
ARL_1 with n=8	2.08	1.00

b)

Being fixed δ, the type II error can be estimated as a function of n with the same expression used in the previous case:

$$\beta = \Pr\left(Z \leq K - \delta\sqrt{n}\right) - \Pr\left(Z \leq -K - \delta\sqrt{n}\right)$$

The resulting $ARL_1(n)$ curves for the two given mean shifts are the following:



	$n = 3$	$n = 6$
ARL_1 with δ = 1	7.94	2.99
ARL_1 with δ = 2	1.39	1.02

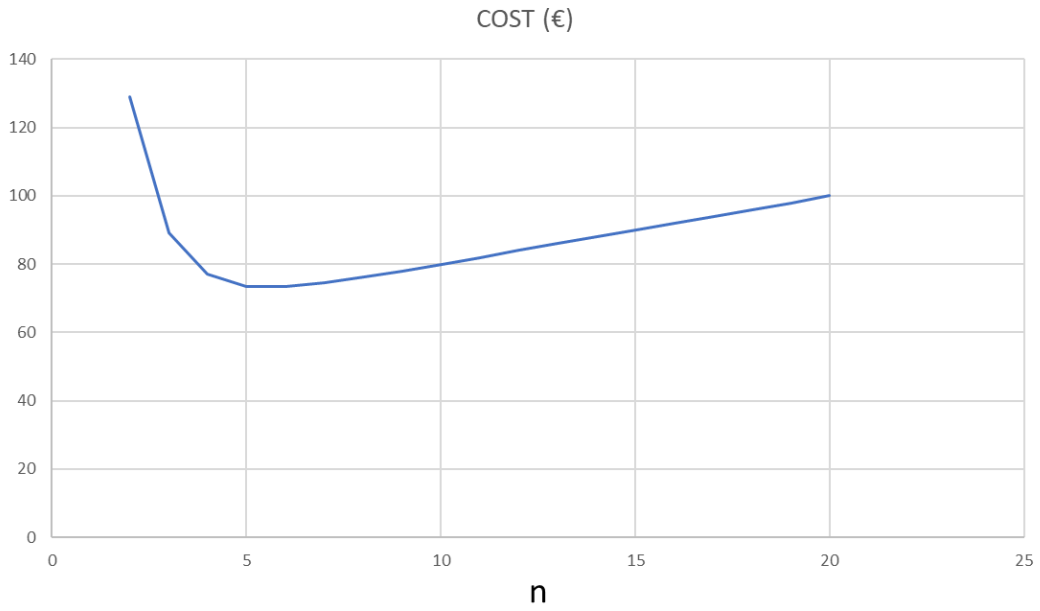
c)

The function to be minimized is the following:

$$C(n) = C1 * n + C2 * ATS(n) = 2 * n + 15 * ATS(n)$$

Where ATS = $h \cdot ARL_1$, where h is the time between the collection of two consecutive samples, i.e., $h = 4\ h$.

The cost function for $\delta = 2$ is shown below:



The late detection cost predominates at smaller values of n, whereas the inspection cost predominates at larger values of n. The optimal values of the sample size is n=6.

Exercise 3 (solution)

Given a stationary AR(1) model $X_t = \xi + \phi_1 X_{t-1} + \varepsilon_t,\ \ \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, it is known that:

$$\mu = \frac{\xi}{1-\phi_1}$$
$$\sigma^2 = \frac{\sigma_\varepsilon^2}{1-{\phi_1}^2}$$

Therefore:

$$1-\phi_1 = \frac{\xi}{\mu}$$
$$1-{\phi_1}^2 = \frac{\sigma_\varepsilon^2}{\sigma^2}$$

By solving the two equations with two unknowns:

$$\phi_1 = \sqrt{1-\frac{\sigma_\varepsilon^2}{\sigma^2}}$$

$$\xi = \mu\left(1-\sqrt{1-\frac{\sigma_\varepsilon^2}{\sigma^2}}\right)$$

QUALITY DATA ANALYSIS

06/06/2024

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
- **Multichance students should skip: point b) in Exercise 1, point a) in Exercise 2**

Exercise 1 (15 points)

The concentration of a contaminant (measured in ppm) in the production of synthetic rubber is monitored over time. '230609_ex1.csv' contains the measurements collected in 50 consecutive samples.

a) Being known that a negative value is the result of a temporary miscalibration of the measuring device, fit a suitable model to these data;
b) Based on the result of point a), estimate the 95% prediction interval for the contaminant concentration in the next sample.
c) Based on the result of point a), design an appropriate control chart for these data with $ARL_0 = 250$.
d) From historical data, it is known that the most appropriate model for this process yielded a standard deviation of residuals equal to $\sigma_\varepsilon = 2.5$. Determine, with a statistical test, if the model fitted at point a) is such that the standard deviation of residuals is greater than this value (report also the p-value of the test). Discuss the result.

Exercise 2 (15 points)

A company produces aluminum laminates. The quality control department has recently introduced a statistical monitoring tool to keep under control the planarity of the laminates. It consists of an $\bar{X}$ control chart designed such that the number of samples before a false alarm is equal to 250.

a) Estimate and draw the curves of ARL_1 as a function of the mean shift δ expressed in standard deviation units with a sample size $n = 4$ and $n = 8$, respectively (show the two curves for $\delta \in [0\ 2]$ and report the ARL_1 values for $\delta = 1$ and $\delta = 2$).
b) Estimate and draw the curves of ARL_1 as a function of the sample size n for two values of the shift, $\delta = 1$ and $\delta = 2$, where δ is expressed in standard deviation units (show the two curves for $n \in [2\ 20]$ and report the ARL_1 values for $n = 3$ and $n = 6$).
c) The head of the quality control department is interested in selecting an optimal sample size n to minimize the lack of quality costs in the presence of a mean shift equal to $\delta = 2$ standard deviation units. Knowing that samples are gathered every 4 hours, the cost of planarity measurements for each laminate is $C_1 = 2$ € and an extra cost equal to $C_2 = 15$ € is due for each hour spent in the out-of-control state, determine the optimal sample size that minimizes the overall expected costs (assume the cost of the process in its in-control state as a reference baseline). Discuss the results.

Exercise 3 (3 points)

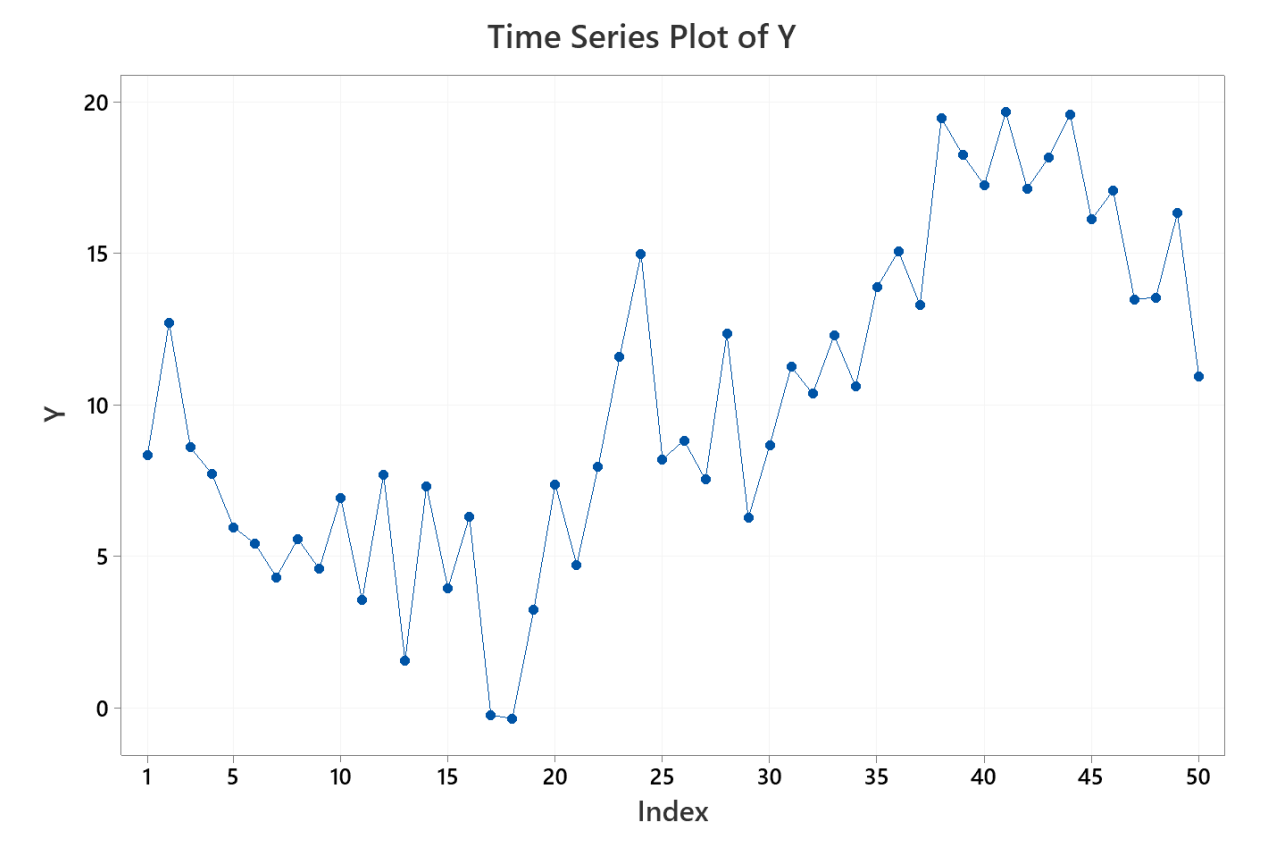
A quality characteristic X_t follows a stationary AR(1) model $X_t = \xi + \phi_1 X_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ with positive autocorrelation coefficient and known σ_ε^2. Let $E(X_t) = \mu$ and $V(X_t) = \sigma^2$. Compute the expressions of ξ and ϕ_1 as functions of μ, σ^2 and σ_ε^2.

Solutions

Exercise 1

a)

Time series plot of the temperature series:



It is present a meandering pattern. Negative values were observed in sample 17 and 18.

Runs test: null hypothesis is not accepted:

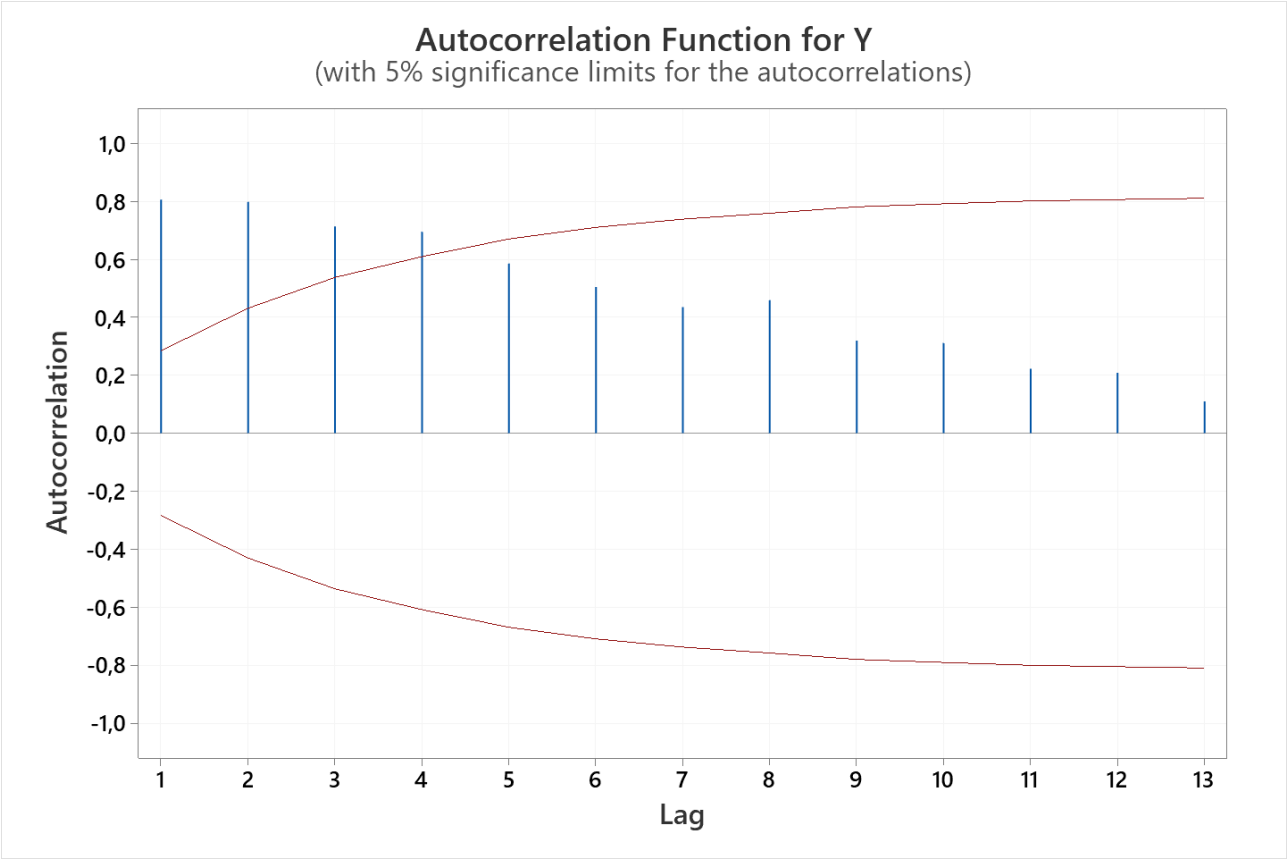
Test

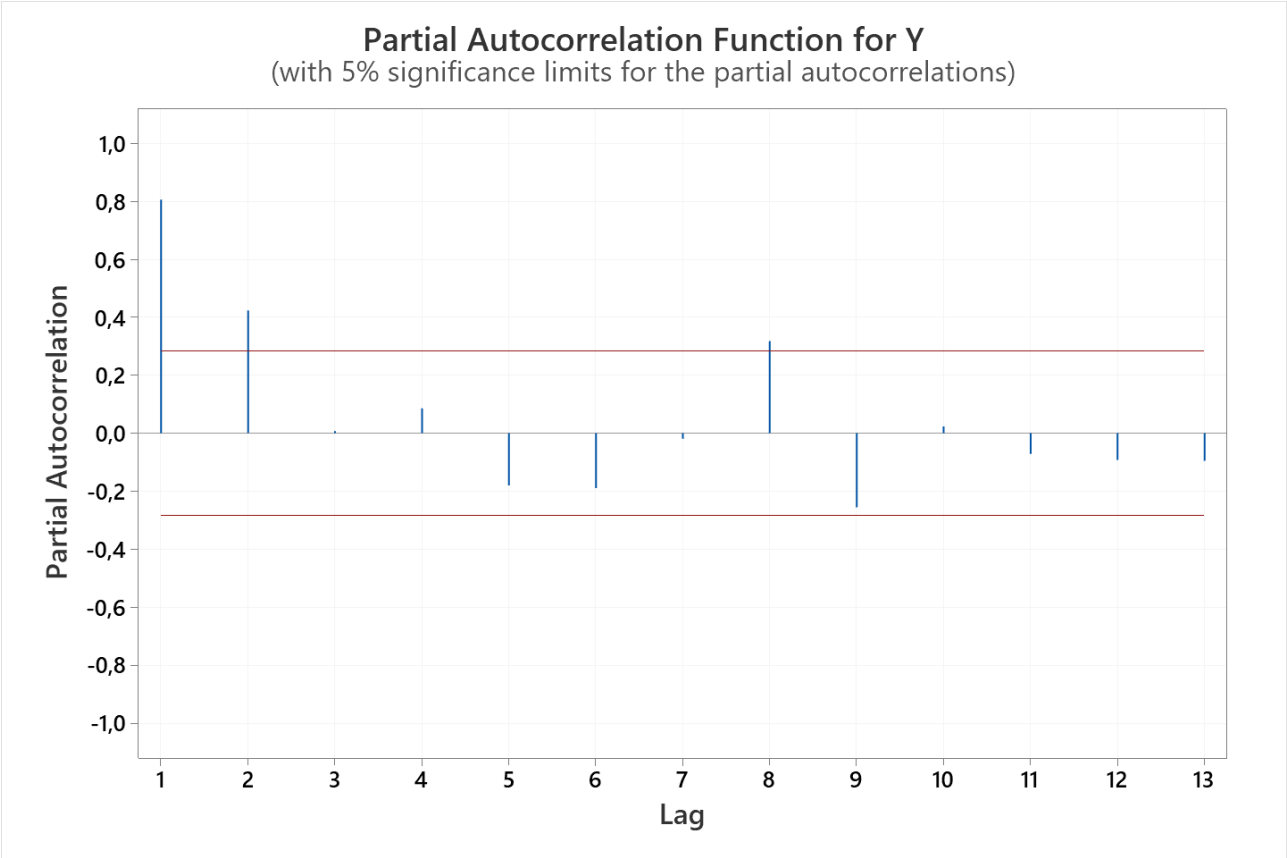
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs

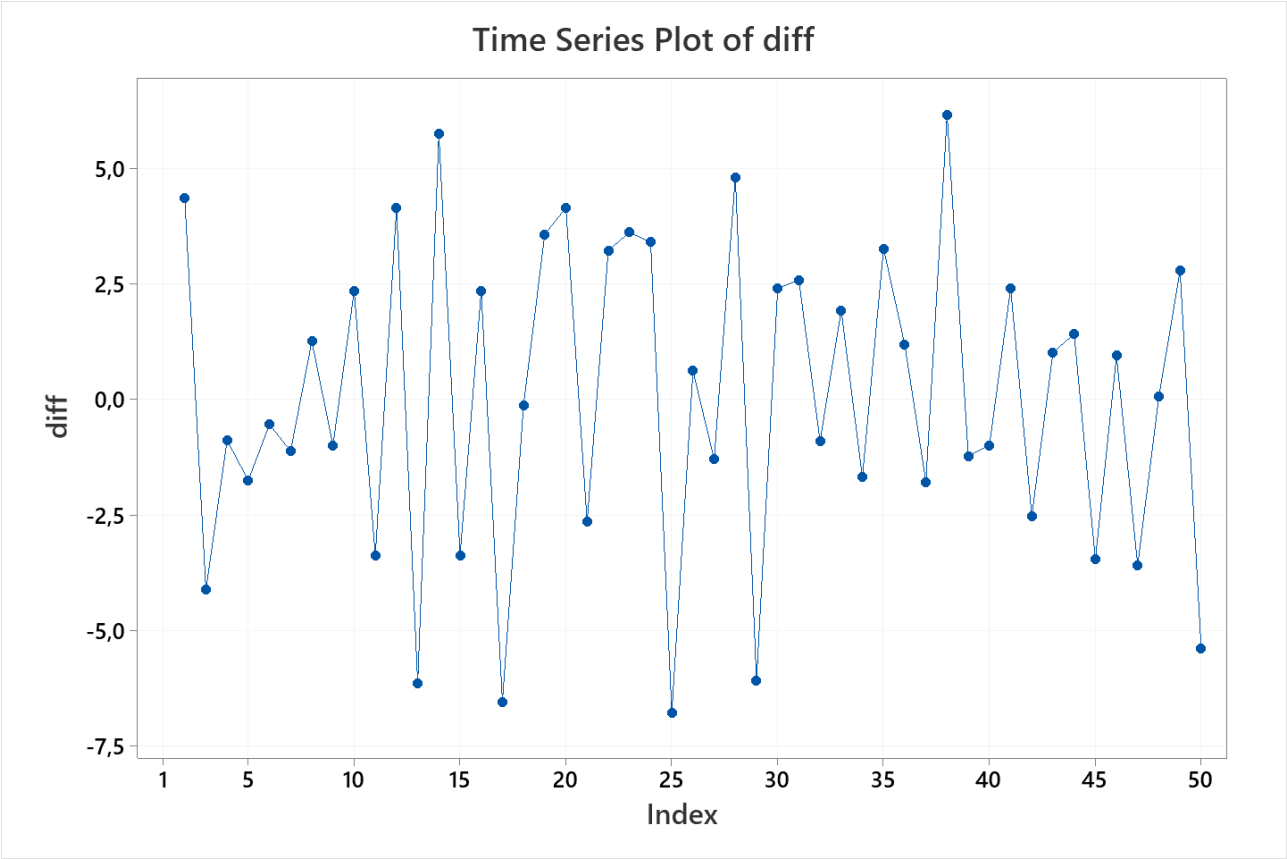
Observed	Expected	P-Value
8	25,96	0,000

Sample autocorrelation and partial autocorrelation functions:

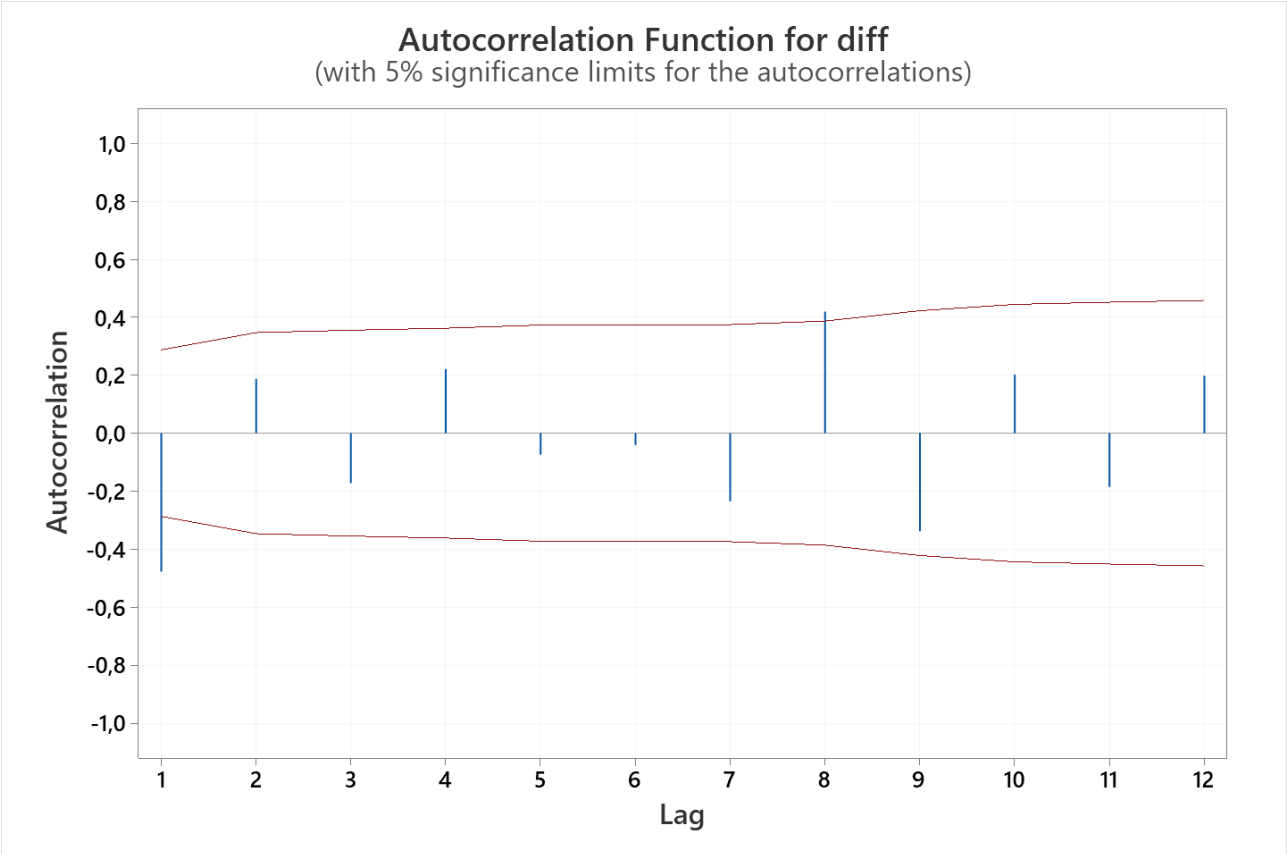


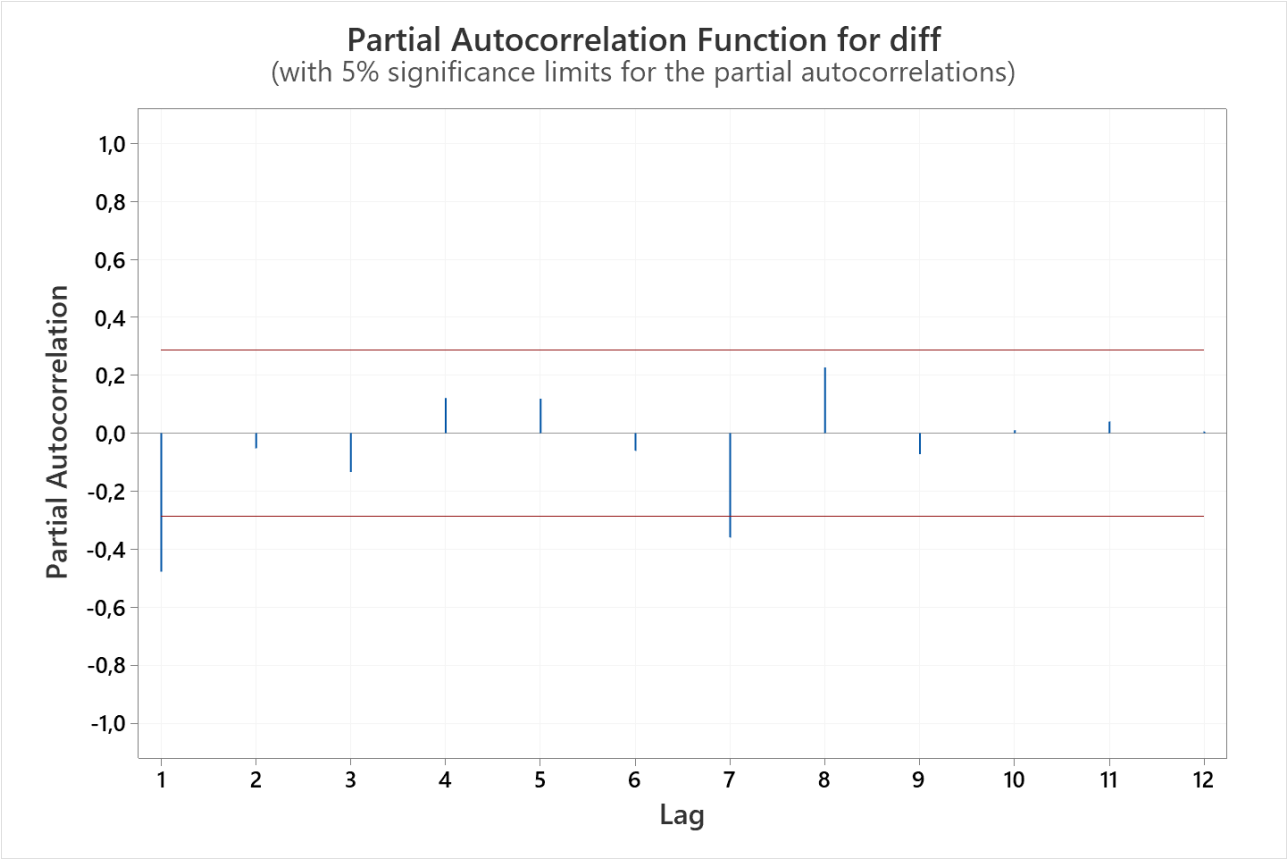


A slow decay of the SACF is present, which suggests a non-stationarity of the process. By differencing the timeseries we get:



The SACF and SPACF of the data after the differencing operation are the following:





A suitable model for the temperature time series is therefore an ARIMA(1,1,0). However, we should keep in mind that two negative values are present, caused by a temporary miscalibration of the sensor. Thus, a dummy variable that is equal to 1 for these two samples and 0 for all other samples can be included in the model.

WORKSHEET 1

Regression Analysis: diff versus AR1; dummy

Method

Categorical predictor coding	(1; 0)
Rows unused	2

Regression Equation

dummy	
0	diff = 0,251 - 0,546 AR1
1	diff = -4,47 - 0,546 AR1

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	0,251	0,413	(-0,581; 1,083)	0,61	0,547	
AR1	-0,546	0,125	(-0,797; -0,295)	-4,38	0,000	1,02
dummy						
1	-4,72	2,04	(-8,83; -0,62)	-2,32	0,025	1,02

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2,79333	32,91%	29,92%	387,706	25,92%	240,66	247,22

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	172,21	32,91%	172,21	86,106	11,04	0,000
AR1	1	130,36	24,91%	149,58	149,580	19,17	0,000
dummy	1	41,85	8,00%	41,85	41,854	5,36	0,025
Error	45	351,12	67,09%	351,12	7,803		
Total	47	523,33	100,00%				

The constant term is not significant, thus we may remove it:

WORKSHEET 1

Regression Analysis: diff versus AR1; dummy

Method

Categorical predictor coding	(1; 0)
Rows unused	2

Regression Equation

dummy	
0	diff = 0,0 - 0,540 AR1
1	diff = -4,46 - 0,540 AR1

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
AR1	-0,540	0,123	(-0,789; -0,292)	-4,37	0,000	1,02
dummy						
1	-4,46	1,98	(-8,44; -0,48)	-2,25	0,029	1,02

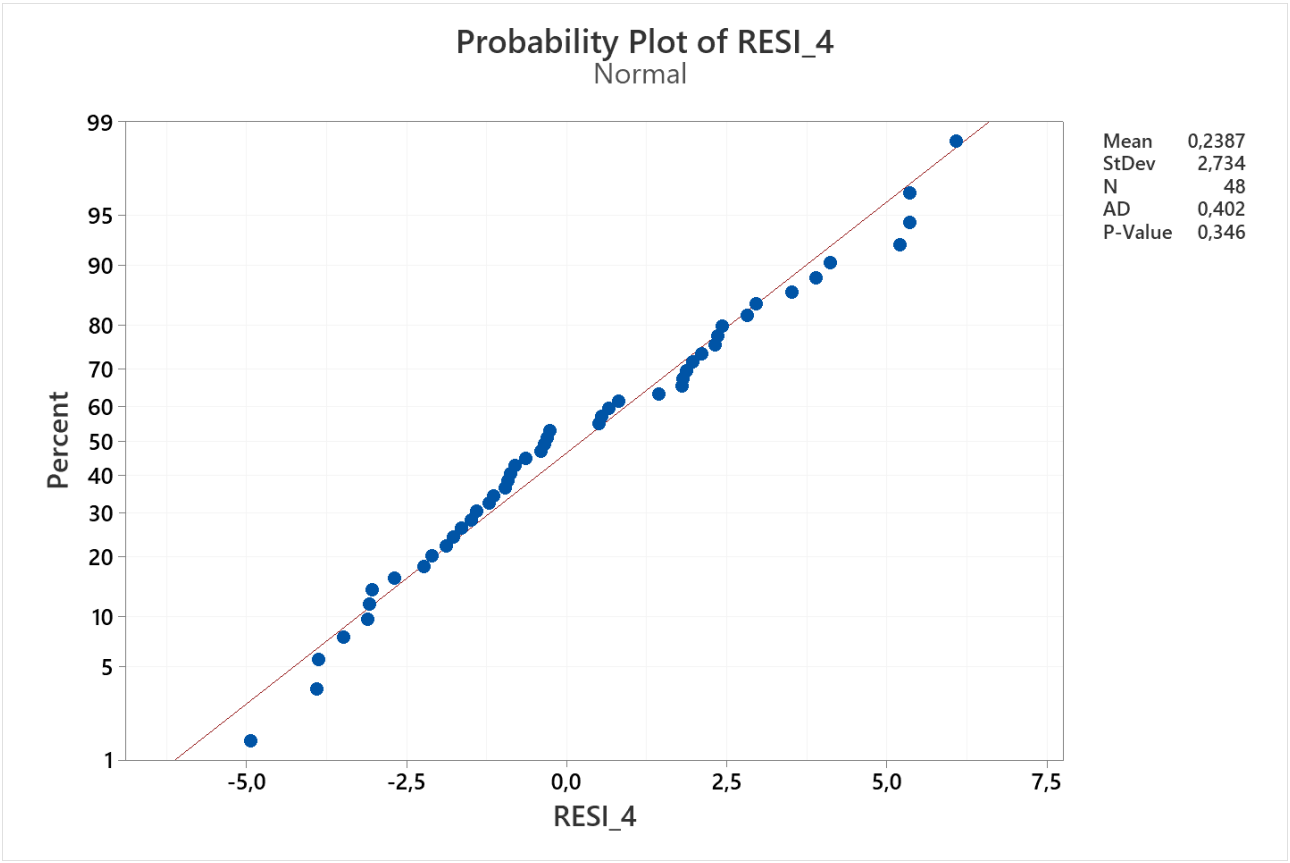
Model Summary

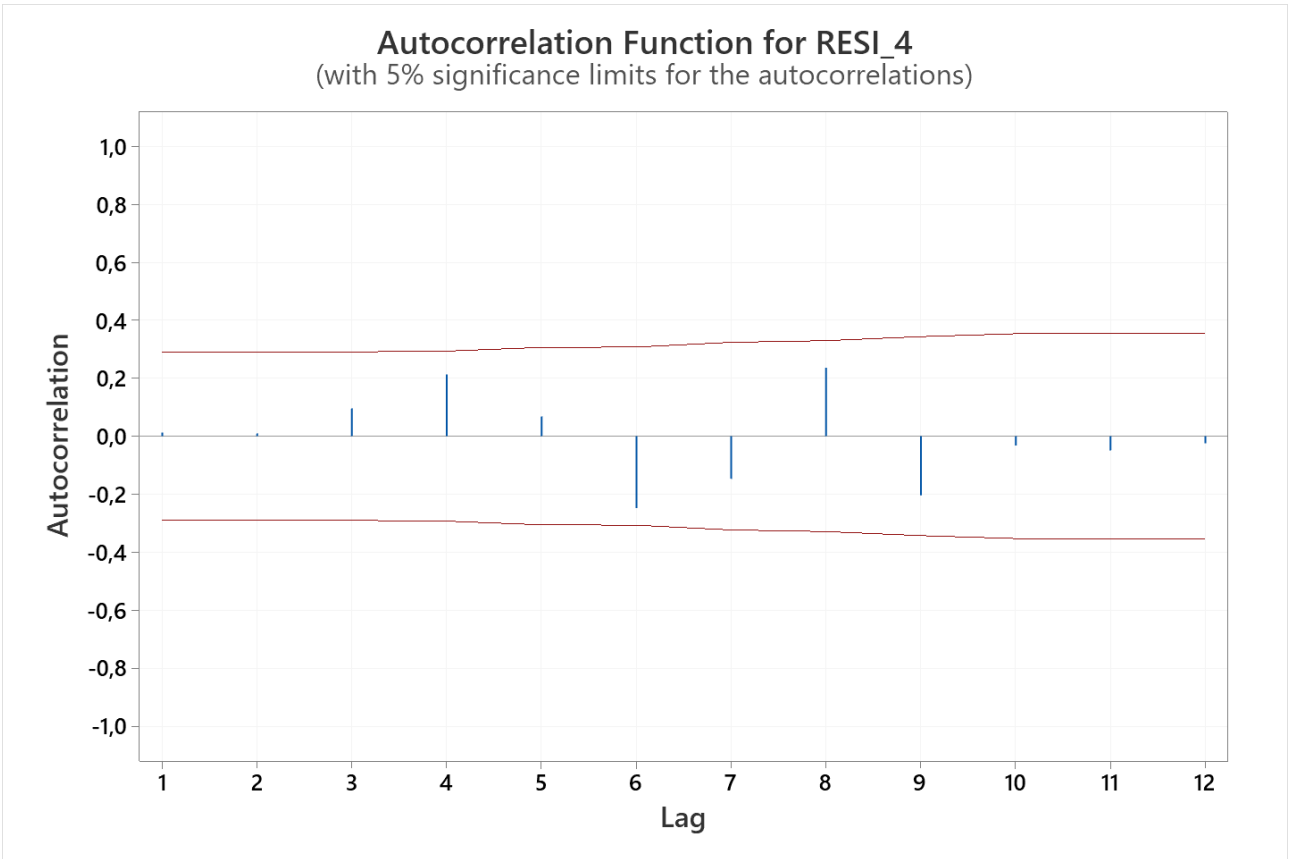
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2,77408	32,37%	29,43%	374,471	28,45%	238,67	243,74

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	169,41	32,37%	169,41	84,703	11,01	0,000
AR1	1	130,32	24,90%	147,23	147,227	19,13	0,000
dummy	1	39,09	7,47%	39,09	39,087	5,08	0,029
Error	46	353,99	67,63%	353,99	7,696		
Total	48	523,40	100,00%				

Check of residuals:


Probability Plot of RESI_4
Normal
Mean 0,2387
StDev 2,734
N 48
AD 0,402
P-Value 0,346
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-5,0
-2,5
0,0
2,5
5,0
7,5
RESI_4


Autocorrelation Function for RESI_4
(with 5% significance limits for the autocorrelations)
Autocorrelation
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,6
-0,8
-1,0
1
2
3
4
5
6
7
8
9
10
11
12
Lag

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
29	24,83	0,221

The residuals are normal and independent. The model is adequate.

b)

The 95% prediction interval for the differenced time series for observation 51 is the following:

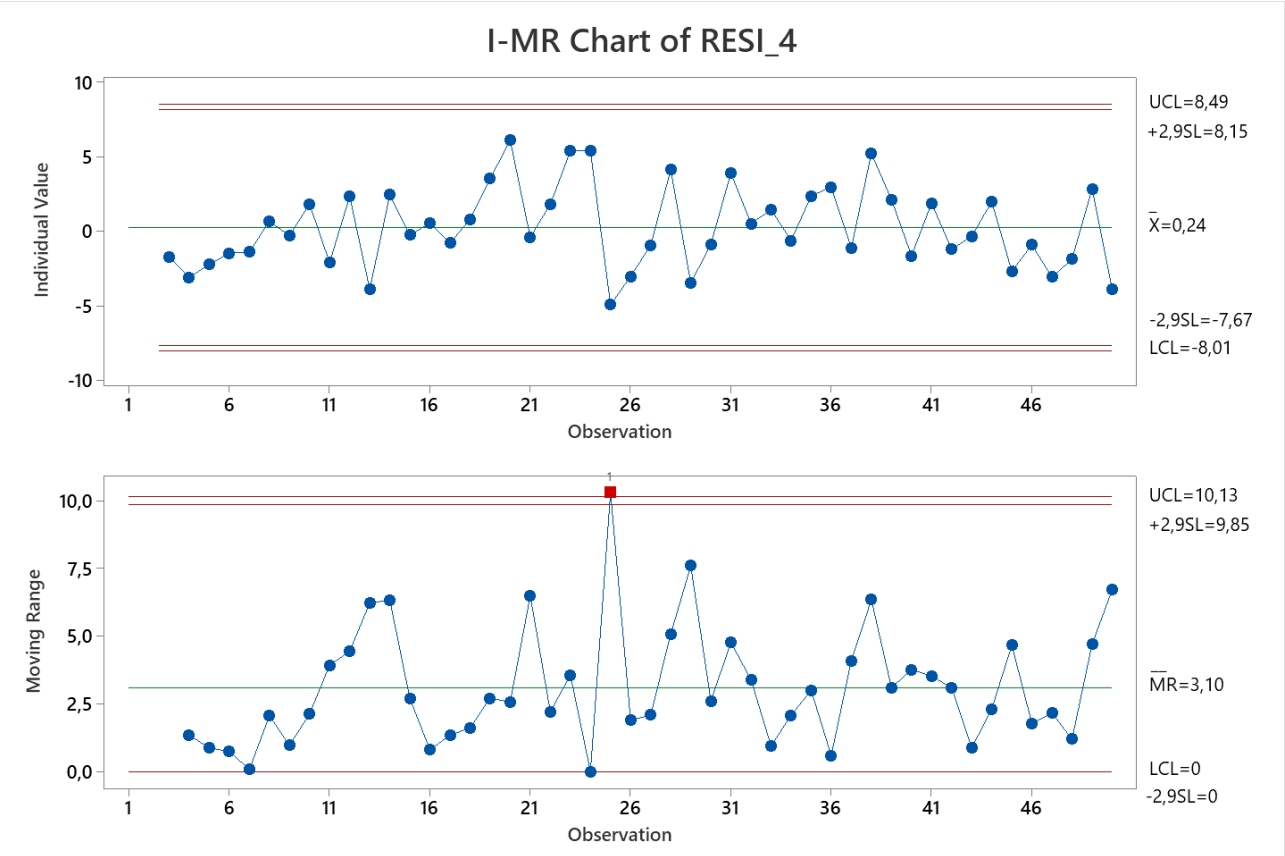
95% PI
(-2,83103; 8,65381)

This is a prediction interval on the differenced data. To obtain the prediction interval on the original data (contaminant concentration in ppm) we must sum the value of the variable at the 50th sample, i.e., $Y = 10{,}95$, thus:

$$8.119\ ppm\ \leq Y \leq 19.604\ ppm$$

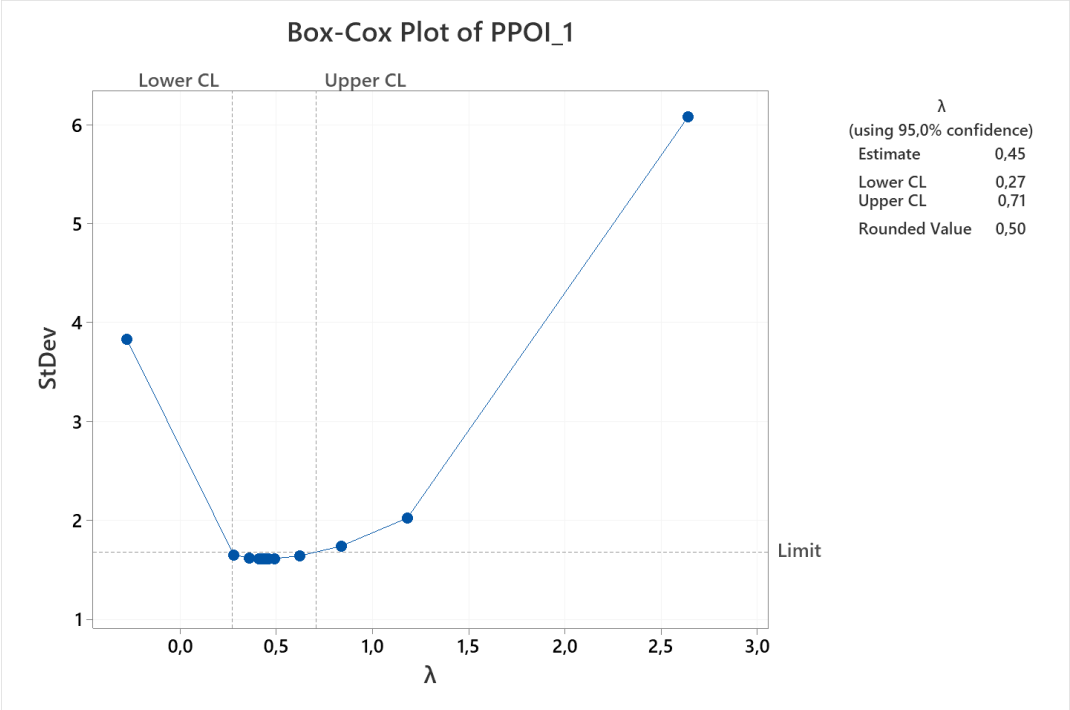
c)

The Type I error corresponding to $ARL_0 = 250$ is $\alpha = 0{,}004$, which corresponds to $k = z_{\alpha/2} = 2{,}878$. The resulting I-MR control chart for the model residuals is the following:

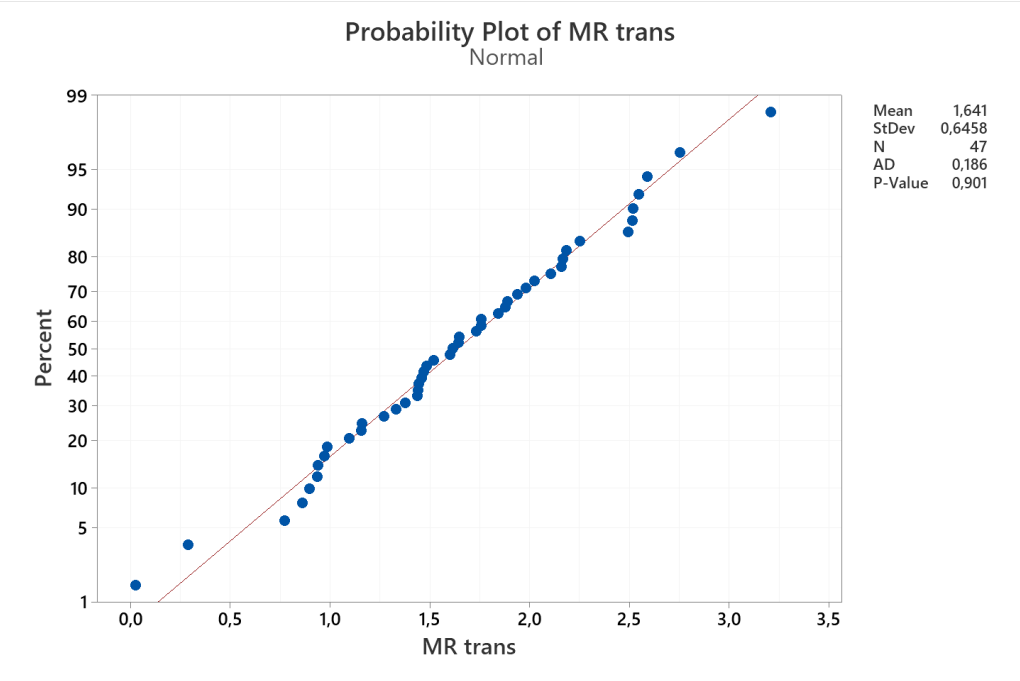


Sample 25 yields an OOC in the MR control chart. It is possible to verify if this OOC is the consequence of a violation of assumptions in the MR chart. One possible way is to transform MR data to normality and redesign the chart as follows:

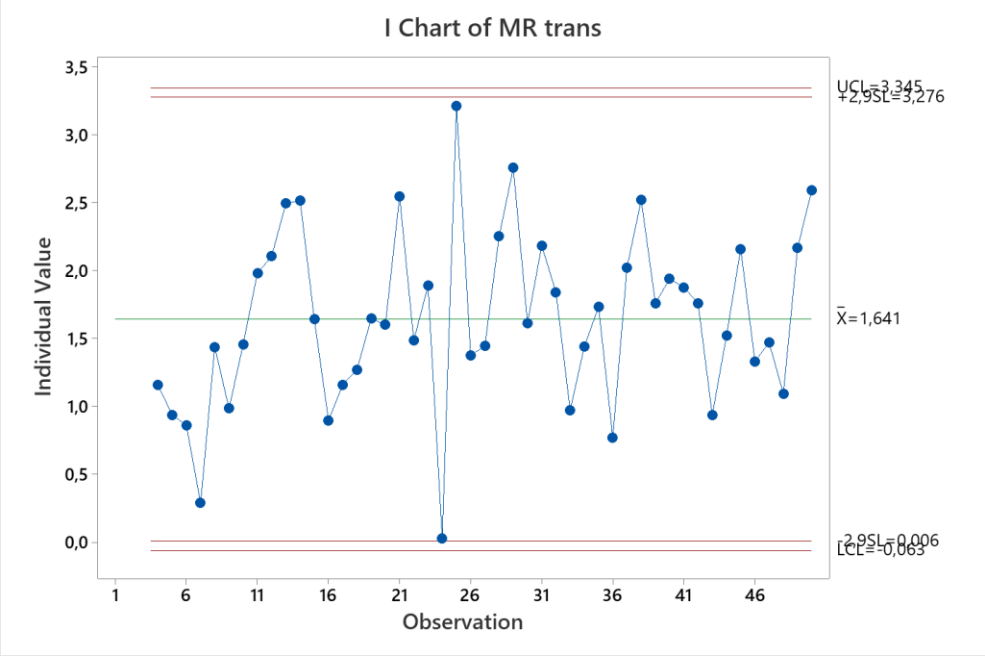
Box-Cox transformation:



Normality of MR statistic after transformation:



New MR control chart:



The OOC in the MR control chart was caused by a violation of assumptions of the chart itself.

The process is in-control.

d)

Since model residuals are normal and independent, it is possible to perform a one sample chi-squared test as follows.

By estimating the standard deviation of the model residuals as $\hat{\sigma}_\varepsilon = \sqrt{MSE} = 2.774$.

The test is such that:

$H_0: \sigma_\varepsilon = 2.5$

$H_1: \sigma_\varepsilon > 2.5$

The test statistic is $X^2 = \frac{(n-p)\hat{\sigma}_\varepsilon{}^2}{\sigma_\varepsilon{}^2} \sim X^2_{n-p}$, where $p = 2$ is the number of model terms, and $n - p = 46$.

Under H_0 we get $X^2 = 56.636$. The corresponding p-value is 0.135.

At 95% confidence, the standard deviation of residuals of the model fitted in point a) is not statistically larger than the one observed on historical data.

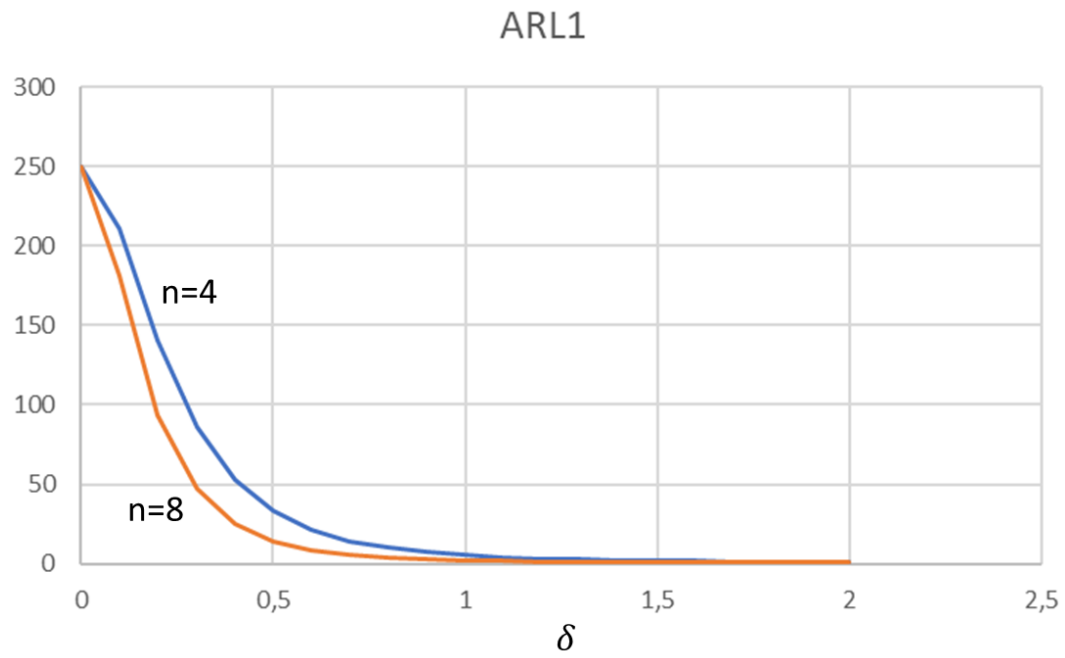
Exercise 2

The value of $K = z_{\alpha/2}$ with $\alpha = \frac{1}{250} = 0.004$ is: $K = 2.878$.

The Type II error as a function of the mean shift in standard deviation units is given by:

$\beta = \Pr(Z \leq K - \delta\sqrt{n}) - \Pr(Z \leq -K - \delta\sqrt{n})$, where $\delta = \frac{\mu_1 - \mu_0}{\sigma}$

Being, $ARL_1(\delta) = \frac{1}{1-\beta}$. The $ARL_1(\delta)$ curves for $n = 4$ and $n = 8$ are the following:



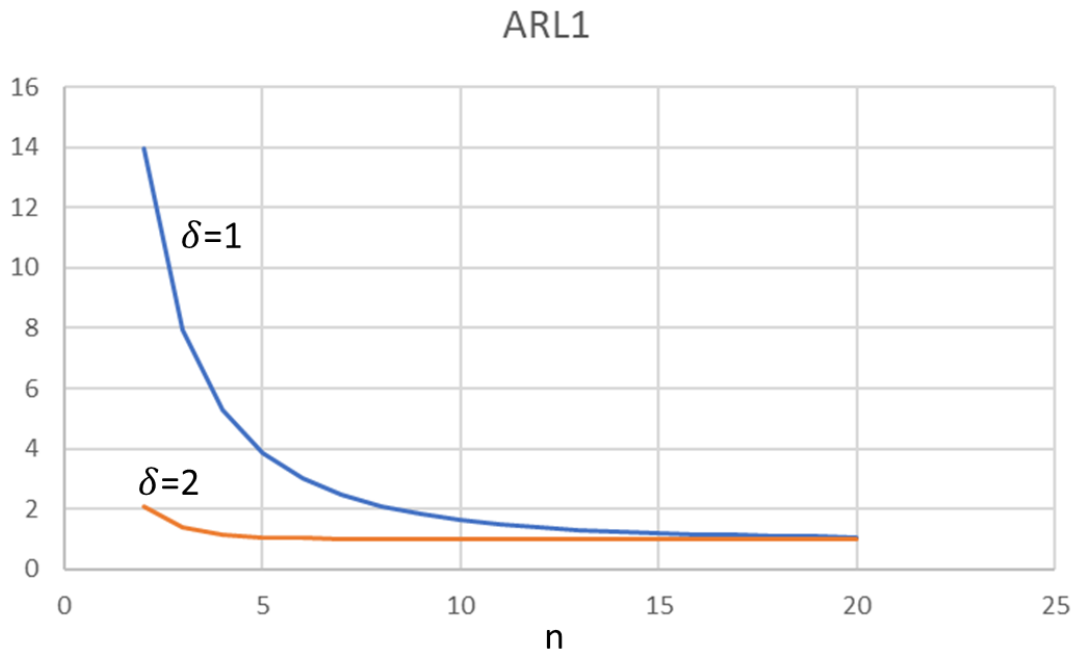
	$\delta = 1$	$\delta = 2$
ARL_1 with n=4	5.26	1.15
ARL_1 with n=8	2.08	1.00

b)

Being fixed δ, the type II error can be estimated as a function of n with the same expression used in the previous case:

$$\beta = \Pr\left(Z \leq K - \delta\sqrt{n}\right) - \Pr\left(Z \leq -K - \delta\sqrt{n}\right)$$

The resulting $ARL_1(n)$ curves for the two given mean shifts are the following:



	$n = 3$	$n = 6$
ARL_1 with δ = 1	7.94	2.99
ARL_1 with δ = 2	1.39	1.02

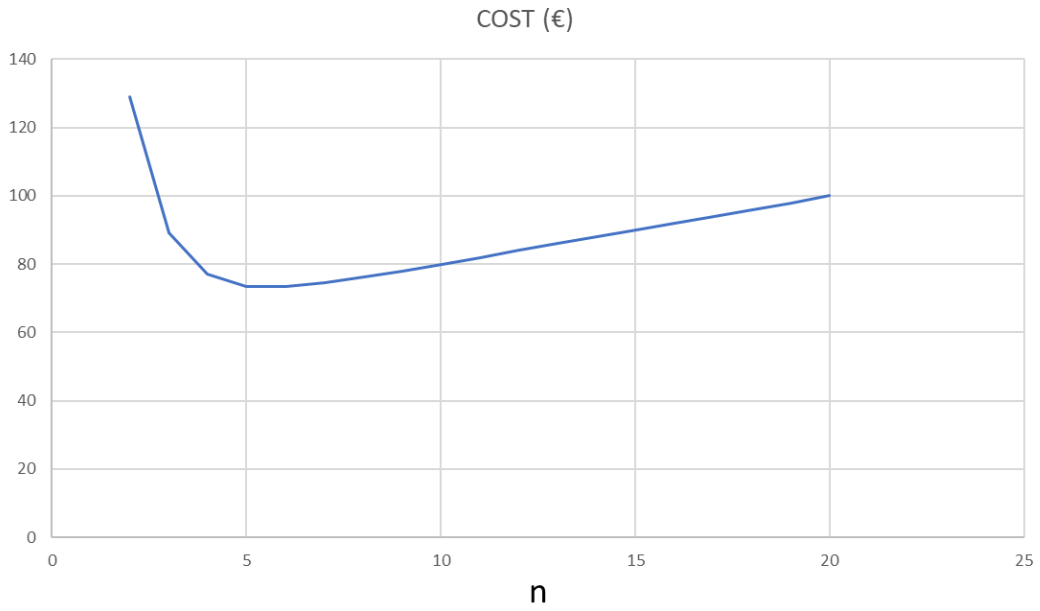
c)

The function to be minimized is the following:

$$C(n) = C1 * n + C2 * ATS(n) = 2 * n + 15 * ATS(n)$$

Where ATS = $h \cdot ARL_1$, where h is the time between the collection of two consecutive samples, i.e., $h = 4\ h$.

The cost function for $\delta = 2$ is shown below:



The late detection cost predominates at smaller values of n, whereas the inspection cost predominates at larger values of n. The optimal values of the sample size is n=6.

Exercise 3 (solution)

Given a stationary AR(1) model $X_t = \xi + \phi_1 X_{t-1} + \varepsilon_t,\ \ \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, it is known that:

$$\mu = \frac{\xi}{1-\phi_1}$$
$$\sigma^2 = \frac{\sigma_\varepsilon^2}{1-{\phi_1}^2}$$

Therefore:

$$1-\phi_1 = \frac{\xi}{\mu}$$
$$1-{\phi_1}^2 = \frac{\sigma_\varepsilon^2}{\sigma^2}$$

By solving the two equations with two unknowns:

$$\phi_1 = \sqrt{1-\frac{\sigma_\varepsilon^2}{\sigma^2}}$$

$$\xi = \mu\left(1-\sqrt{1-\frac{\sigma_\varepsilon^2}{\sigma^2}}\right)$$

QUALITY DATA ANALYSIS

30/08/2022

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
- **Multichance students should skip: point b) in Exercise 1, point a) in Exercise 2**

Exercise 1 (15 points)

The concentration of a contaminant (measured in ppm) in the production of synthetic rubber is monitored over time. Table 1 shows the measurements collected in 50 consecutive samples.

Table 1

Sample	Concentration	Sample	Concentration
1	8,36	26	8,83
2	12,72	27	7,54
3	8,60	28	12,35
4	7,72	29	6,27
5	5,97	30	8,68
6	5,43	31	11,27
7	4,32	32	10,37
8	5,58	33	12,3
9	4,59	34	10,62
10	6,94	35	13,89
11	3,56	36	15,08
12	7,71	37	13,3
13	1,57	38	19,47
14	7,32	39	18,25
15	3,95	40	17,26
16	6,31	41	19,67
17	-0,23	42	17,15
18	-0,35	43	18,17
19	3,23	44	19,59
20	7,38	45	16,13
21	4,73	46	17,08
22	7,96	47	13,49
23	11,58	48	13,55
24	14,99	49	16,34
25	8,21	50	10,95

a) Being known that a negative value is the result of a temporary miscalibration of the measuring device, fit a suitable model to these data;
b) Based on the result of point a), estimate the 95% prediction interval for the contaminant concentration in the next sample.
c) Based on the result of point a), design an appropriate control chart for these data with $ARL_0 = 250$.
d) From historical data, it is known that the most appropriate model for this process yielded a standard deviation of residuals equal to $\sigma_\varepsilon = 2.5$. Determine, with a statistical test, if the model fitted at point a) is such that the standard deviation of residuals is greater than this value (report also the p-value of the test). Discuss the result.

Exercise 2 (15 points)

A company produces aluminum laminates. The quality control department has recently introduced a statistical monitoring tool to keep under control the planarity of the laminates. It consists of an $\bar{X}$ control chart designed such that the number of samples before a false alarm is equal to 250.

a) Estimate and draw the curves of ARL_1 as a function of the mean shift δ expressed in standard deviation units with a sample size $n = 4$ and $n = 8$, respectively (show the two curves for $\delta \in [0\ 2]$ and report the ARL_1 values for $\delta = 1$ and $\delta = 2$).
b) Estimate and draw the curves of ARL_1 as a function of the sample size n for two values of the shift, $\delta = 1$ and $\delta = 2$, where δ is expressed in standard deviation units (show the two curves for $n \in [2\ 20]$ and report the ARL_1 values for $n = 3$ and $n = 6$).
c) The head of the quality control department is interested in selecting an optimal sample size n to minimize the lack of quality costs in the presence of a mean shift equal to $\delta = 2$ standard deviation units. Knowing that samples are gathered every 4 hours, the cost of planarity measurements for each laminate is $C_1 = 2$ € and an extra cost equal to $C_2 = 15$ € is due for each hour spent in the out-of-control state, determine the optimal sample size that minimizes the overall expected costs (assume the cost of the process in its in-control state as a reference baseline). Discuss the results.

Exercise 3 (3 points)

A company that produces thermal cameras is interested in monitoring the calibration curves of their devices. The calibration curve can be modelled by a linear model $y = \beta_0 + \beta_1 x + \varepsilon_t$ where the regressor x is the infrared counts measured by the sensor, whereas y is the temperature shown as output by the camera. All calibration curves are generated by using the same infrared counts levels for the regressor; moreover, the intercept $\hat{\beta}_0 = b_0$ and the slope $\hat{\beta}_1 = b_1$ are estimated using ordinary least squares.

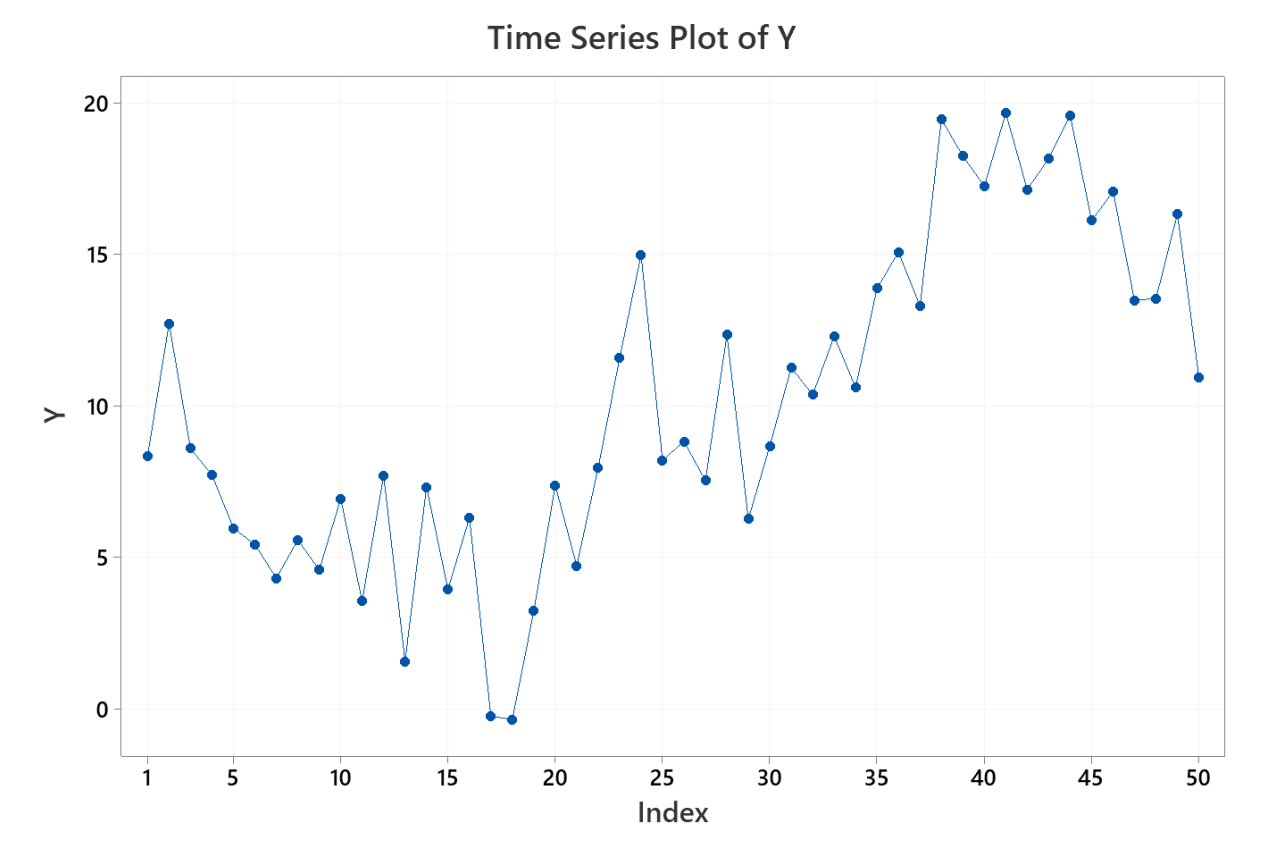
Assuming $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, and assuming that β_0 and β_1 and σ_ε^2 are known, write down the expression of the control limits of a control chart for monitoring the slope of calibration curves.

Solutions

Exercise 1

a)

Time series plot of the temperature series:



It is present a meandering pattern. Negative values were observed in sample 17 and 18.

Runs test: null hypothesis is not accepted:

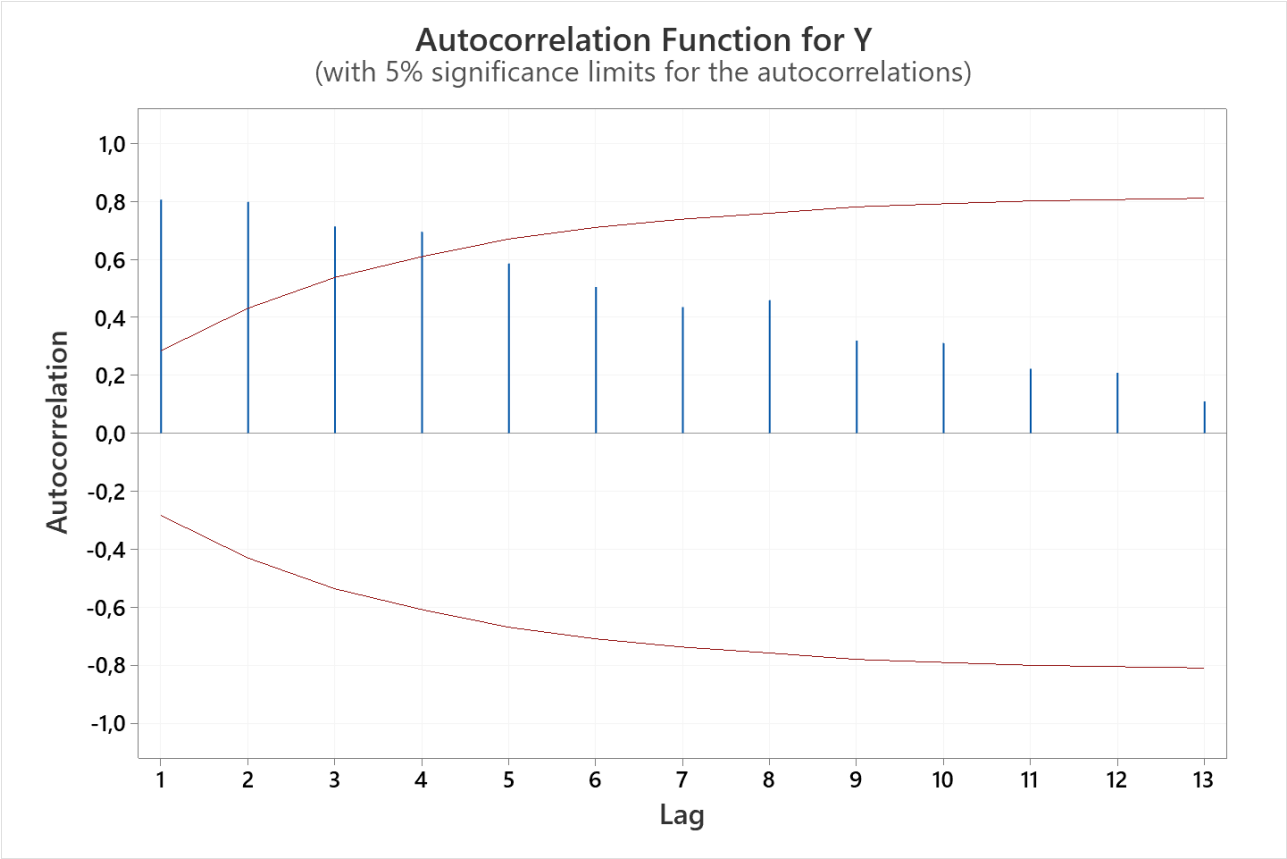
Test

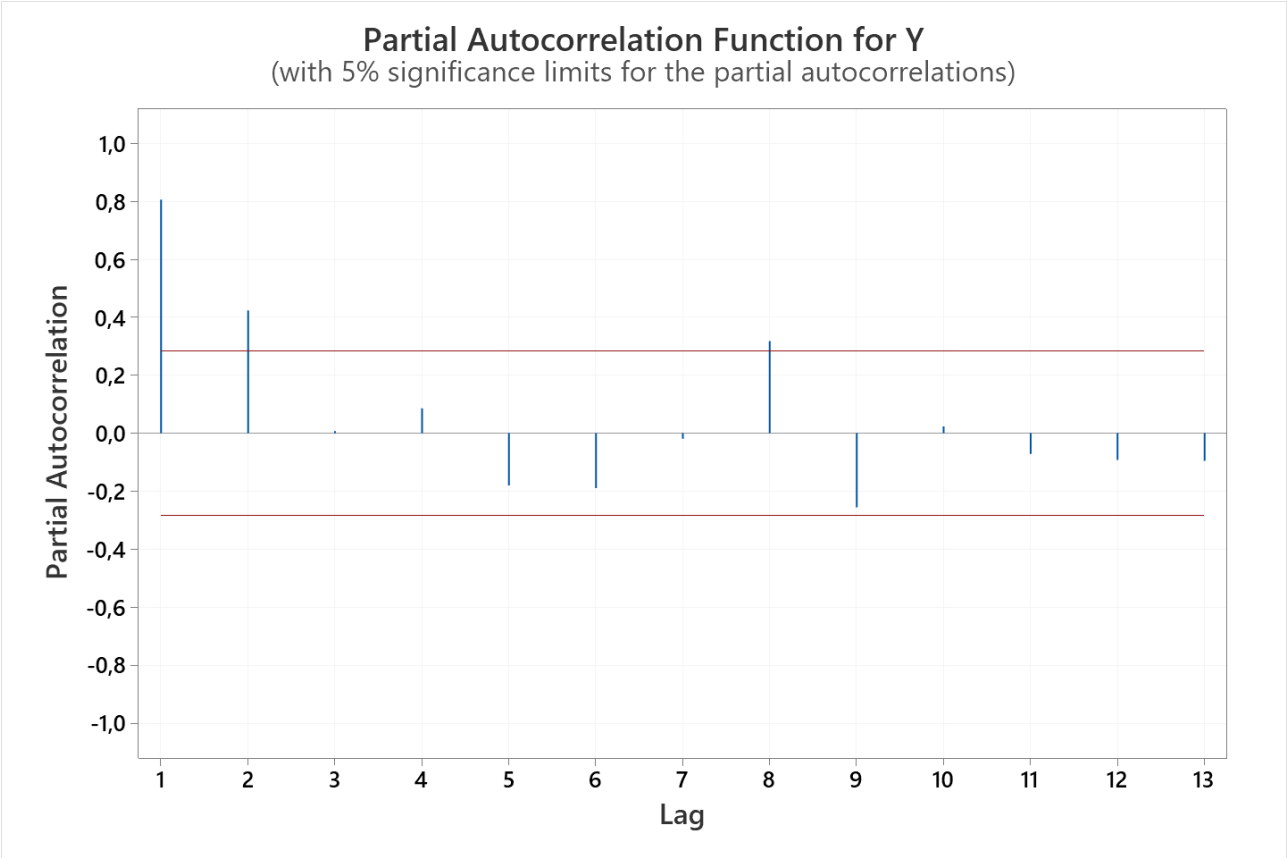
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs

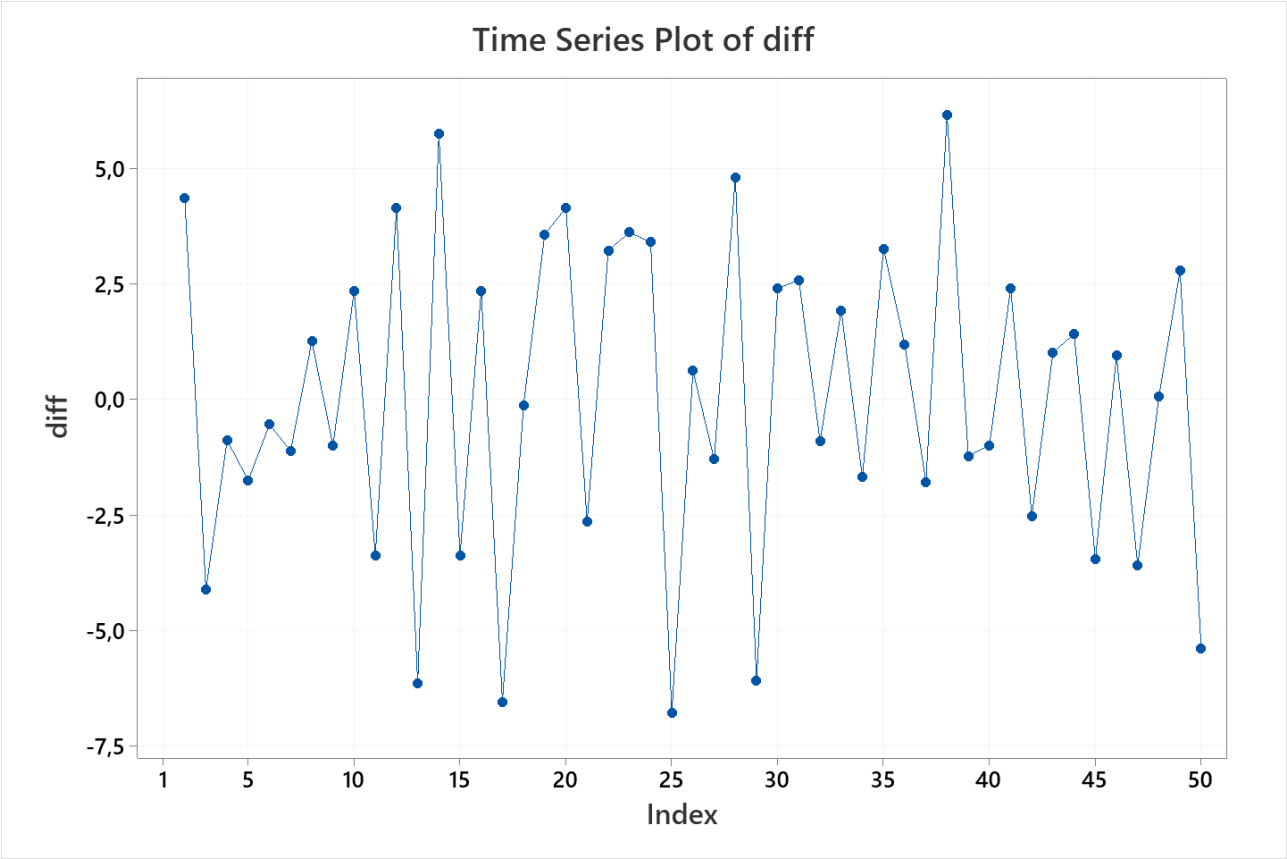
Observed	Expected	P-Value
8	25,96	0,000

Sample autocorrelation and partial autocorrelation functions:

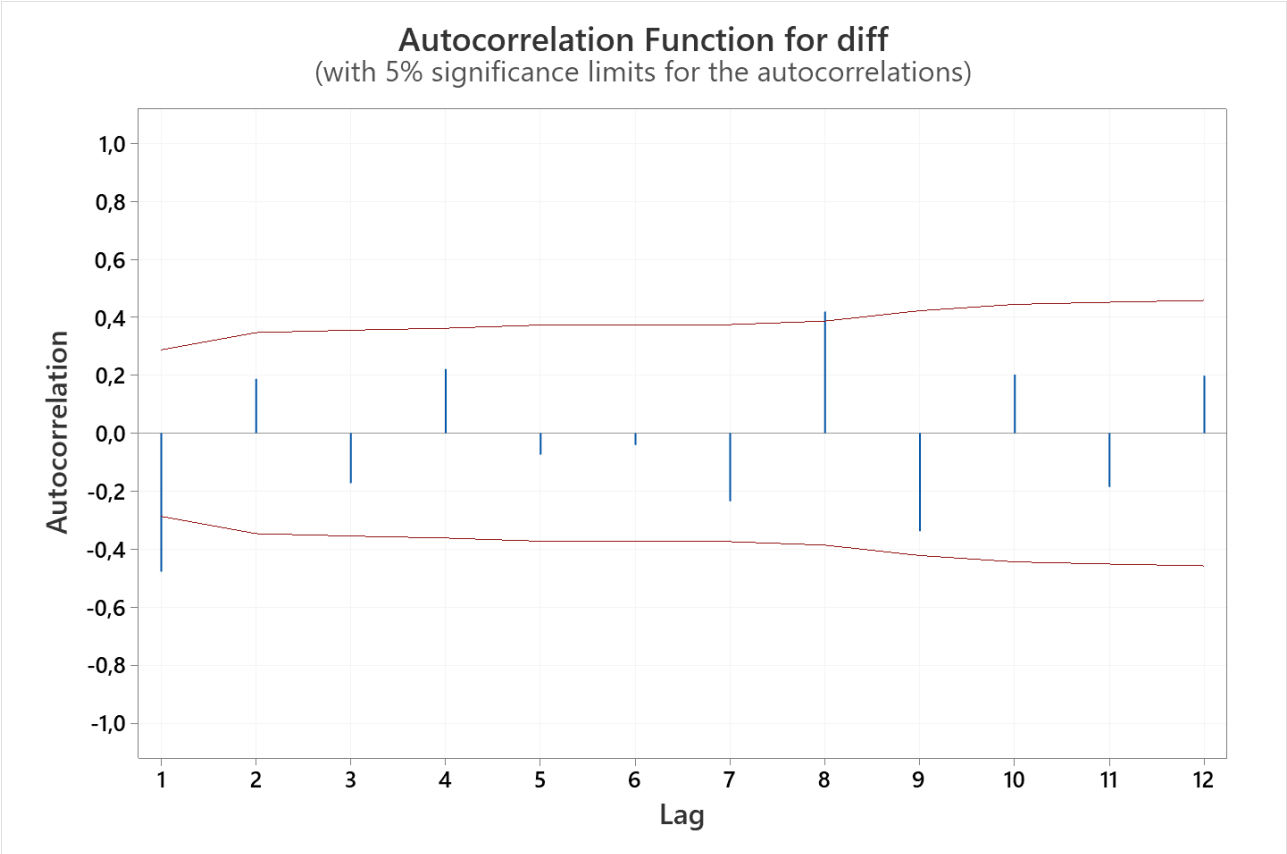


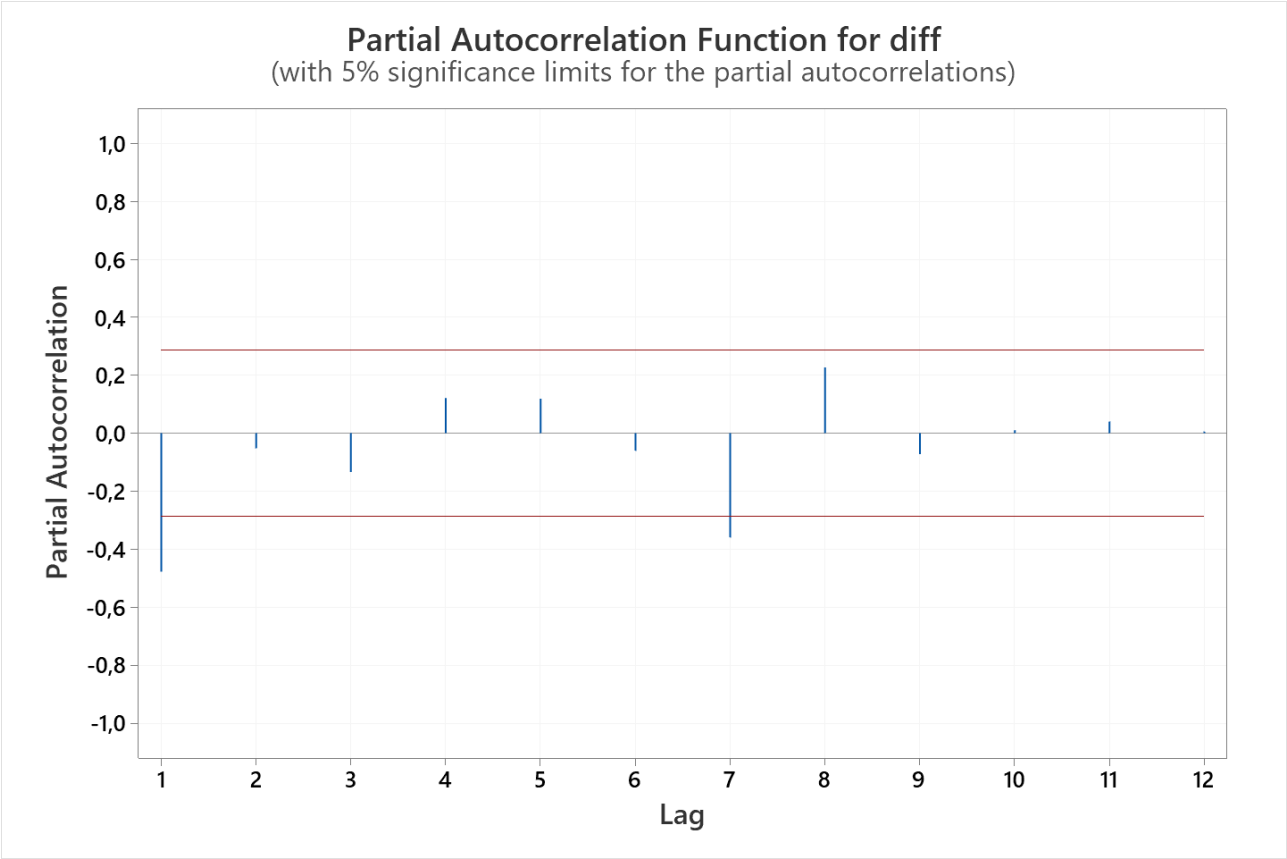


A slow decay of the SACF is present, which suggests a non-stationarity of the process. By differencing the timeseries we get:



The SACF and SPACF of the data after the differencing operation are the following:





A suitable model for the temperature time series is therefore an ARIMA(1,1,0). However, we should keep in mind that two negative values are present, caused by a temporary miscalibration of the sensor. Thus, a dummy variable that is equal to 1 for these two samples and 0 for all other samples can be included in the model.

WORKSHEET 1

Regression Analysis: diff versus AR1; dummy

Method

Categorical predictor coding	(1; 0)
Rows unused	2

Regression Equation

dummy	
0	diff = 0,251 - 0,546 AR1
1	diff = -4,47 - 0,546 AR1

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	0,251	0,413	(-0,581; 1,083)	0,61	0,547	
AR1	-0,546	0,125	(-0,797; -0,295)	-4,38	0,000	1,02
dummy						
1	-4,72	2,04	(-8,83; -0,62)	-2,32	0,025	1,02

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2,79333	32,91%	29,92%	387,706	25,92%	240,66	247,22

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	172,21	32,91%	172,21	86,106	11,04	0,000
AR1	1	130,36	24,91%	149,58	149,580	19,17	0,000
dummy	1	41,85	8,00%	41,85	41,854	5,36	0,025
Error	45	351,12	67,09%	351,12	7,803		
Total	47	523,33	100,00%				

The constant term is not significant, thus we may remove it:

WORKSHEET 1

Regression Analysis: diff versus AR1; dummy

Method

Categorical predictor coding	(1; 0)
Rows unused	2

Regression Equation

dummy	
0	diff = 0,0 - 0,540 AR1
1	diff = -4,46 - 0,540 AR1

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
AR1	-0,540	0,123	(-0,789; -0,292)	-4,37	0,000	1,02
dummy						
1	-4,46	1,98	(-8,44; -0,48)	-2,25	0,029	1,02

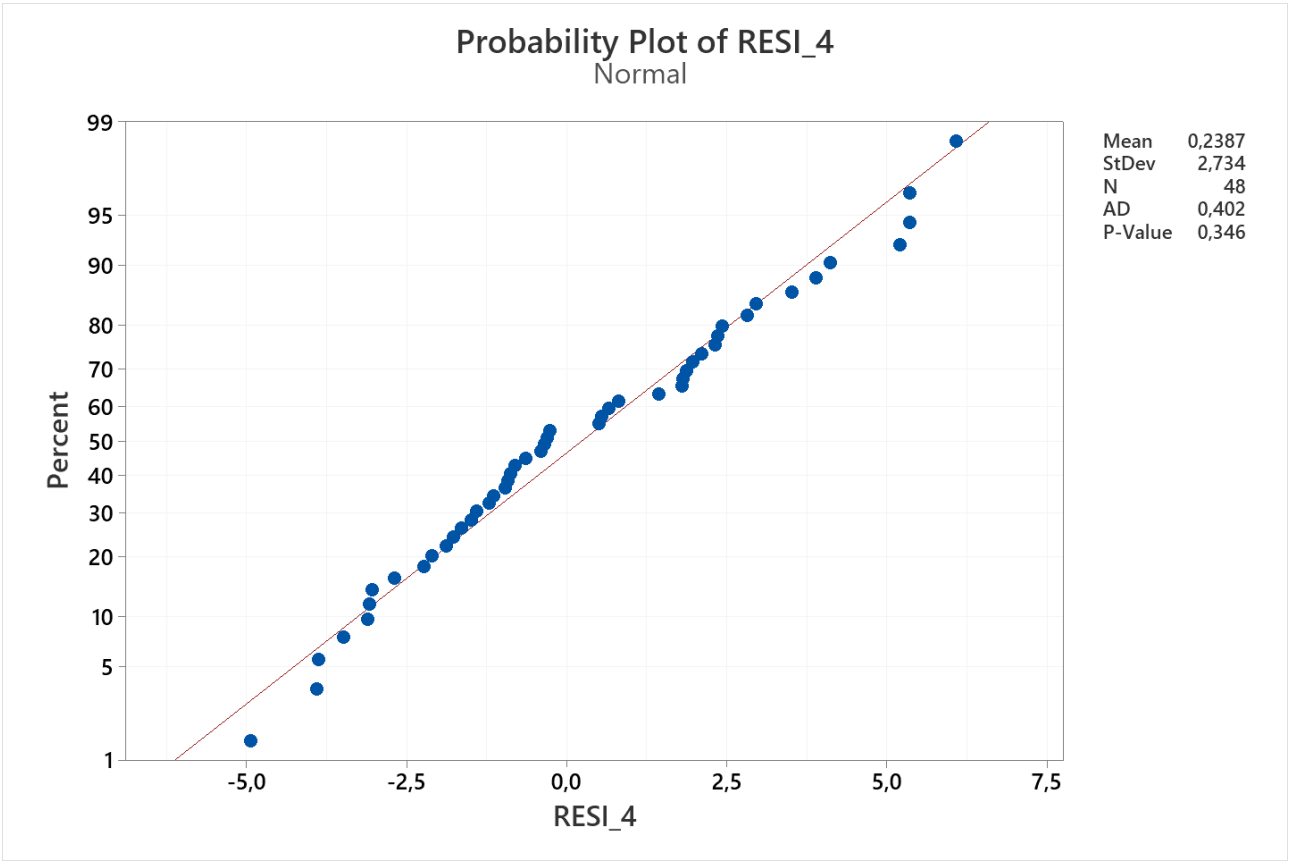
Model Summary

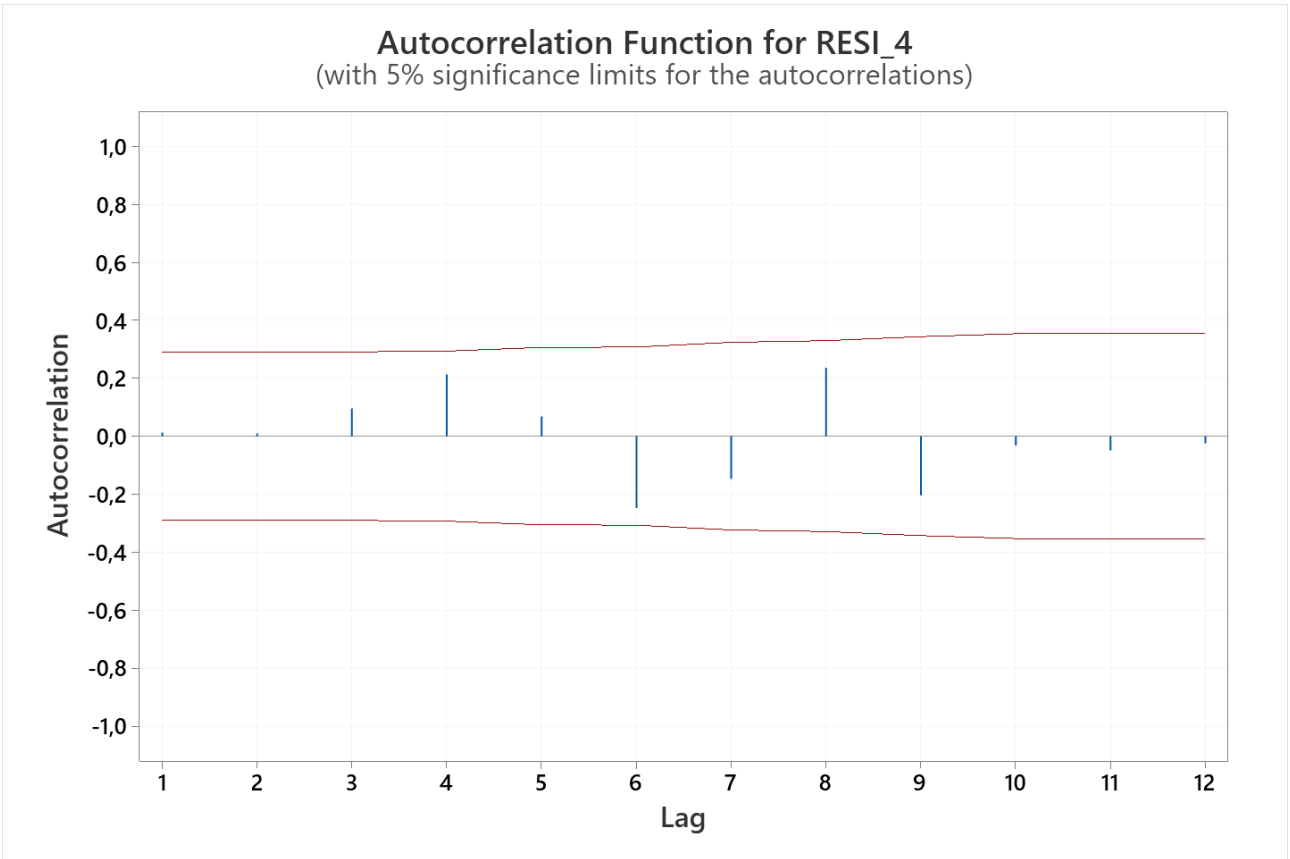
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2,77408	32,37%	29,43%	374,471	28,45%	238,67	243,74

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	169,41	32,37%	169,41	84,703	11,01	0,000
AR1	1	130,32	24,90%	147,23	147,227	19,13	0,000
dummy	1	39,09	7,47%	39,09	39,087	5,08	0,029
Error	46	353,99	67,63%	353,99	7,696		
Total	48	523,40	100,00%				

Check of residuals:


Probability Plot of RESI_4
Normal
Mean 0,2387
StDev 2,734
N 48
AD 0,402
P-Value 0,346
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-5,0
-2,5
0,0
2,5
5,0
7,5
RESI_4


Autocorrelation Function for RESI_4
(with 5% significance limits for the autocorrelations)
Autocorrelation
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,6
-0,8
-1,0
1
2
3
4
5
6
7
8
9
10
11
12
Lag

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
29	24,83	0,221

The residuals are normal and independent. The model is adequate.

b)

The 95% prediction interval for the differenced time series for observation 51 is the following:

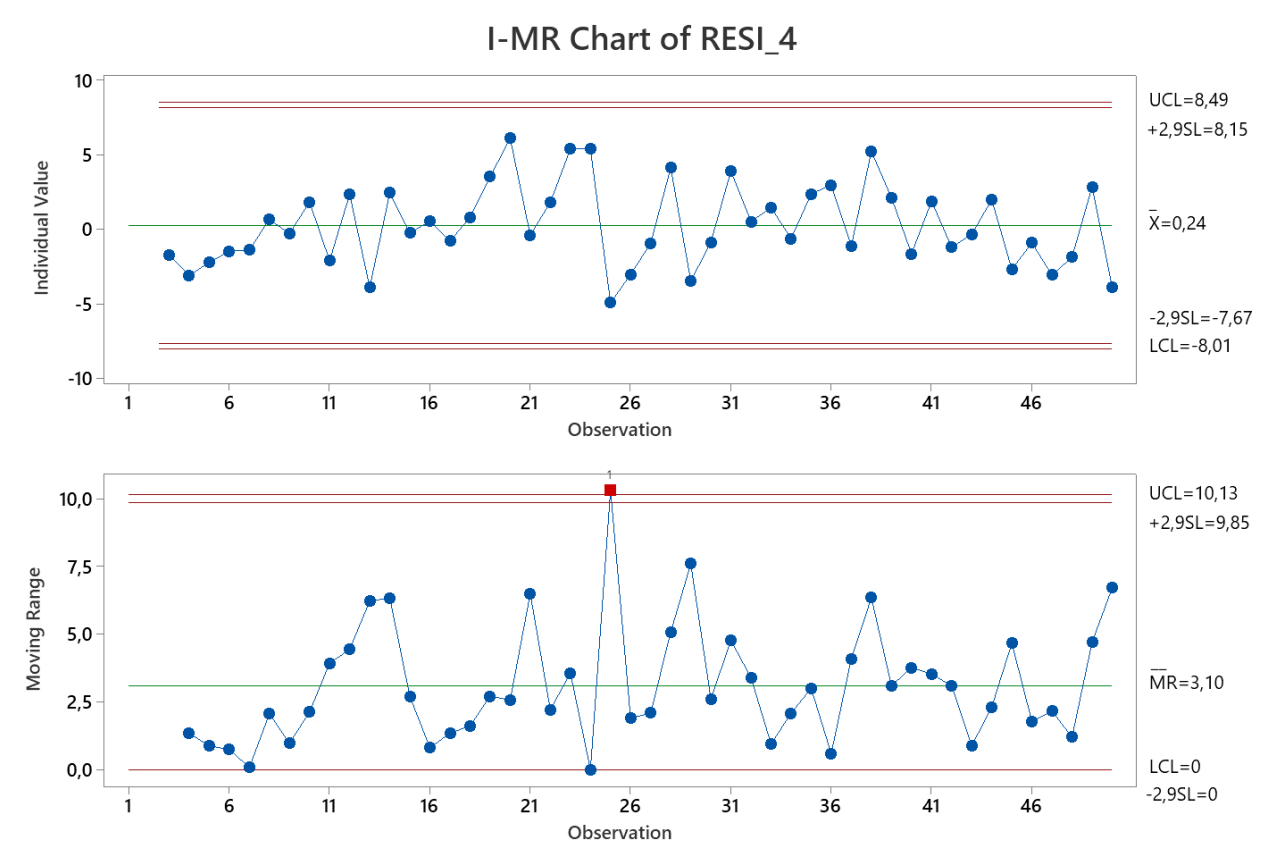
95% PI
(-2,83103; 8,65381)

This is a prediction interval on the differenced data. To obtain the prediction interval on the original data (contaminant concentration in ppm) we must sum the value of the variable at the 50th sample, i.e., $Y = 10{,}95$, thus:

$$8.119\ ppm \leq Y \leq 19.604\ ppm$$

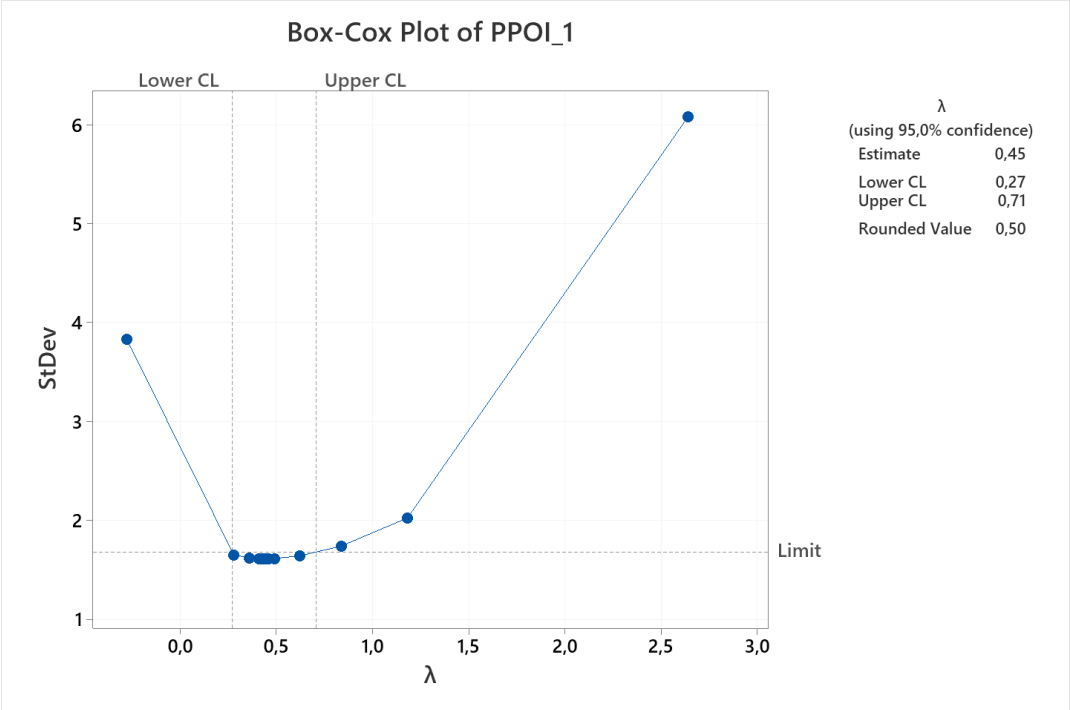
c)

The Type I error corresponding to $ARL_0 = 250$ is $\alpha = 0{,}004$, which corresponds to $k = z_{\alpha/2} = 2{,}878$. The resulting I-MR control chart for the model residuals is the following:

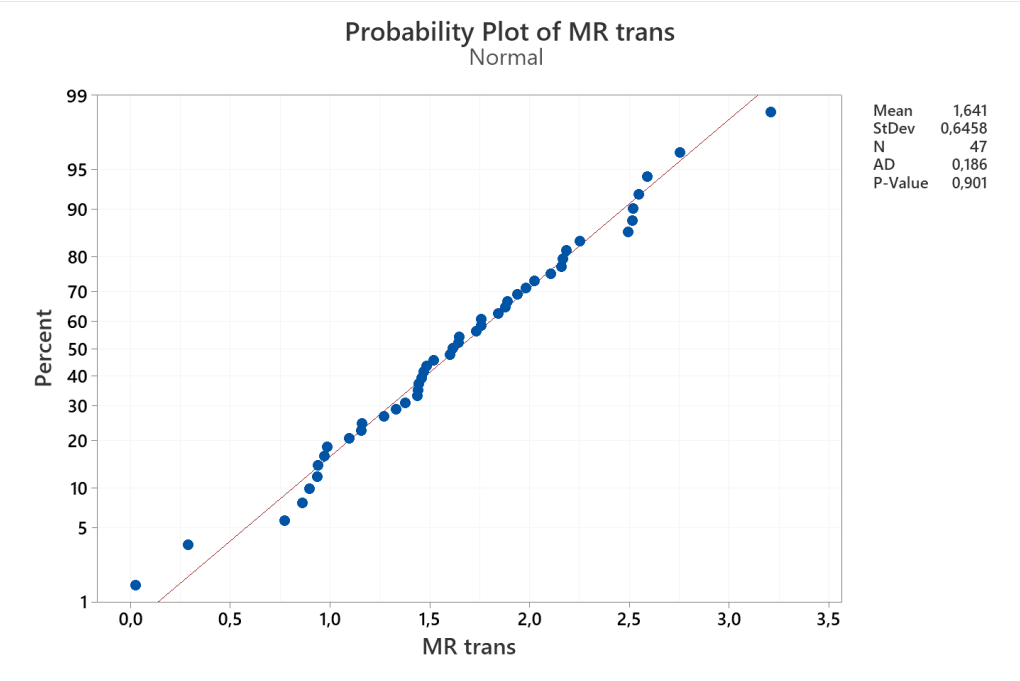


Sample 25 yields an OOC in the MR control chart. It is possible to verify if this OOC is the consequence of a violation of assumptions in the MR chart. One possible way is to transform MR data to normality and redesign the chart as follows:

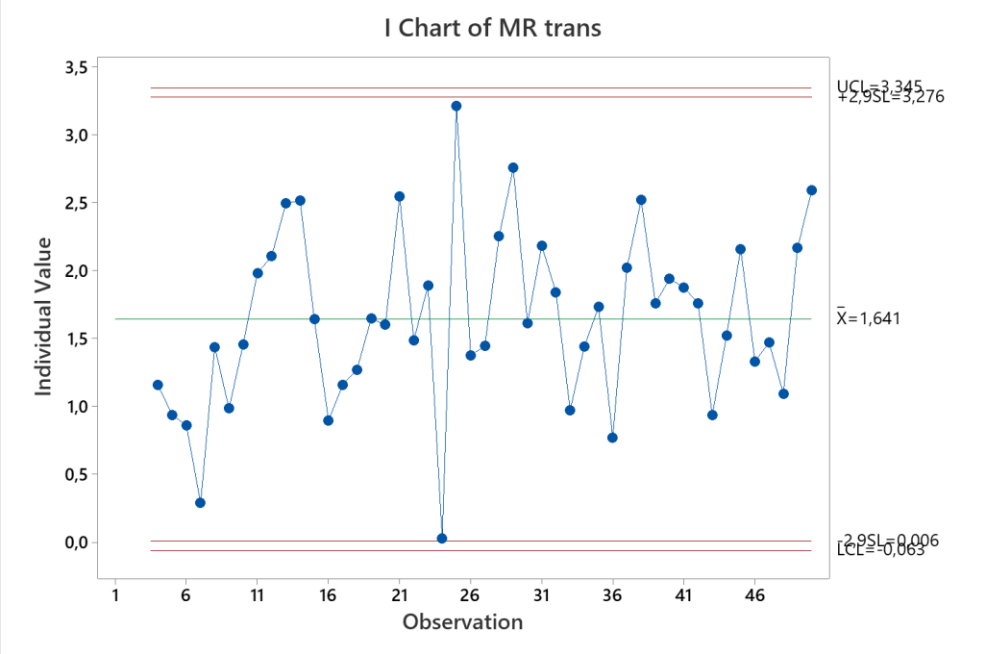
Box-Cox transformation:



Normality of MR statistic after transformation:



New MR control chart:



The OOC in the MR control chart was caused by a violation of assumptions of the chart itself.

The process is in-control.

d)

Since model residuals are normal and independent, it is possible to perform a one sample chi-squared test as follows.

By estimating the standard deviation of the model residuals as $\hat{\sigma}_\varepsilon = \sqrt{MSE} = 2.774$.

The test is such that:

$H_0: \sigma_\varepsilon = 2.5$

$H_1: \sigma_\varepsilon > 2.5$

The test statistic is $X^2 = \frac{(n-p)\hat{\sigma}_\varepsilon{}^2}{\sigma_\varepsilon{}^2} \sim X^2_{n-p}$, where $p = 2$ is the number of model terms, and $n - p = 46$.

Under H_0 we get $X^2 = 56.636$. The corresponding p-value is 0.135.

At 95% confidence, the standard deviation of residuals of the model fitted in point a) is not statistically larger than the one observed on historical data.

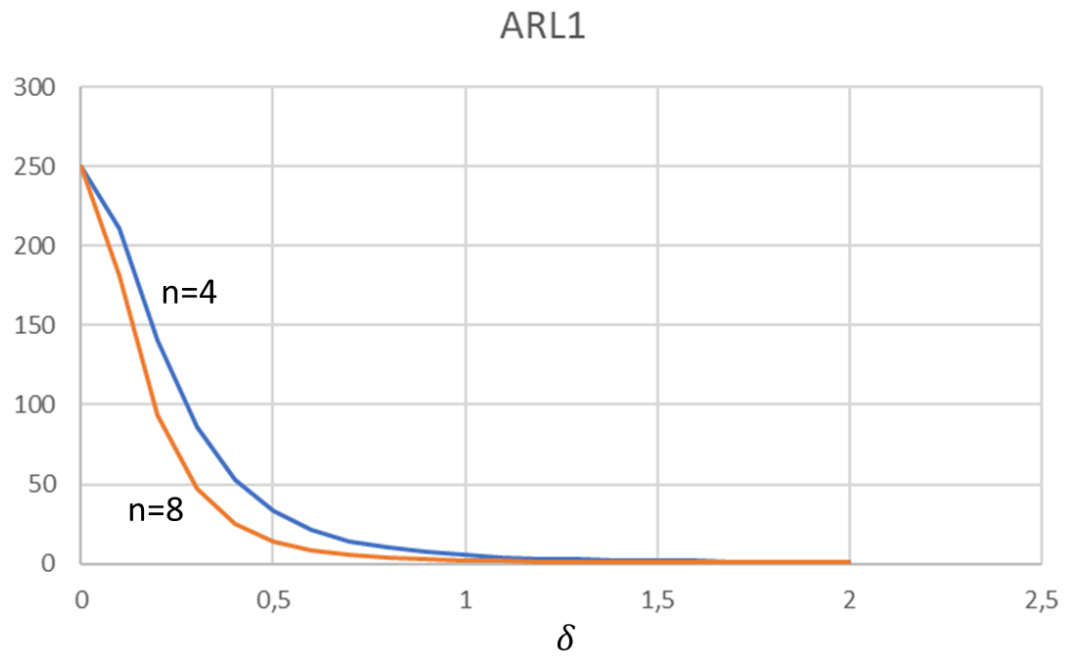
Exercise 2

The value of $K = z_{\alpha/2}$ with $\alpha = \frac{1}{250} = 0.004$ is: $K = 2.878$.

The Type II error as a function of the mean shift in standard deviation units is given by:

$\beta = \Pr(Z \leq K - \delta\sqrt{n}) - \Pr(Z \leq -K - \delta\sqrt{n})$, where $\delta = \frac{\mu_1 - \mu_0}{\sigma}$

Being, $ARL_1(\delta) = \frac{1}{1-\beta}$. The $ARL_1(\delta)$ curves for $n = 4$ and $n = 8$ are the following:



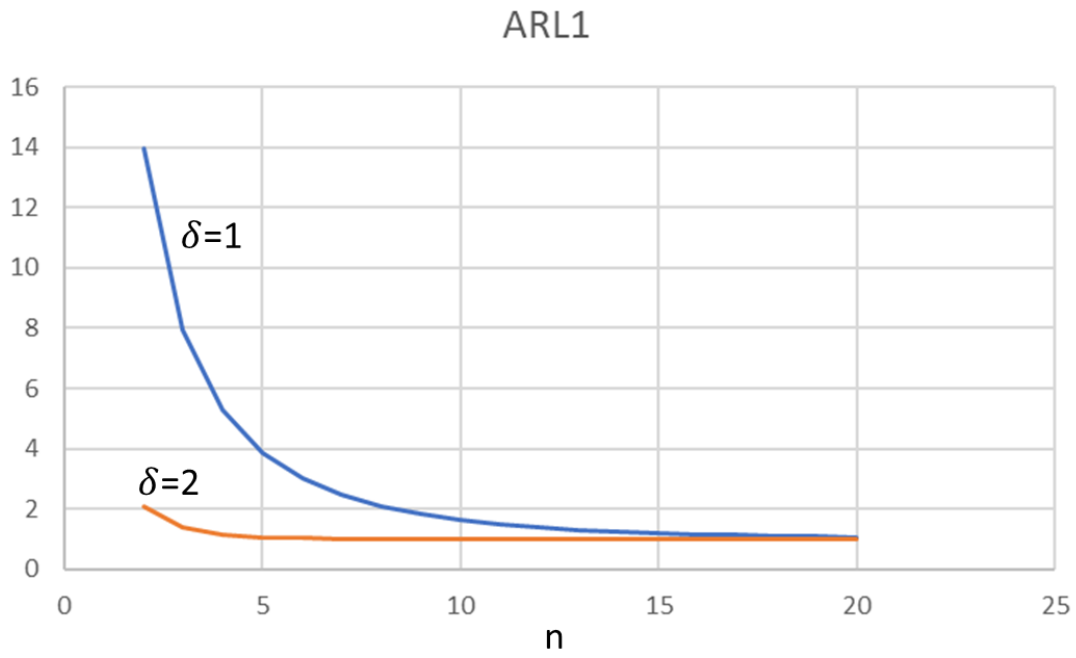
	$\delta = 1$	$\delta = 2$
ARL_1 with n=4	5.26	1.15
ARL_1 with n=8	2.08	1.00

b)

Being fixed δ, the type II error can be estimated as a function of n with the same expression used in the previous case:

$$\beta = \Pr\left(Z \leq K - \delta\sqrt{n}\right) - \Pr\left(Z \leq -K - \delta\sqrt{n}\right)$$

The resulting $ARL_1(n)$ curves for the two given mean shifts are the following:



	$n = 3$	$n = 6$
ARL_1 with $\delta = 1$	7.94	2.99
ARL_1 with $\delta = 2$	1.39	1.02

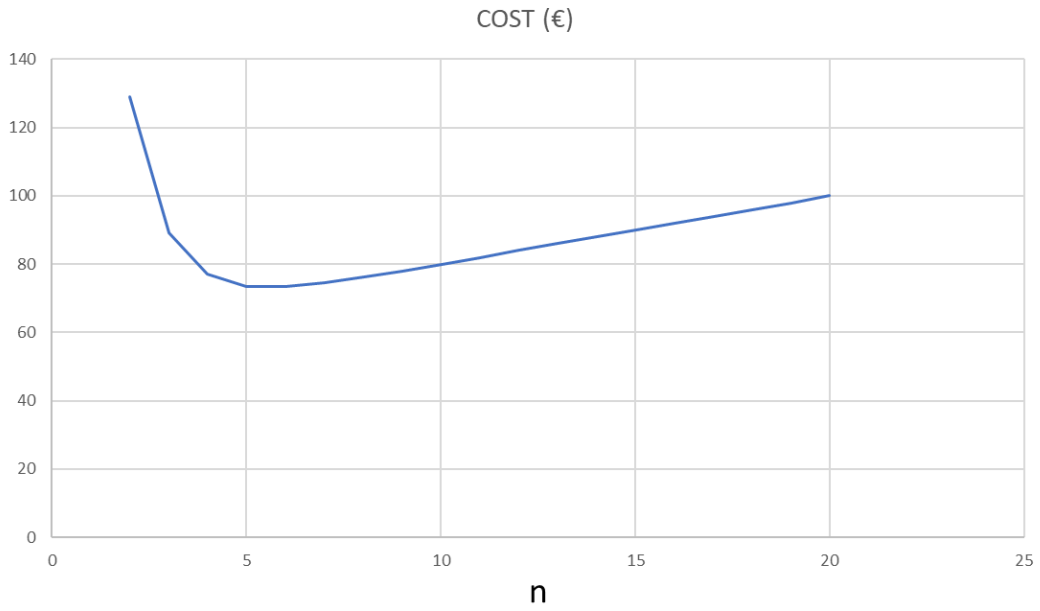
c)

The function to be minimized is the following:

$$C(n) = C1 * n + C2 * ATS(n) = 2 * n + 15 * ATS(n)$$

Where ATS = $h \cdot ARL_1$, where h is the time between the collection of two consecutive samples, i.e., $h = 4\ h$.

The cost function for $\delta = 2$ is shown below:



The late detection cost predominates at smaller values of n, whereas the inspection cost predominates at larger values of n. The optimal values of the sample size is n=6.

Exercise 3

The estimated slope b_1 is a random variable such that:

$E(b_1) = \beta_1, V(b_1) = \frac{\sigma_\varepsilon^2}{S_{xx}}$ where:

- σ_ε^2 is the variance of the normal error term
- $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

By using the Shewhart's scheme and assuming known parameters, the control chart for b_1 can be designed as follows:

$$UCL = \beta_1 + z_{\alpha/2}\sqrt{\frac{\sigma_\varepsilon^2}{S_{xx}}}$$

$$CL = \beta_1$$

$$LCL = \beta_1 - z_{\alpha/2}\sqrt{\frac{\sigma_\varepsilon^2}{S_{xx}}}$$

Where α is the Type I error.

The control charts can be used to monitor the stability over time of the calibration curves' slope for different sensors. It can be possibly combined with a control chart on $\hat{\sigma}_\varepsilon^2$, to monitor the model residuals as well.

QUALITY DATA ANALYSIS

30/08/2022

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
- **Multichance students should skip: point a) in Exercise 1, point b) in Exercise 3**

Exercise 1 (15 points)

A company produces aluminum laminates. The quality control department has recently introduced a statistical monitoring tool to keep under control the planarity of the laminates. It consists of an $\bar{X}$ control chart designed such that the number of samples before a false alarm is equal to 300.

a) Estimate and draw the curves of ARL_1 as a function of the mean shift δ expressed in standard deviation units with a sample size $n = 5$ and $n = 10$, respectively (show the two curves for $\delta \in [0\ 2]$ and report the ARL_1 values for $\delta = 1$ and $\delta = 2$).

b) Estimate and draw the curves of ARL_1 as a function of the sample size n for two values of the shift, $\delta = 1$ and $\delta = 2$, where δ is expressed in standard deviation units (show the two curves for $n \in [2\ 20]$ and report the ARL_1 values for $n = 6$ and $n = 12$).

c) The head of the quality control department is interested in selecting an optimal sample size n to minimize the lack of quality costs in the presence of a mean shift equal to $\delta = 2$ standard deviation units. Knowing that samples are gathered every 4 hours, the cost of planarity measurements for each laminate is $C_1 = 2$ € and an extra cost equal to $C_2 = 15$ € is due for each hour spent in the out-of-control state, determine the optimal sample size that minimizes the overall expected costs (assume the cost of the process in its in-control state as a reference baseline). Discuss the results.

Exercise 2 (3 points)

A company that produces thermal cameras is interested in monitoring the calibration curves of their devices. The calibration curve can be modelled by a linear model $y = \beta_0 + \beta_1 x + \varepsilon_t$ where the regressor x is the infrared counts measured by the sensor, whereas y is the temperature shown as output by the camera. All calibration curves are generated by using the same infrared counts levels for the regressor; moreover, the intercept $\hat{\beta}_0 = b_0$ and the slope $\hat{\beta}_1 = b_1$ are estimated using ordinary least squares.

Assuming $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, and assuming that β_0 and β_1 and σ_ε^2 are known, write down the expression of the control limits of a control chart for monitoring the slope of calibration curves.

Exercise 3 (15 points)

The concentration of a contaminant (measured in ppm) in the production of synthetic rubber is monitored over time. Table 1 shows the measurements collected in 50 consecutive samples.

Table 1

Sample	Concentration	Sample	Concentration
1	23,41	26	24,72
2	35,62	27	21,11
3	24,08	28	34,58
4	21,62	29	17,56
5	16,72	30	24,3
6	15,2	31	31,56
7	12,1	32	29,04
8	15,62	33	34,44
9	12,85	34	29,74
10	19,43	35	38,89
11	9,97	36	42,22
12	21,59	37	37,24
13	4,4	38	54,52
14	20,5	39	51,1
15	11,06	40	48,33
16	17,67	41	55,08
17	-3,64	42	48,02
18	-3,98	43	50,88
19	-3,88	44	54,85
20	20,66	45	45,16
21	13,24	46	47,82
22	22,29	47	37,77
23	32,42	48	37,94
24	41,97	49	45,75
25	22,99	50	30,66

a) Being known that a negative value is the result of a temporary miscalibration of the measuring device, fit a suitable model to these data;
b) Based on the result of point a), estimate the 95% prediction interval for the contaminant concentration in the next sample.
c) Based on the result of point a), design an appropriate control chart for these data with $ARL_0 = 250$.
d) From historical data, it is known that the most appropriate model for this process yielded a standard deviation of residuals equal to $\sigma_\varepsilon = 8.0$. Determine, with a statistical test, if the model fitted at point a) is such that the standard deviation of residuals is greater than this value (report also the p-value of the test). Discuss the result.

Solutions

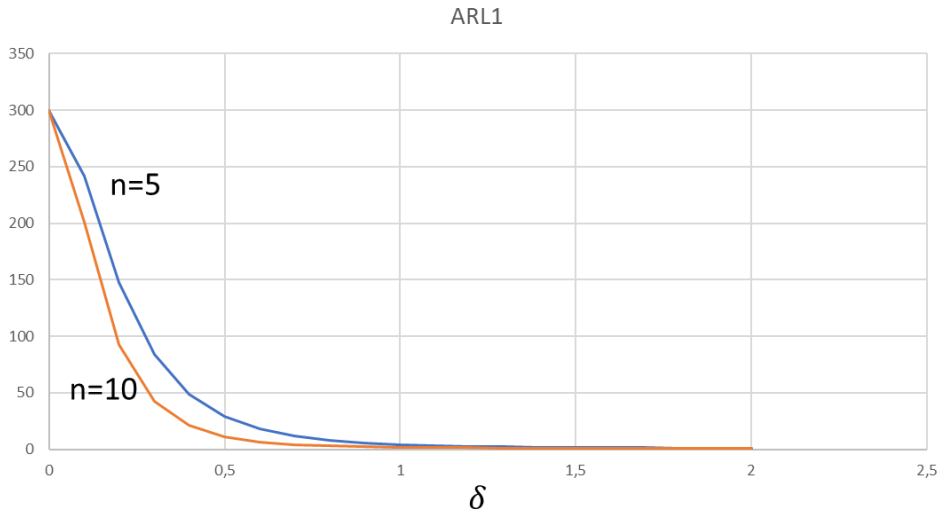
Exercise 1

The value of $K = z_{\alpha/2}$ with $\alpha = \frac{1}{300} = 0.0033$ is: $K = 2.935$.

The Type II error as a function of the mean shift in standard deviation units is given by:

$\beta = \Pr(Z \leq K - \delta\sqrt{n}) - \Pr(Z \leq -K - \delta\sqrt{n})$, where $\delta = \frac{\mu_1 - \mu_0}{\sigma}$

Being, $ARL_1(\delta) = \frac{1}{1-\beta}$. The $ARL_1(\delta)$ curves for $n = 5$ and $n = 10$ are the following:



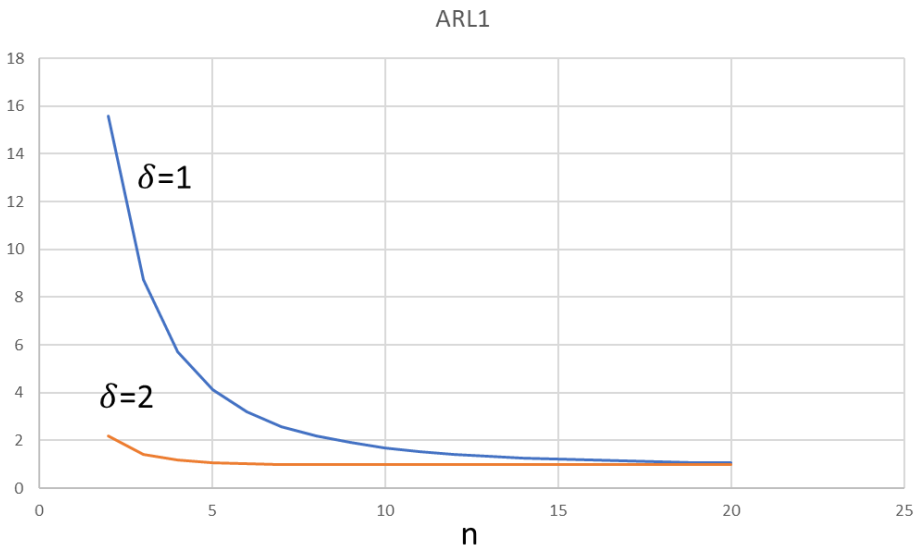
	$\delta = 1$	$\delta = 2$
ARL_1 with n=5	4.13	1.07
ARL_1 with n=10	1.70	1.00

b)

Being fixed δ, the type II error can be estimated as a function of n with the same expression used in the previous case:

$\beta = \Pr(Z \leq K - \delta\sqrt{n}) - \Pr(Z \leq -K - \delta\sqrt{n})$

The resulting $ARL_1(n)$ curves for the two given mean shifts are the following:



	$n = 6$	$n = 12$
ARL_1 with $\delta = 1$	3.19	1.03
ARL_1 with $\delta = 2$	1.43	1.00

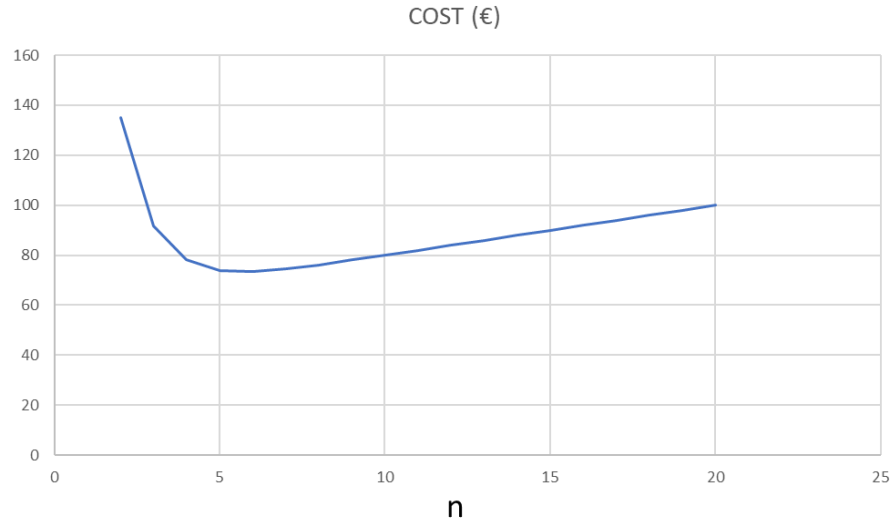
c)

The function to be minimized is the following:

$$C(n) = C1 * n + C2 * ATS(n) = 2 * n + 15 * ATS(n)$$

Where ATS = $h \cdot ARL_1$, where h is the time between the collection of two consecutive samples, i.e., $h = 4\ h$.

The cost function for $\delta = 2$ is shown below:



The late detection cost predominates at smaller values of n, whereas the inspection cost predominates at larger values of n. The optimal values of the sample size is n=6.

Exercise 2

The estimated slope b_1 is a random variable such that:

$E(b_1) = \beta_1, V(b_1) = \frac{\sigma_\varepsilon^2}{S_{xx}}$ where:

- σ_ε^2 is the variance of the normal error term
- $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

By using the Shewhart's scheme and assuming known parameters, the control chart for b_1 can be designed as follows:

$$UCL = \beta_1 + z_{\alpha/2}\sqrt{\frac{\sigma_\varepsilon^2}{S_{xx}}}$$

$$CL = \beta_1$$

$$LCL = \beta_1 - z_{\alpha/2}\sqrt{\frac{\sigma_\varepsilon^2}{S_{xx}}}$$

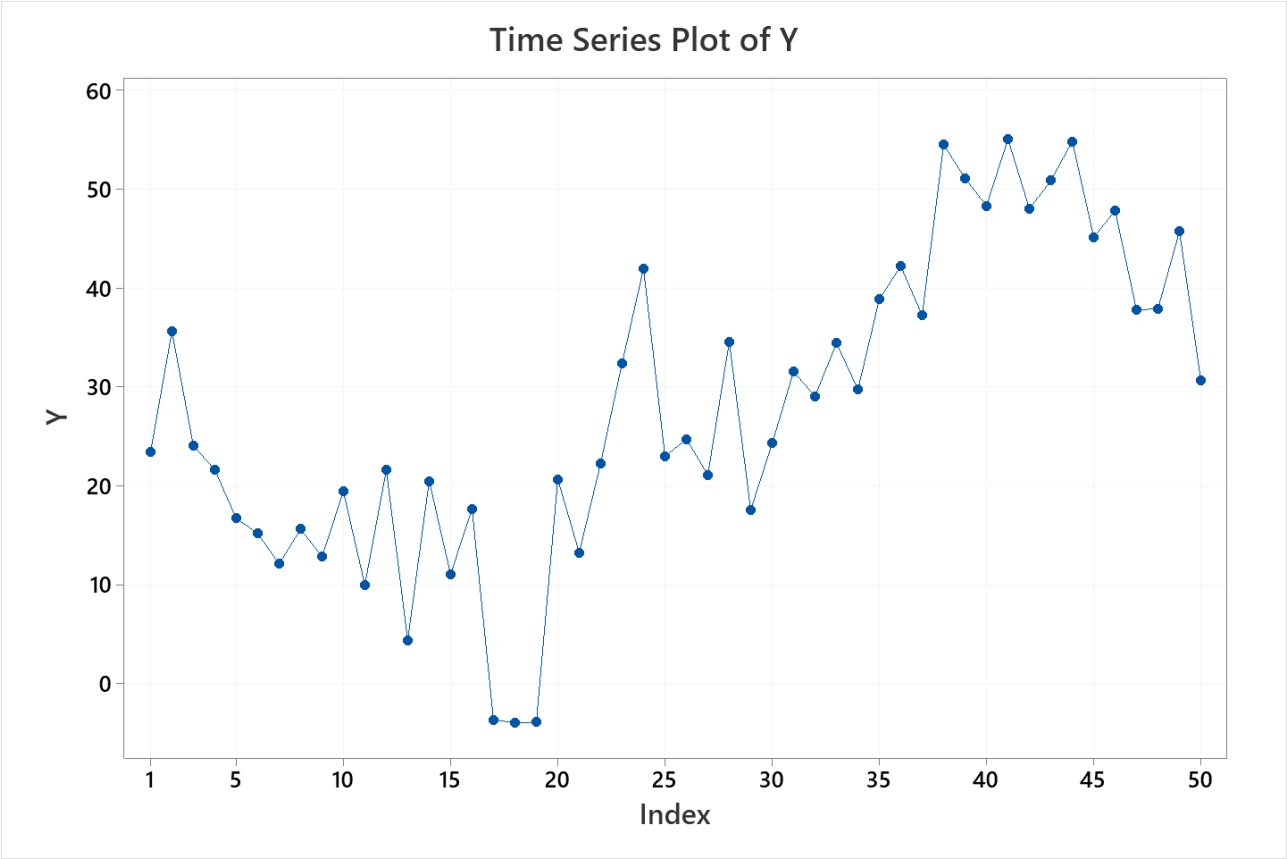
Where α is the Type I error.

The control charts can be used to monitor the stability over time of the calibration curves' slope for different sensors. It can be possibly combined with a control chart on $\hat{\sigma}_\varepsilon^2$, to monitor the model residuals as well.

Exercise 3

a)

Time series plot of the temperature series:



It is present a meandering pattern. Negative values were observed in sample 17, 18 and 19.

Runs test: null hypothesis is not accepted:

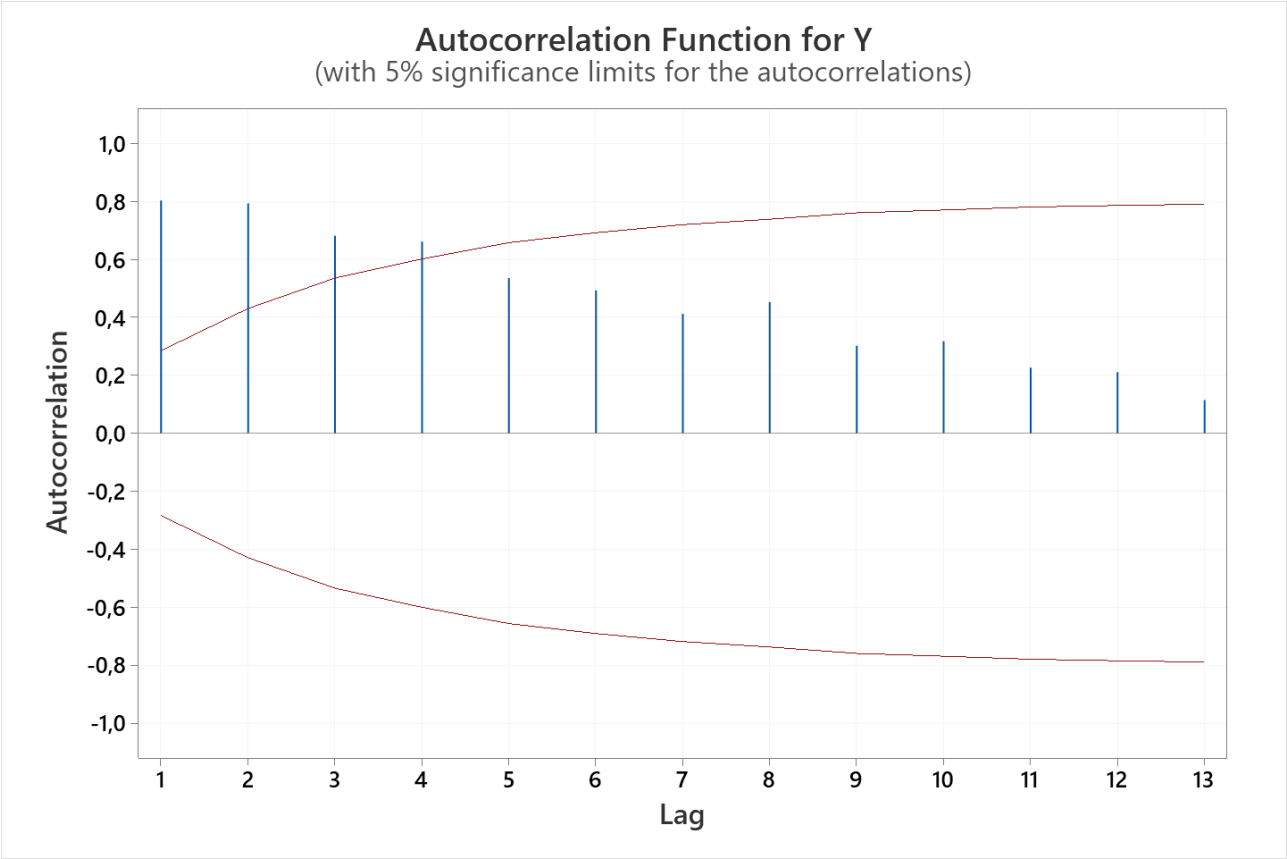
Test

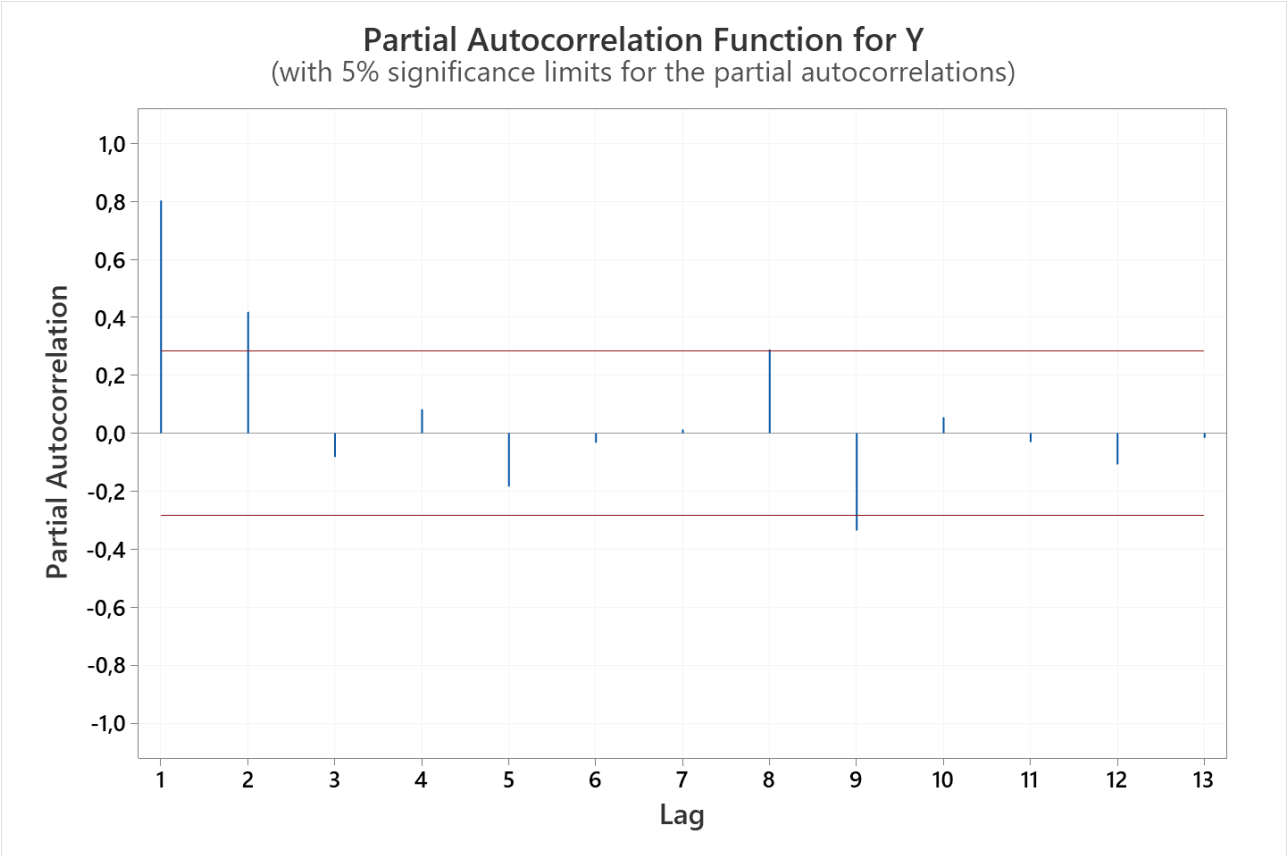
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs

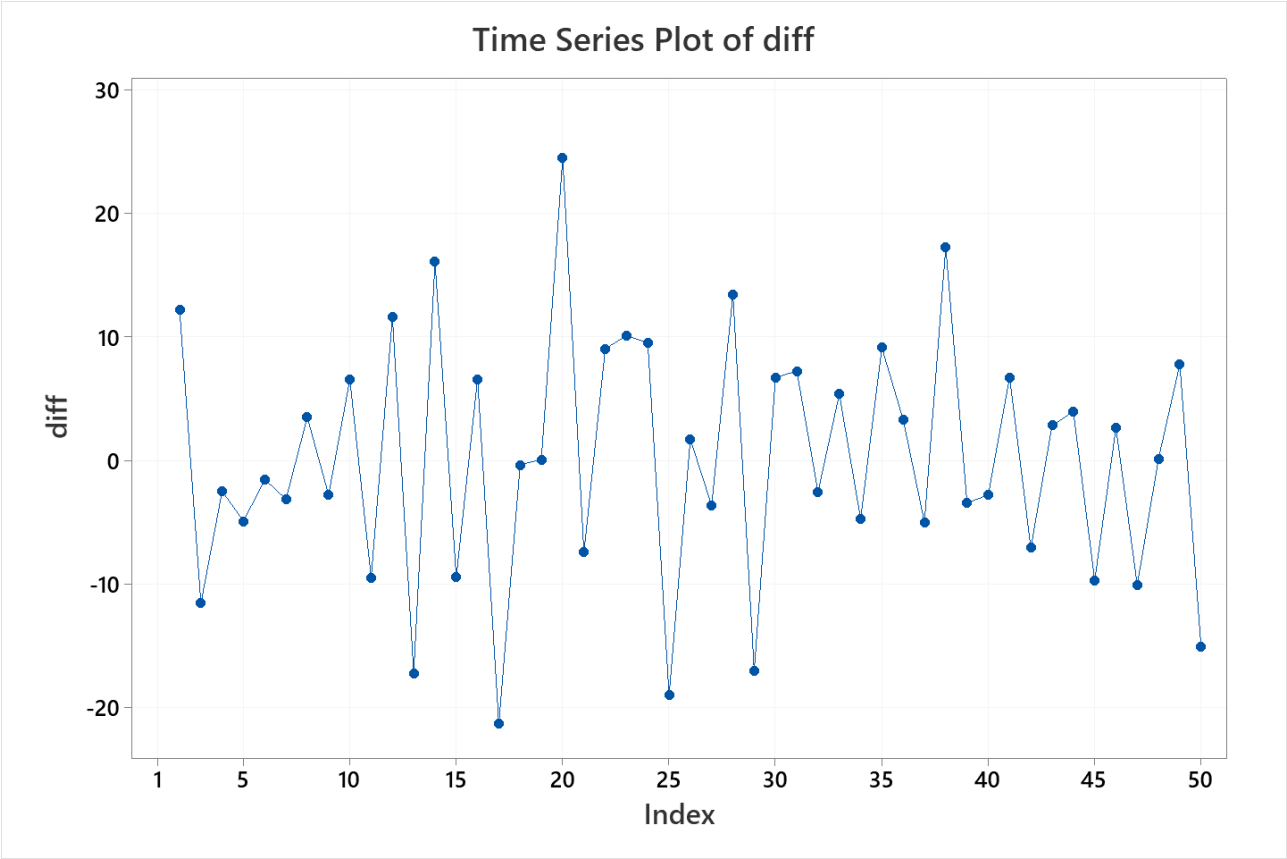
Observed	Expected	P-Value
8	25,96	0,000

Sample autocorrelation and partial autocorrelation functions:

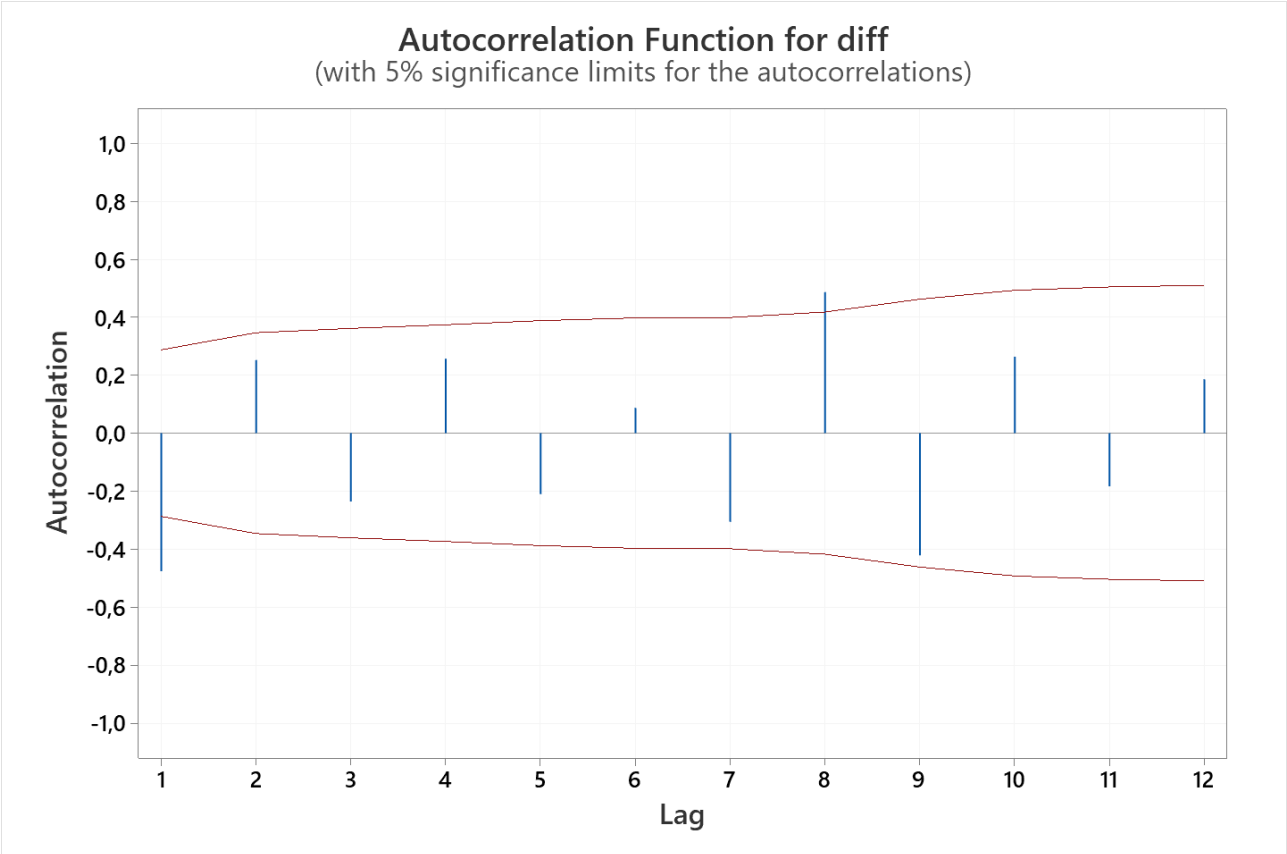


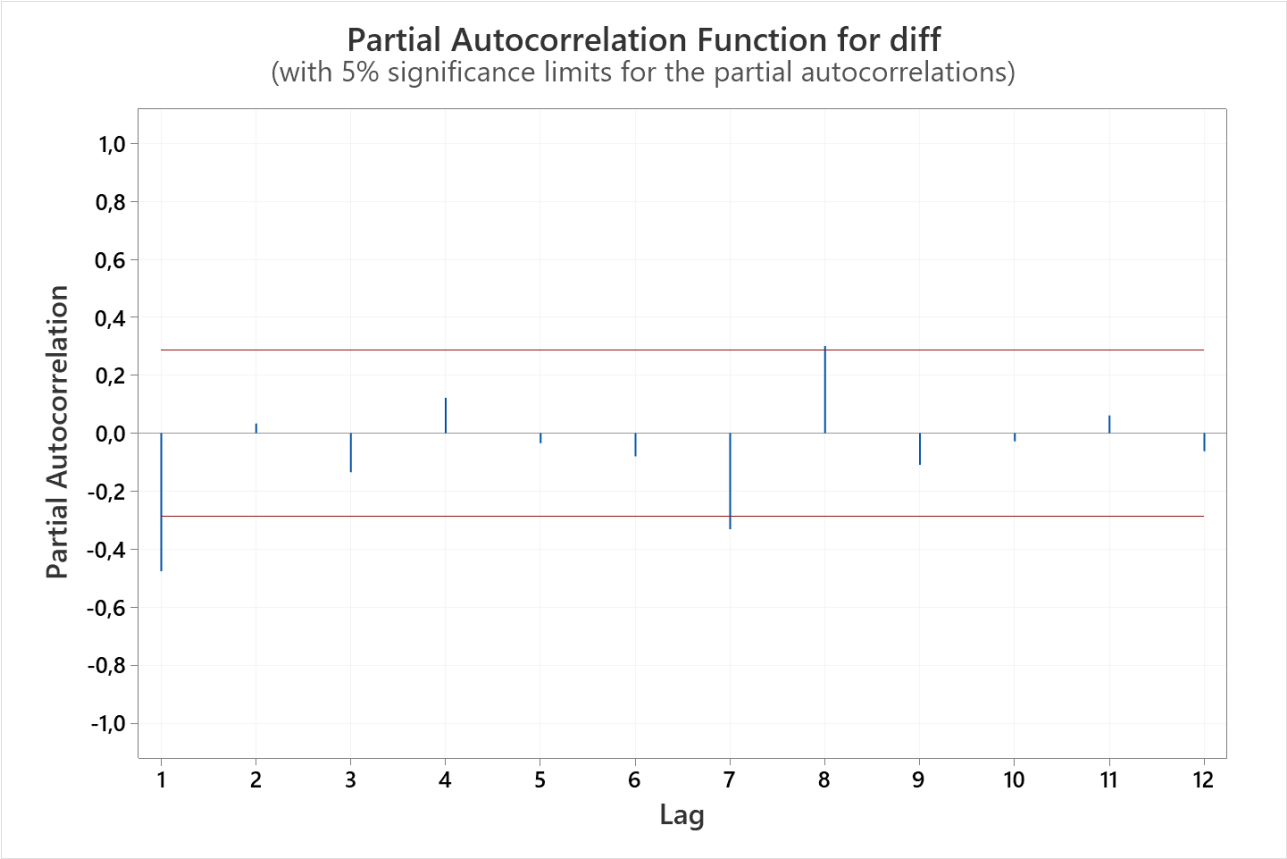


A slow decay of the SACF is present, which suggests a non-stationarity of the process. By differencing the timeseries we get:



The SACF and SPACF of the data after the differencing operation are the following:





A suitable model for the temperature time series is therefore an ARIMA(1,1,0). However, we should keep in mind that three negative values are present, caused by a temporary miscalibration of the sensor. Thus, a dummy variable that is equal to 1 for these three samples and 0 for all other samples can be included in the model.

VERSIONE 2

Regression Analysis: diff versus AR1; dummy

Method

Categorical predictor coding	(1; 0)
Rows unused	2

Regression Equation

dummy	
0	diff = 0,82 - 0,539 AR1
1	diff = -9,89 - 0,539 AR1

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	0,82	1,25	(-1,70; 3,34)	0,65	0,517	
AR1	-0,539	0,126	(-0,793; -0,285)	-4,28	0,000	1,02
dummy						
1	-10,70	5,04	(-20,86; -0,55)	-2,12	0,039	1,02

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
8,36277	31,41%	28,36%	3615,48	21,20%	345,93	352,49

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	1440,9	31,41%	1440,9	720,45	10,30	0,000
AR1	1	1125,6	24,53%	1280,5	1280,49	18,31	0,000
dummy	1	315,3	6,87%	315,3	315,33	4,51	0,039
Error	45	3147,1	68,59%	3147,1	69,94		
Total	47	4588,0	100,00%				

The constant term is not significant, thus we may remove it:

VERSIONE 2

Regression Analysis: diff versus AR1; dummy

Method

Categorical predictor coding	(1; 0)
Rows unused	2

Regression Equation

dummy	
0	diff = 0,0 - 0,532 AR1
1	diff = -9,85 - 0,532 AR1

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
AR1	-0,532	0,125	(-0,784; -0,281)	-4,27	0,000	1,02
dummy						
1	-9,85	4,84	(-19,59; -0,11)	-2,04	0,048	1,02

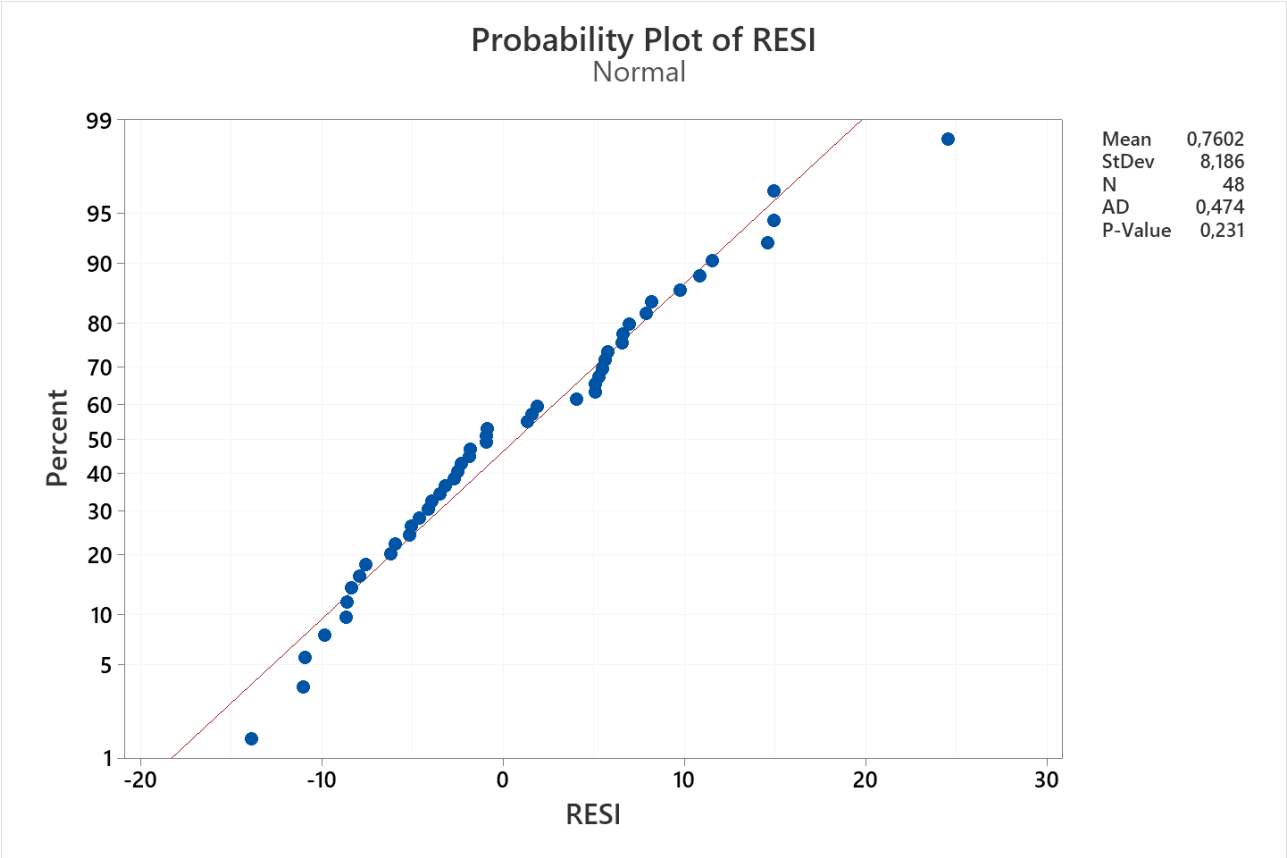
Model Summary

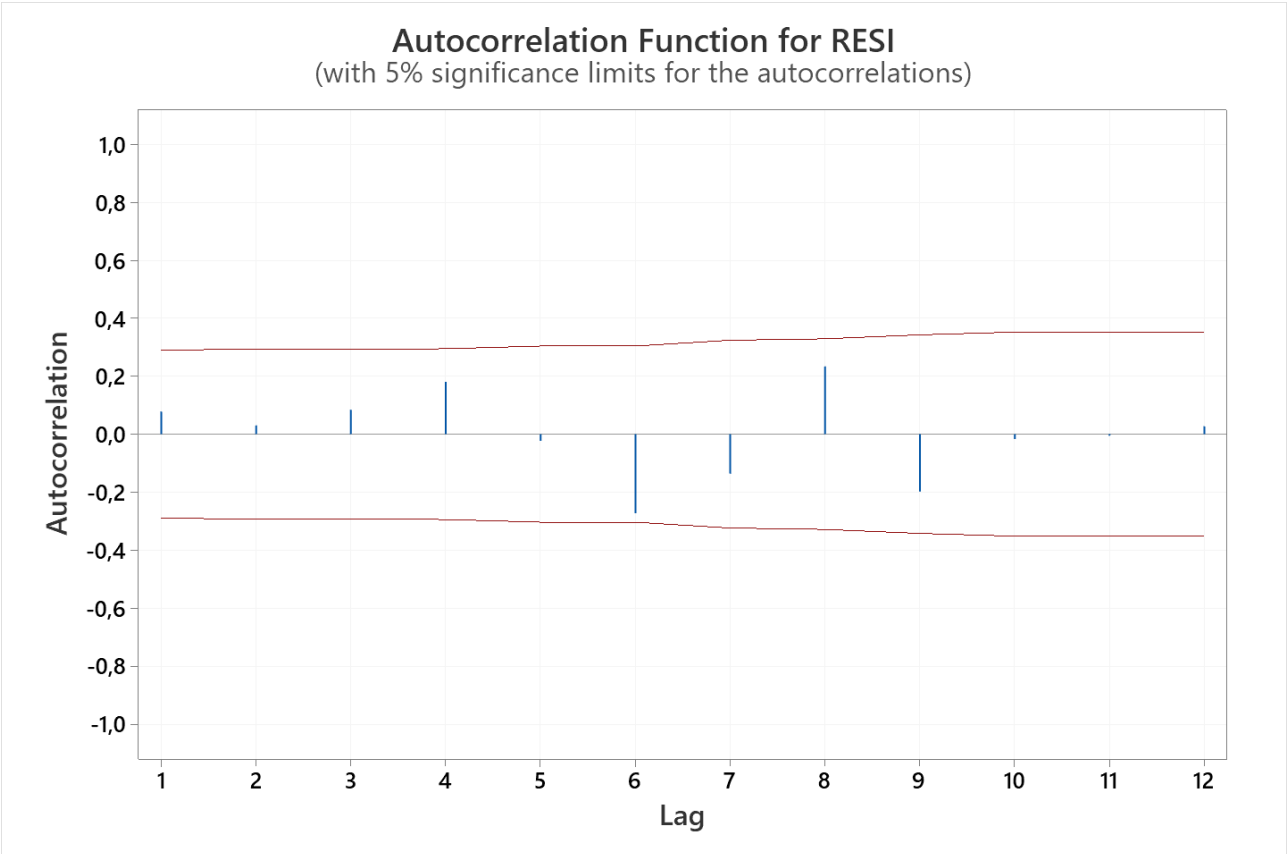
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
8,31044	30,76%	27,75%	3503,43	23,65%	344,00	349,07

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	1411,6	30,76%	1411,6	705,81	10,22	0,000
AR1	1	1125,3	24,52%	1256,8	1256,81	18,20	0,000
dummy	1	286,3	6,24%	286,3	286,33	4,15	0,048
Error	46	3176,9	69,24%	3176,9	69,06		
Total	48	4588,5	100,00%				

Check of residuals:





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
27	24,83	0,524

The residuals are normal and independent. The model is adequate.

b)

The 95% prediction interval for the differenced time series for observation 51 is the following:

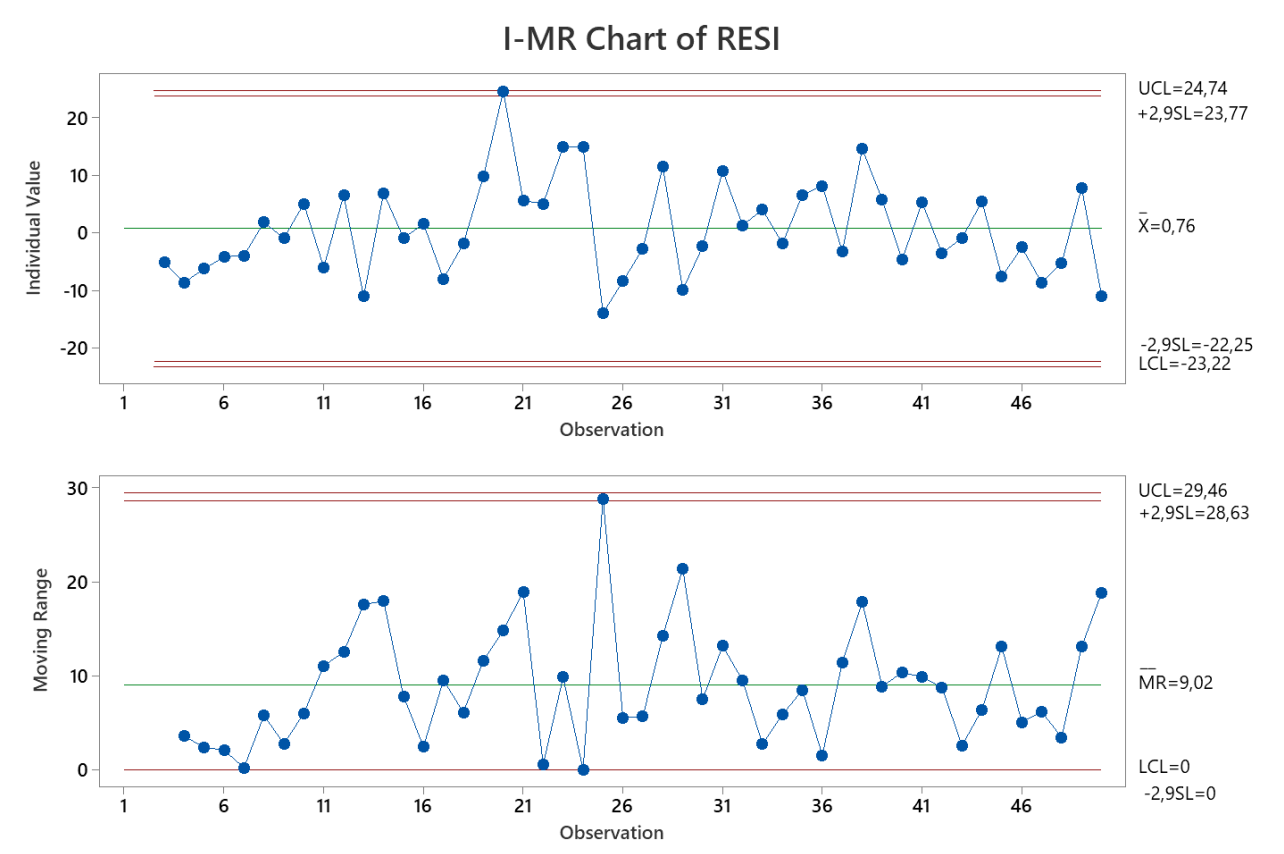
95% PI
(-9,11870; 25,1856)

This is a prediction interval on the differenced data. To obtain the prediction interval on the original data (contaminant concentration in ppm) we must sum the value of the variable at the 50th sample, i.e., $Y = 30{,}66$, thus:

$$21.541\ ppm\ \leq Y \leq 55.846\ ppm$$

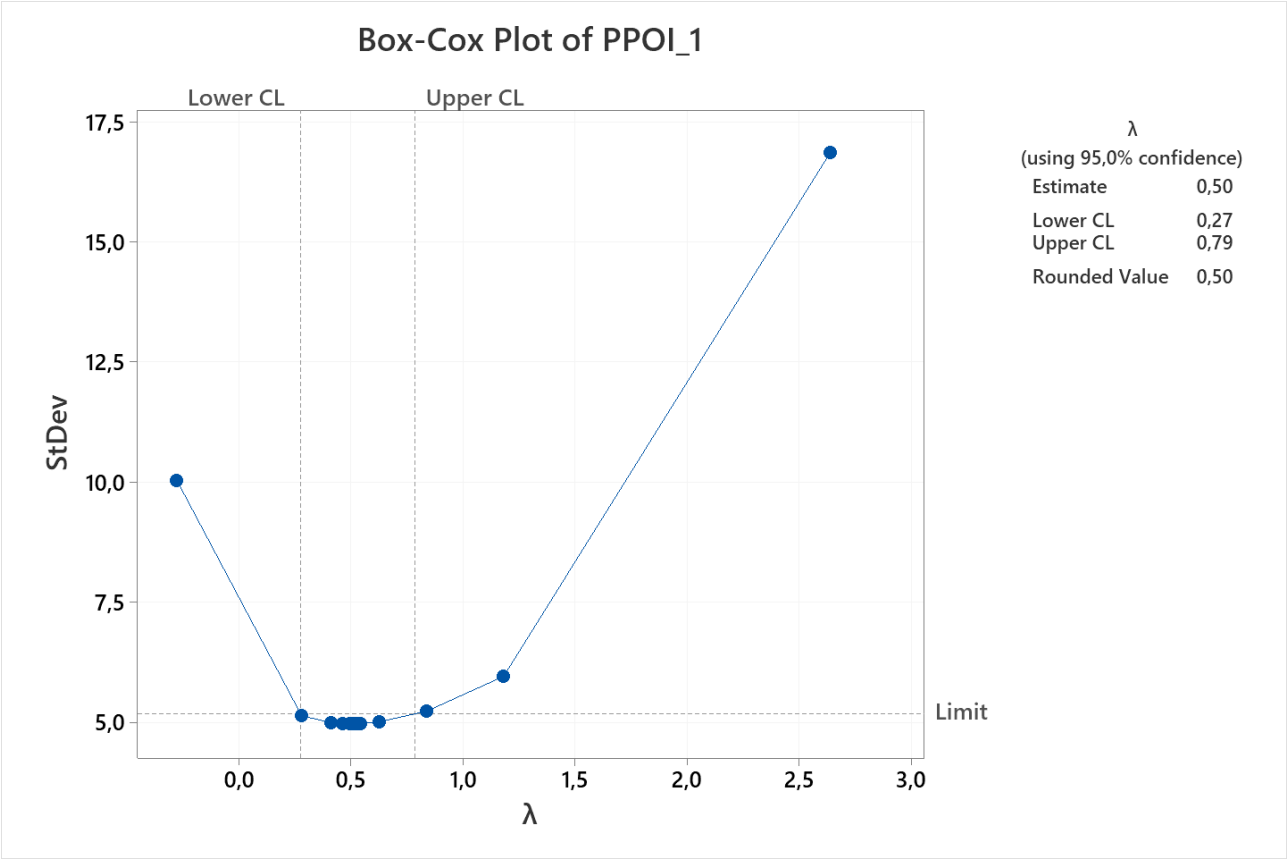
c)

The Type I error corresponding to $ARL_0 = 250$ is $\alpha = 0{,}004$, which corresponds to $k = z_{\alpha/2} = 2{,}878$. The resulting I-MR control chart for the model residuals is the following:

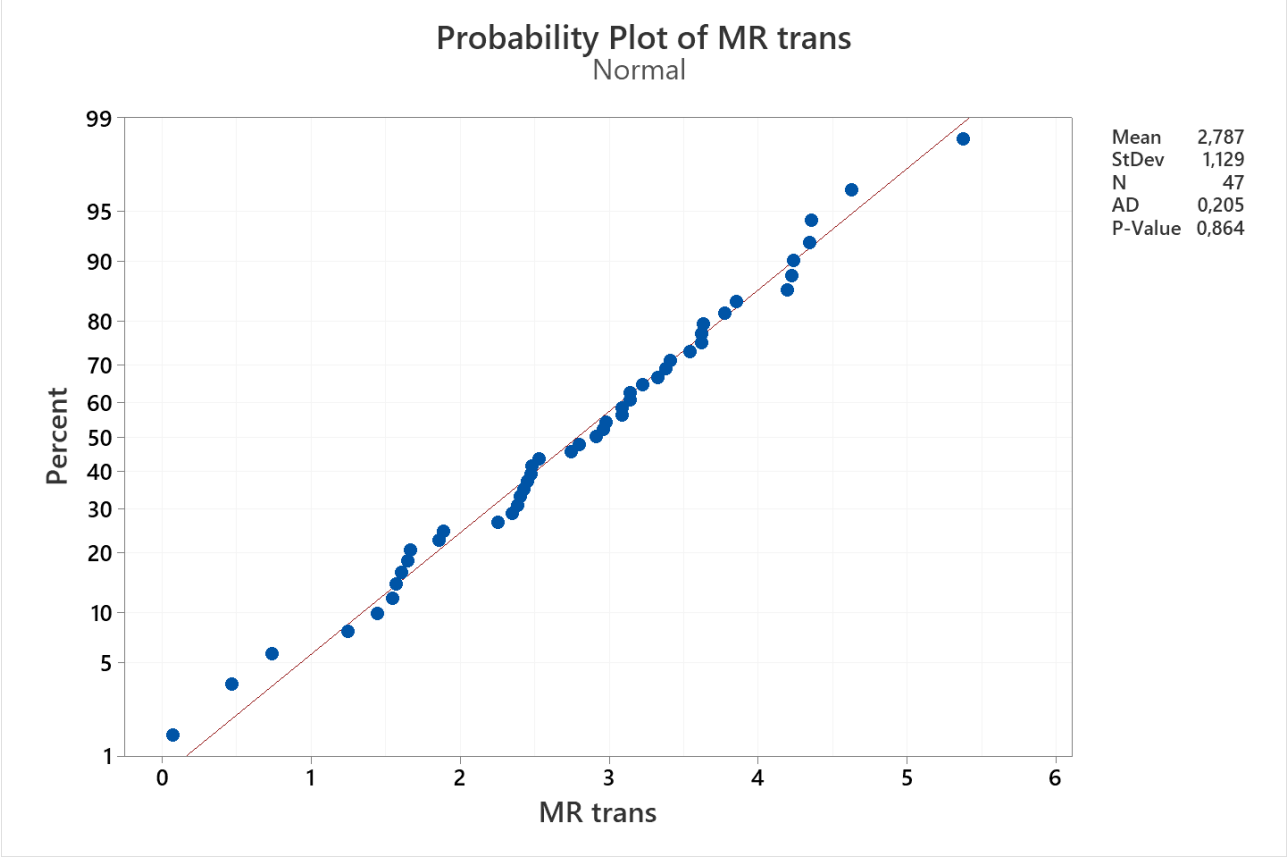


Sample 20 yields an OOC in the I control chart, whereas sample 25 yields an OOC in the MR control chart. It is possible to verify if this latter OOC is the consequence of a violation of assumptions in the MR chart. One possible way is to transform MR data to normality and redesign the chart as follows:

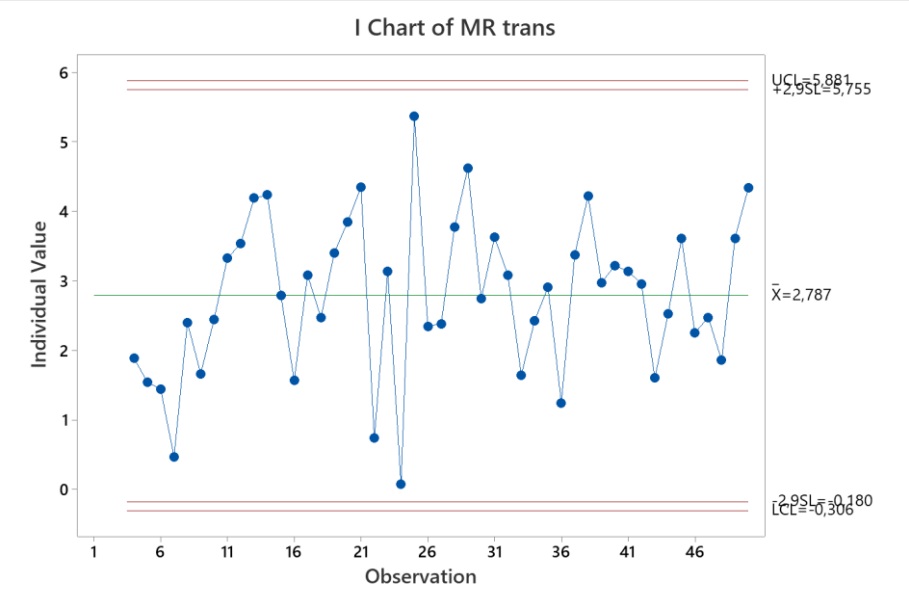
Box-Cox transformation:



Normality of MR statistic after transformation:



New MR control chart:



The OOC in the MR control chart was caused by a violation of assumptions of the chart itself. Regarding the OOC in the I control chart, in the absence of any information about an assignable cause we may deem it a false alarm. The control chart design is over.

d)

Since model residuals are normal and independent, it is possible to perform a one sample chi-squared test as follows.

By estimating the standard deviation of the model residuals as $\hat{\sigma}_\varepsilon = \sqrt{MSE} = 8.31$.

The test is such that:

$H_0: \sigma_\varepsilon = 8.0$

$H_1: \sigma_\varepsilon > 8.0$

The test statistic is $X^2 = \frac{(n-p)\hat{\sigma}_\varepsilon{}^2}{\sigma_\varepsilon{}^2} \sim X^2_{n-p}$, where $p = 2$ is the number of model terms, and $n - p = 46$.

Under H_0 we get $X^2 = 49.634$. The corresponding p-value is 0.331.

At 95% confidence, the standard deviation of residuals of the model fitted in point a) is not statistically larger than the one observed on historical data.