QUALITY DATA ANALYSIS

18/06/2021

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

Exercise 1 (15 points)

The energy consumption of a thermal treatment facility is known to follow a time series model $x_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon$, with $\varepsilon \sim N(0, \sigma_\varepsilon = 0.65)$. Based on historical process data, the unknown coefficients estimates are available $\widehat{\beta_0} = b_0$ and $\widehat{\beta_1} = b_1$ and one can assume that $E(b_0) = 9.5$, $V(b_0) = 6.5$, $E(b_1) = 0.64$, $V(b_1) = 0.0196$. Every month, 30 consecutive measurements (in KWh) are collected. The data collected in June and July 2020 are reported in Table 1[1].

Table 1

Month	Order	Measured value	Order	Measured value
June	1	26,01	16	26,47
	2	26,72	17	26,12
	3	25,24	18	26,64
	4	25,63	19	27,02
	5	25,66	20	27,32
	6	24,87	21	27,19
	7	24,68	22	26,15
	8	24,92	23	26,28
	9	26,72	24	26,84
	10	27,76	25	26,71
	11	26,54	26	26,89
	12	27,75	27	26,87
	13	27,56	28	26,35
	14	27,02	29	26,23
	15	26,97	30	25,59
Month	Order	Measured value	Order	Measured value
July	1	26,5	16	26
	2	23,74	17	24,64
	3	25,83	18	26,64
	4	25,47	19	25,14
	5	25,72	20	25,68
	6	25,6	21	24,6
	7	25,38	22	26,09
	8	25,67	23	25,03
	9	25,16	24	25,78
	10	26,13	25	25,25
	11	24,44	26	25,93
	12	26,36	27	24,84
	13	24,56	28	26,01
	14	25,9	29	25,02
	15	25,54	30	26,61

[1] An implicit use of conditional distribution is done, considering that the estimate relies on having observed the previous datapoint.

1) Determine if the energy consumption data in Table 1 are in-control or not by using a special cause control chart (use $ARL_0 = 371$).
2) The head of the quality department aims to replace the special cause control chart with a control chart designed to keep under control the estimated coefficients b_0 and b_1. To this aim, the model coefficients are estimated every month, i.e., the i-th values of b_0 and b_1 in the i-th month are estimated by fitting the model to the measurements acquired in the i-th month only. Design two control charts suitable to monitor the mean of the estimated model coefficients (assuming an overall $ARL_0 = 371$).
3) Fit two models for June's and July's data, respectively, and estimate their model coefficients.
4) Based on control charts designed in point 2) and using the model coefficients estimated in point 3), is the process in these two months in-control or not? Discuss and compare the pros and cons of these control charts compared with the control chart used in point 1).

Exercise 2 (15 points)

The mean of piston ring diameters is monitored by means of an $\bar{X}$ control chart. To this aim, a sample of piston rings is collected every 3 hours, and the $\bar{X}$ control chart is designed assuming $ARL_0 = 200$.

1) Estimate the curves of the average time to signal (ATS) 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 $\delta \in [0\ 4]$ and report the ATS values for $\delta = 1$ and $\delta = 2$).
2) Estimate the curves of the ATS as a function of the sample size n for two values of the shift, $\delta = 1$ and $\delta = 2$. (show the two curves for $n \in [2\ 20]$).
3) The head of the quality assurance department is interested in optimizing the process monitoring solution for two mean shifts of major interest, i.e., $\delta = 1$ and $\delta = 2$. Assuming the cost of a process in its in-control state as reference baseline, when the process mean shifts to an out-of-control state, an additional cost is due and equal to C1 = 1€ for each hour spent in the out-of-control state when $\delta = 1$ or equal to C1 = 6€ for each hour spent in the out-of-control state when $\delta = 2$. The inspection cost is C2 = 0.8€ for each measurement of the piston ring diameters. Compute the optimal value of the sample size, n, minimizing the overall expected costs when $\delta = 1$ and when $\delta = 2$, respectively. Discuss the results.

Exercise 3 (3 points)

A piston ring diameter is normally distributed $D \sim N(\mu_0, \sigma^2)$ with $\mu_0 = 20$ mm and σ=1 mm. The lower and upper specification limits are $LSL = 18$ mm and $USL = 22$ mm, respectively. The process produces 1 piston ring per minute and each conforming item costs 0.1€ while each nonconforming item costs 2€.

1) What are the expected costs per hour when the process is in its in-control state?
2) What are the expected costs per hour when the process moves out-of-control with a new mean $\mu_1 = \mu_0 + \delta\,\sigma$ with δ=1?

Multichance students can skip:

Exercise 1 point 4;

Exercise 2 point 1 and

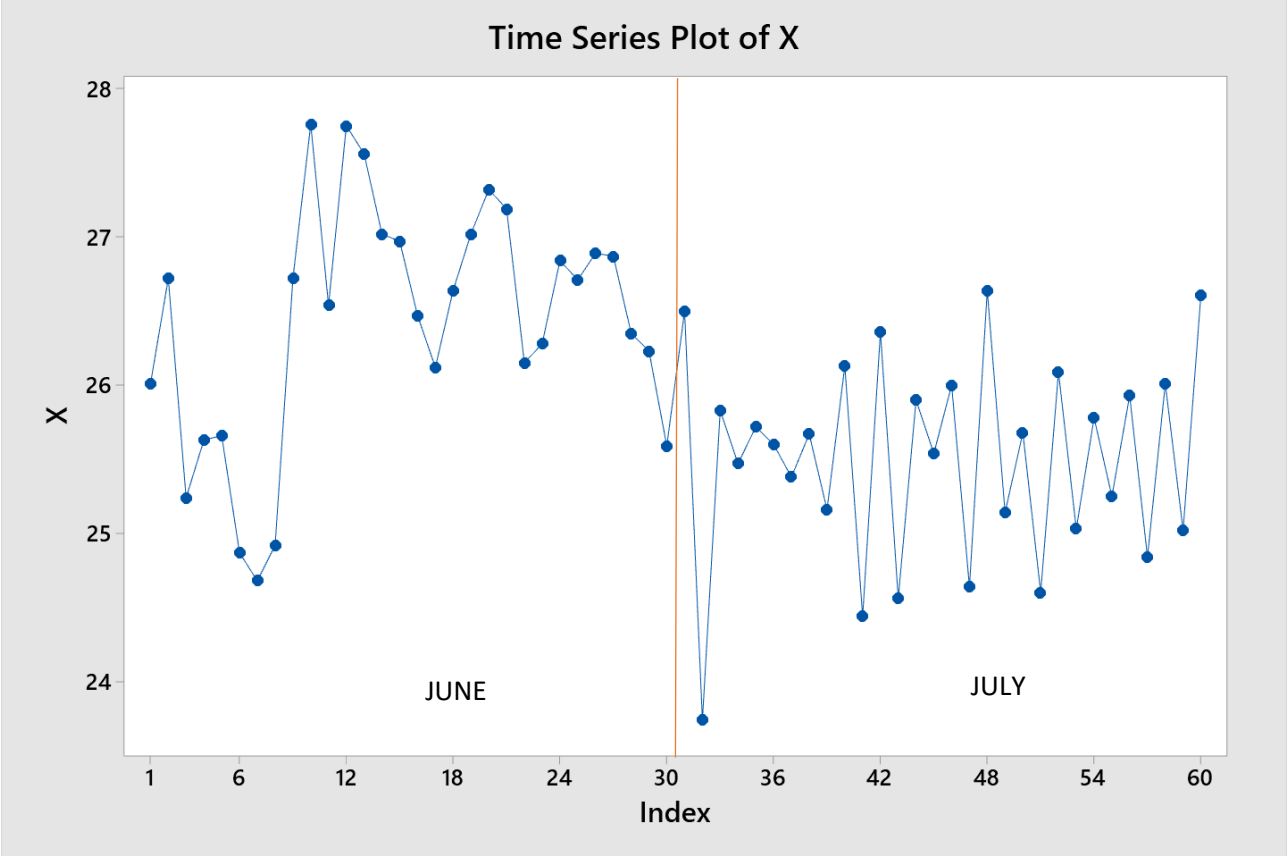
Exercice 3 point 2.

Solutions

Exercise 1

1)

Time series plot:



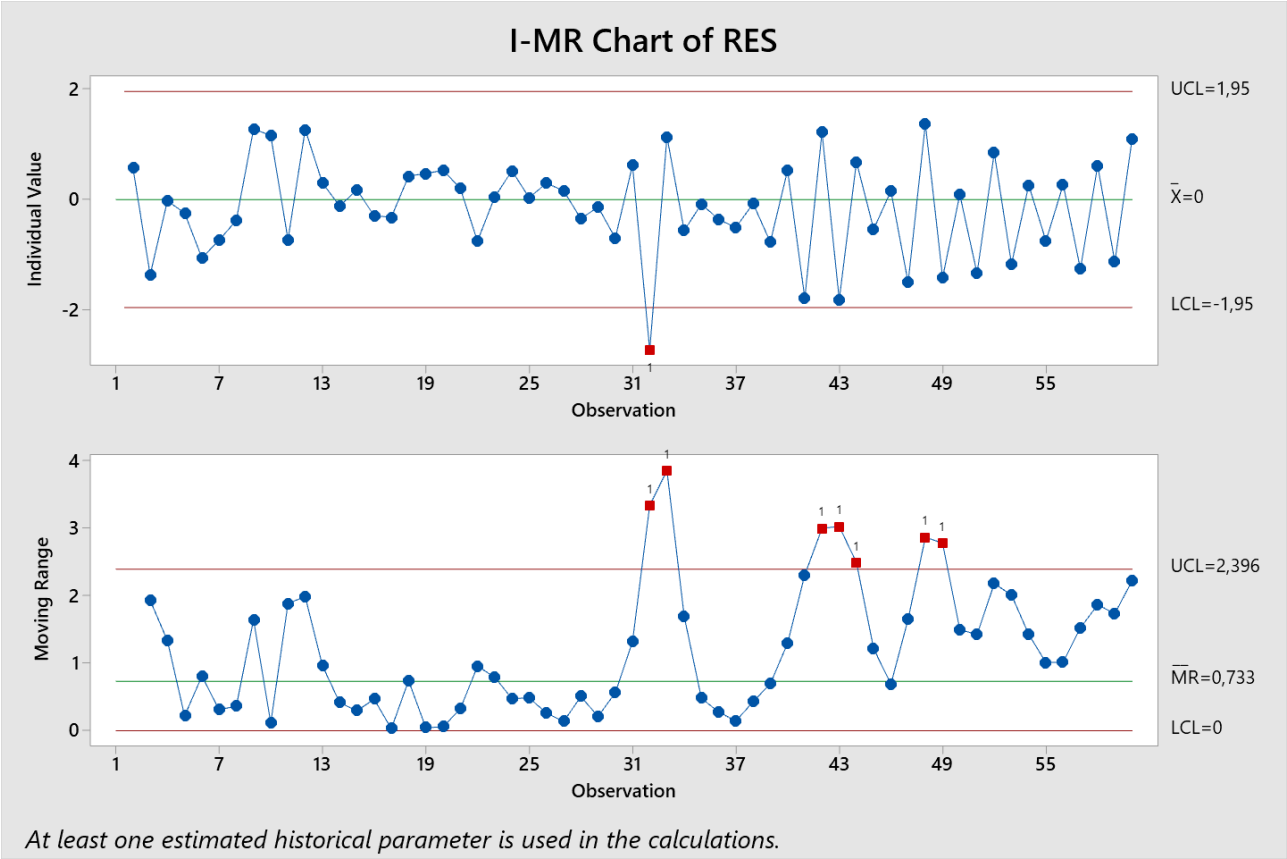
Measured data in June and July exhibit to different patterns, possibly related to a change in the autocorrelation structure.

The model in known, $x_t = 9.5 + 0.64x_{t-1} + \varepsilon$. The corresponding model residuals for June's and July's data are:

	X	AR	FITS	RES
June	26,01			
	26,72	26,01	26,1464	0,5736
	25,24	26,72	26,6008	-1,3608
	25,63	25,24	25,6536	-0,0236
	25,66	25,63	25,9032	-0,2432
	24,87	25,66	25,9224	-1,0524
	24,68	24,87	25,4168	-0,7368
	24,92	24,68	25,2952	-0,3752
	26,72	24,92	25,4488	1,2712
	27,76	26,72	26,6008	1,1592
	26,54	27,76	27,2664	-0,7264
	27,75	26,54	26,4856	1,2644
	27,56	27,75	27,26	0,3
	27,02	27,56	27,1384	-0,1184

	26,97	27,02	26,7928	0,1772
	26,47	26,97	26,7608	-0,2908
	26,12	26,47	26,4408	-0,3208
	26,64	26,12	26,2168	0,4232
	27,02	26,64	26,5496	0,4704
	27,32	27,02	26,7928	0,5272
	27,19	27,32	26,9848	0,2052
	26,15	27,19	26,9016	-0,7516
	26,28	26,15	26,236	0,044
	26,84	26,28	26,3192	0,5208
	26,71	26,84	26,6776	0,0324
	26,89	26,71	26,5944	0,2956
	26,87	26,89	26,7096	0,1604
	26,35	26,87	26,6968	-0,3468
	26,23	26,35	26,364	-0,134
	25,59	26,23	26,2872	-0,6972
July	26,5	25,59	25,8776	0,6224
	23,74	26,5	26,46	-2,72
	25,83	23,74	24,6936	1,1364
	25,47	25,83	26,0312	-0,5612
	25,72	25,47	25,8008	-0,0808
	25,6	25,72	25,9608	-0,3608
	25,38	25,6	25,884	-0,504
	25,67	25,38	25,7432	-0,0732
	25,16	25,67	25,9288	-0,7688
	26,13	25,16	25,6024	0,5276
	24,44	26,13	26,2232	-1,7832
	26,36	24,44	25,1416	1,2184
	24,56	26,36	26,3704	-1,8104
	25,9	24,56	25,2184	0,6816
	25,54	25,9	26,076	-0,536
	26	25,54	25,8456	0,1544
	24,64	26	26,14	-1,5
	26,64	24,64	25,2696	1,3704
	25,14	26,64	26,5496	-1,4096
	25,68	25,14	25,5896	0,0904
	24,6	25,68	25,9352	-1,3352
	26,09	24,6	25,244	0,846
	25,03	26,09	26,1976	-1,1676
	25,78	25,03	25,5192	0,2608
	25,25	25,78	25,9992	-0,7492
	25,93	25,25	25,66	0,27
	24,84	25,93	26,0952	-1,2552
	26,01	24,84	25,3976	0,6124
	25,02	26,01	26,1464	-1,1264
	26,61	25,02	25,5128	1,0972

The resulting special cause control chart with known $\mu_\varepsilon = 0$ and known $\sigma_\varepsilon = 0.65$ is the following:



The process results to be in-control in June (although some hugging may be present in the second part of the month), whereas it results out-of-control in July (several violations of the limits in the MR control chart; only one violation on the first measurement in July in the I control chart, but a systematic pattern is present, highlighting a possible negative autocorrelation of residuals).

2)

The two control charts on the mean of model coefficients can be designed as follow:

Control chart for the constant term:

$$UCL = E(b_0) + K\sqrt{V(b_0)} = 17.67$$

$$CL = E(b_0) = 9.5$$

$$UCL = E(b_0) + K\sqrt{V(b_0)} = 1.33$$

Control chart for the autoregressive term:

$$UCL = E(b_1) + K\sqrt{V(b_1)} = 1{,}09$$

$$CL = E(b_1) = 0.64$$

$$UCL = E(b_1) + K\sqrt{V(b_1)} = 0.19$$

where $K = z_{\alpha/2}$, being $\alpha' = 1/ARL_0$ and $\alpha = \alpha'/2$ according to the Bonferroni's correction, as two control charts are used in parallel for the two model coefficients. Therefore: $K = 3.205$.

3)

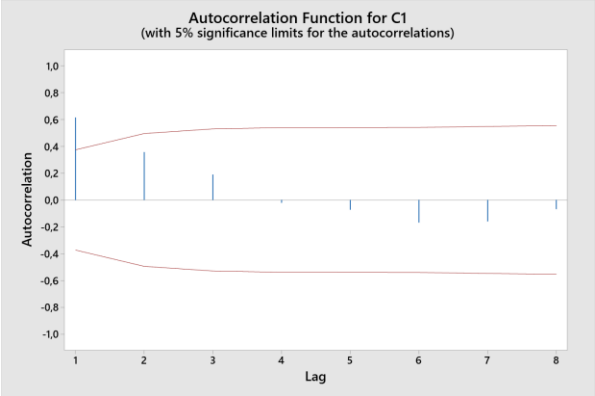
Model fitting for June's data:

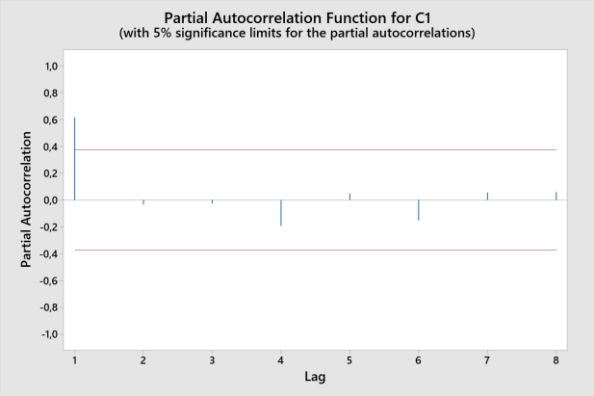
Test

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
9	15,73	0,011





Regression Analysis: AR1 versus C2

Method

Rows 1
unused

Regression Equation

X = 9,58 + 0,637 AR1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	9,58	4,02	2,38	0,024	
C2	0,637	0,152	4,20	0,000	1,00

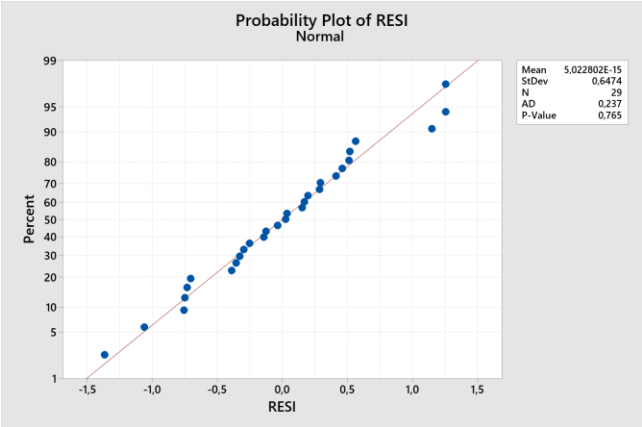
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,659275	39,49%	37,25%	29,07%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	7,660	7,6602	17,62	0,000
C2	1	7,660	7,6602	17,62	0,000
Error	27	11,735	0,4346		
Lack-of-Fit	25	8,499	0,3400	0,21	0,982
Pure Error	2	3,236	1,6182		
Total	28	19,396			

Check of assumptions on model residuals:

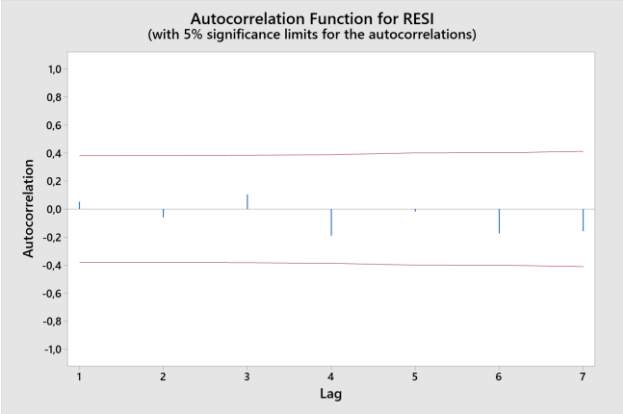


Test

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
12	15,48	0,187



The model is appropriate.

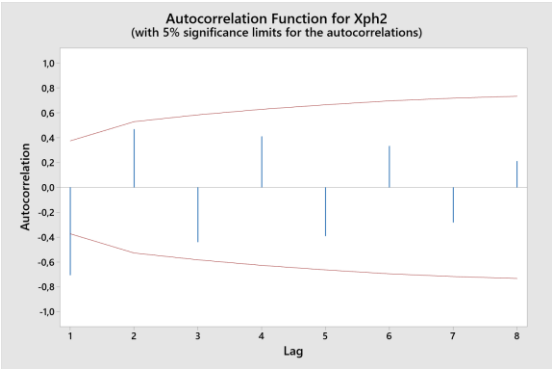
Model fitting for July's data:

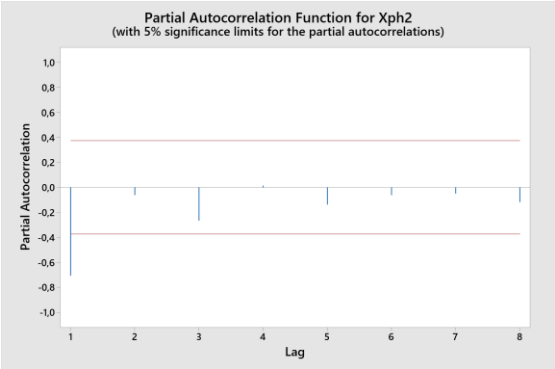
Test

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
27	15,73	0,000





WORKSHEET 1

Regression Analysis: Xph2 versus ARph2

Regression Equation

Xph2 = 45,08 - 0,768 ARph2

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	45,08	3,43	13,15	0,000	
ARph2	-0,768	0,135	-5,71	0,000	1,00

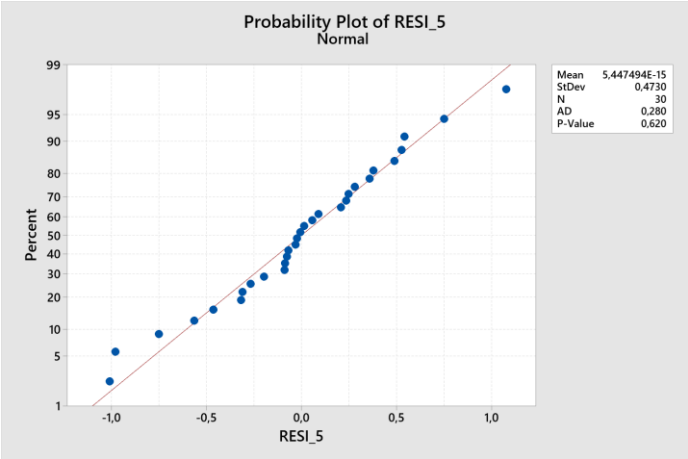
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,481343	53,80%	52,15%	41,79%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	7,555	7,5546	32,61	0,000
ARph2	1	7,555	7,5546	32,61	0,000
Error	28	6,487	0,2317		
Total	29	14,042			

Check of assumptions on model residuals:



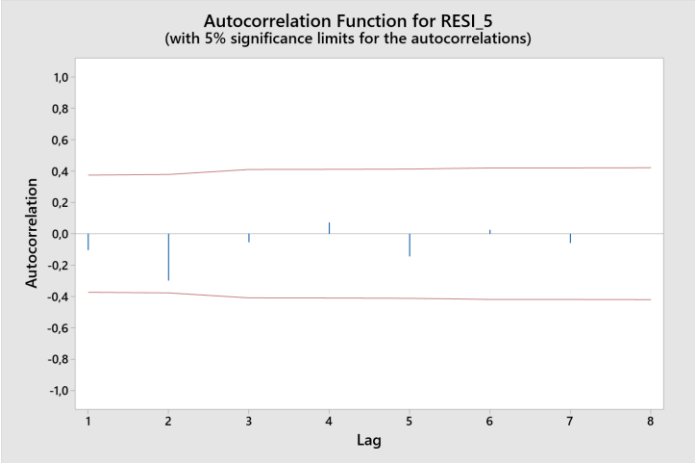
Test

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
19	15,93	0,252



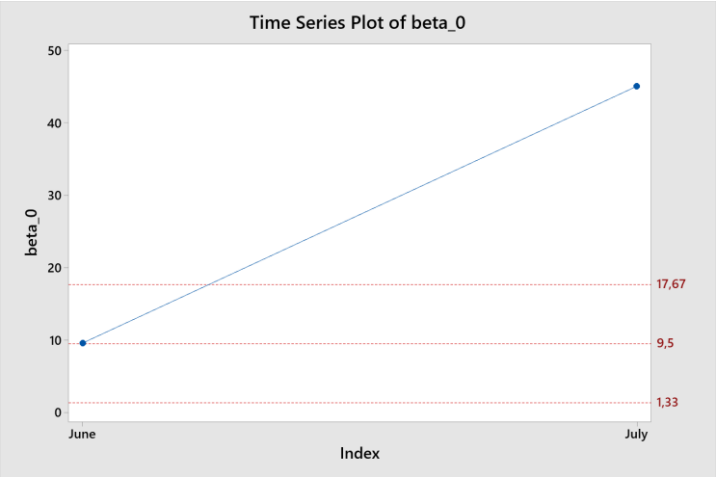
The model is appropriate.

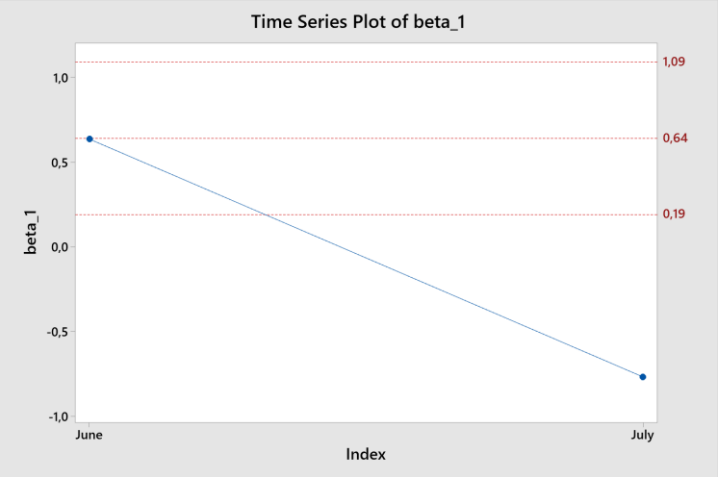
4)

Estimated model coefficients are:

	b_0	b_1
June	9.58	0.637
July	45.08	-0.768

By using the control chart designed in point 2, the test phase in June and July is the following:





The process in July results to be out-of-control both in terms of b_0 and b_1 coefficients. Indeed, the process in July exhibits a negative autocorrelation pattern, differently from the in-control process model.

The special cause control chart has a higher reactivity than the control chart on estimated model parameters, since it allows signaling an alarm at every measured data point, but it is also prone to false alarms related to random fluctuations within the month. The control chart on estimated model parameters is less reactive, as it allows signalling an alarm just at the end of the month, but it can be more effective if the aim is to detect a change in the underlying model despite the slowest reaction.

In real applications, when control charts on estimated model parameters are used, it is a good practice to include also a control chart on the residuals of the model to avoid information losses and enhance the detection capability of the monitoring tool.

Exercise 2

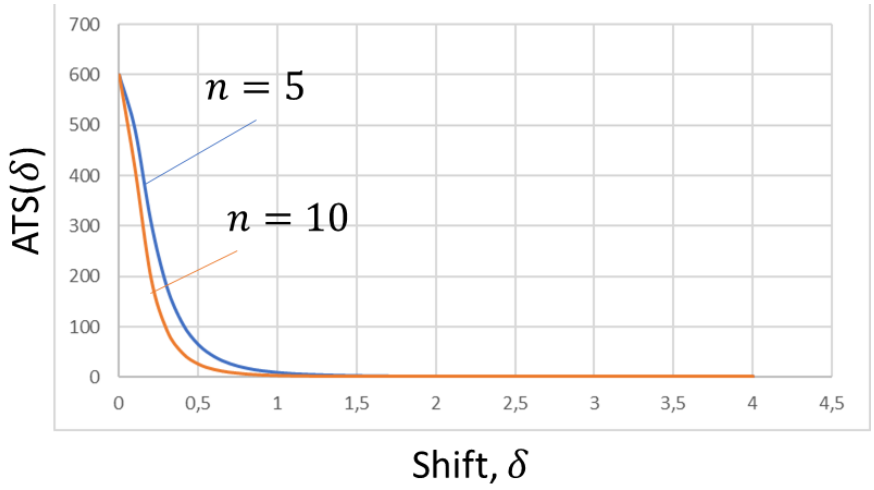
The value of $K = z_{\alpha/2}$ with $\alpha = \frac{1}{200} = 0.005$ is: $K = 2{,}807$.

The Type II error as a function of the mean shift in standard deviation units is given by:

$\beta = \Pr\left(Z \leq K - \delta\sqrt{n}\right) - \Pr\left(Z \leq -K - \delta\sqrt{n}\right)$, where $\delta = \frac{\mu_1 - \mu_0}{\sigma}$

Then, $ATS(\delta) = h \cdot \frac{1}{1-\beta}$ where $h = 3$ hours.

The $ATS(\delta)$ curves for n=5 and n=10 are the following:



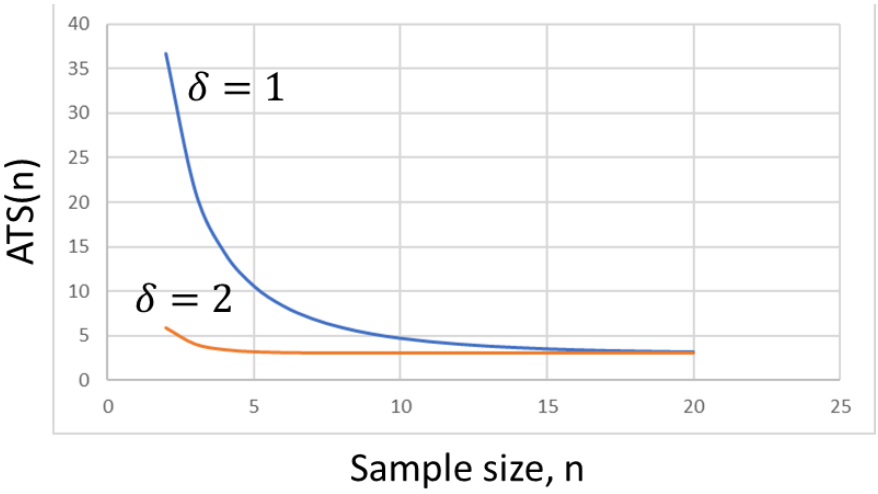
	$\delta = 1$	$\delta = 2$
ATS with n=5	10.56 h	3.15 h
ATS with n=10	4.70 h	3.00 h

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 $ATS(n)$ curves for the two given mean shifts are the following:



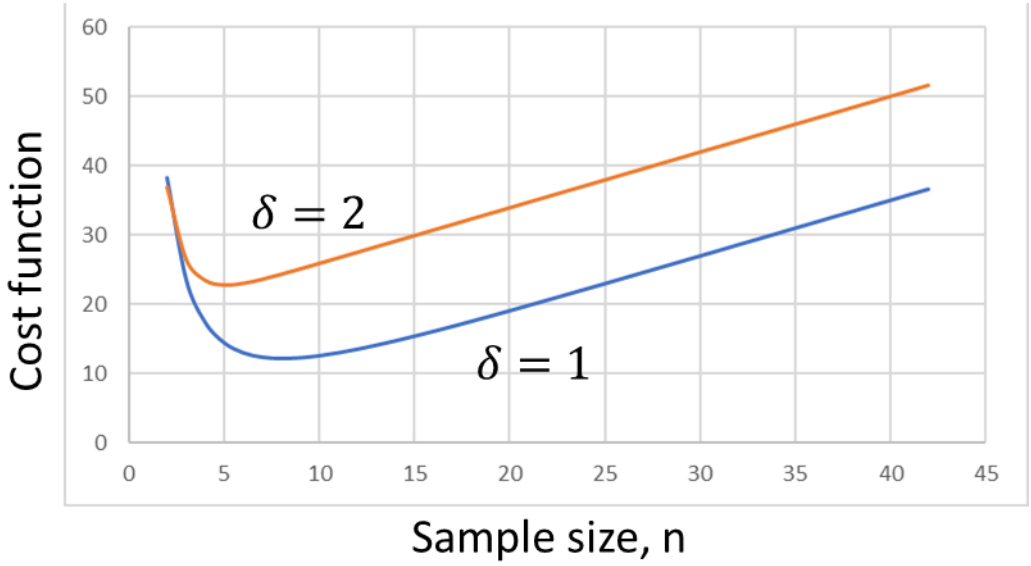
c)

The function to be minimized is the following:

$$C(n) = C1 * ATS(n) + C2 * n$$

For $\delta = 1$, $C(n) = 1ATS(n) + 0.8n$, whereas for $\delta = 2$, $C(n) = 6ATS(n) + 0.8n$.

The cost functions for the two shifts are shown below:



The optimal values of the sample size are:

- n=8 for $\delta = 1$
- n=5 for $\delta = 2$

The late detection cost predominates at smaller values of n, whereas the inspection cost predominates at larger values of n. For a smaller shift, the best compromise is obtained with a larger sample size, due to the higher relative effect of a late detection cost (with higher Type II error), whereas for a larger shift the best compromise is obtained with a smaller sample size.

Exercise 3

a)

When the process is in-control, the probability of non-conforming items is $\gamma = P(D < LSL) + P(D > USL) = 0.0455$, whereas the probability of producing conforming items is $1 - \gamma$.

The cost per each conforming items is therefore $C_c = 0.1(1 - \gamma) = 0.095€$, whereas the cost per each non-conforming item is $C_{nc} = 2(\gamma) = 0.091€$.

Being 60 the number of parts produced per hour, when the process is in-control ($\mu_0 = 20$ mm) the cost per hour is $C_{IC} = 60(C_c + C_{nc}) = 11.19€/h$.

b)

When the process is out-of-control, the probability of non-conforming items is $\gamma = P(D < LSL) + P(D > USL) = 0.16$, whereas the probability of producing conforming items is $1 - \gamma$.

The cost per each conforming items is therefore $C_c = 0.1(1 - \gamma) = 0.084€$, whereas the cost per each non-conforming item is $C_{nc} = 2(\gamma) = 0.32€$.

Being 60 the number of parts produced per hour, when the process is out-of-control the cost per hour is $C_{OOC} = 60(C_c + C_{nc}) = 24.24€/h$.