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
- <u>**Multichance students should skip: point b) in Exercise 1 and point c) in Exercise 2**</u>

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

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)

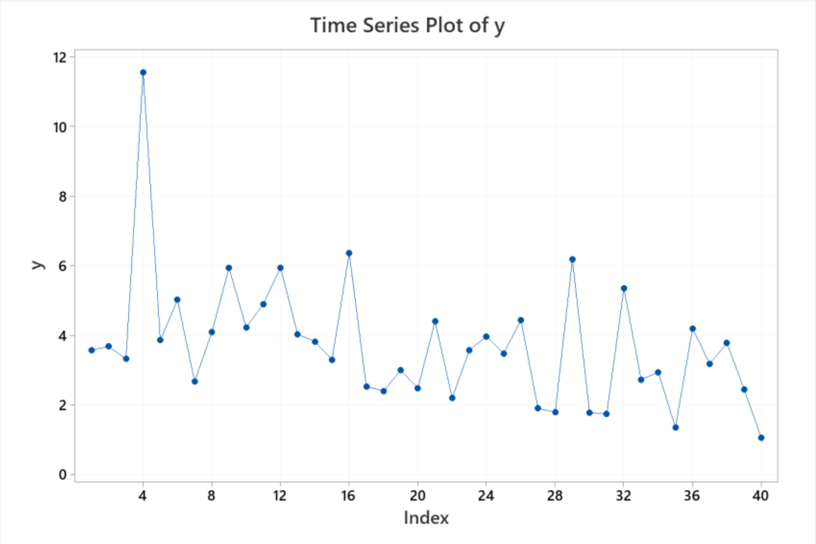
A chemical process for the production of jelly is monitored by means of a $\bar{X} - S$ control chart with known parameters. Based on historical evidence, it is known that an out-of-control increase of the mean always occurs with a simultaneous increase of the standard deviation of the process.

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

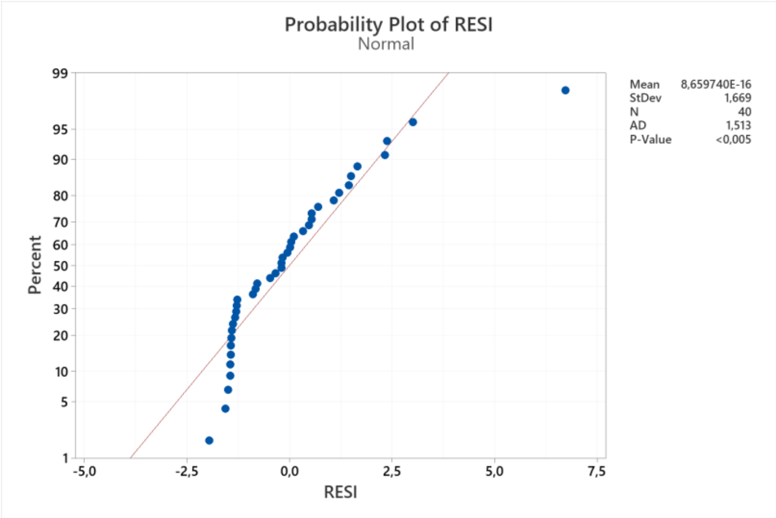
Model Summary

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

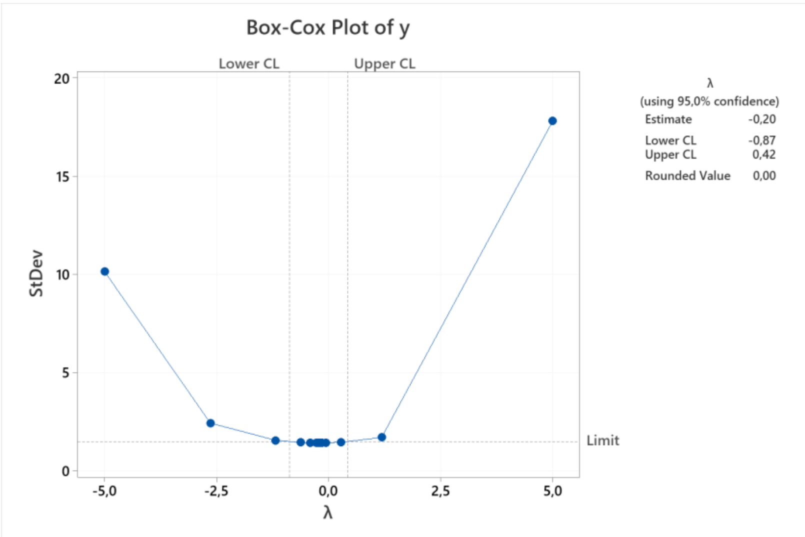
Analysis of Variance

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			

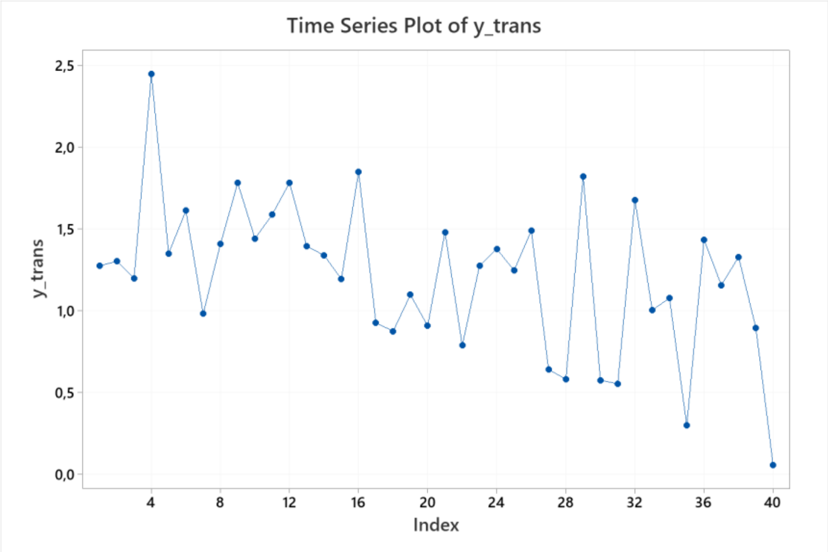
But there is a violation of the normality assumption affects the model residuals:



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

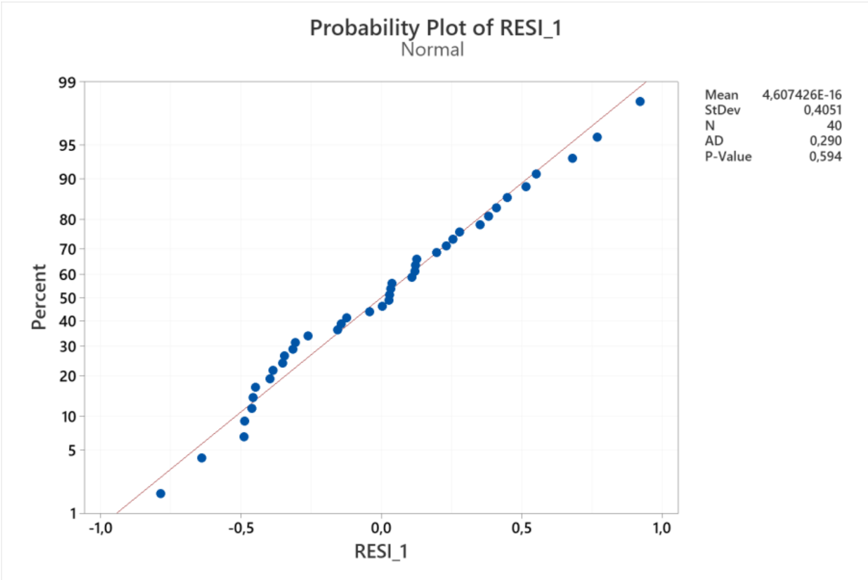
Model Summary

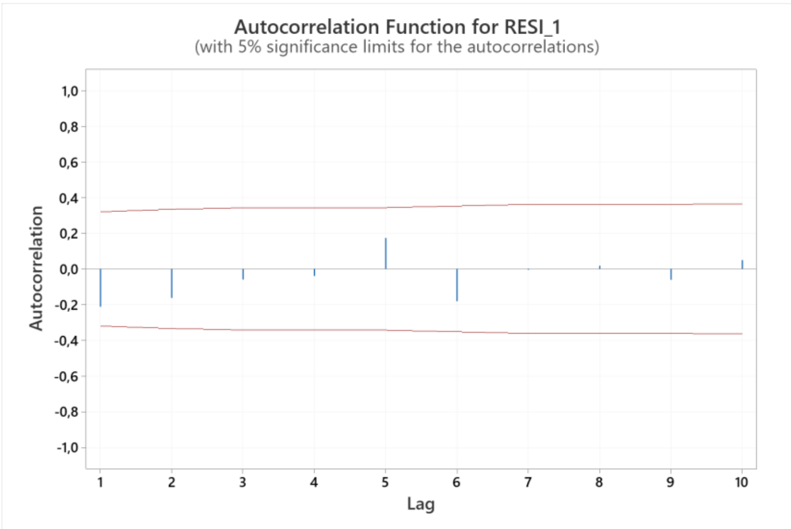
S	R-sq	R-sq(adj)	R-sq(pred)
0,410438	22,98%	20,96%	13,40%

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	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	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:

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

$$UCL = b_0 + b_1 t$$

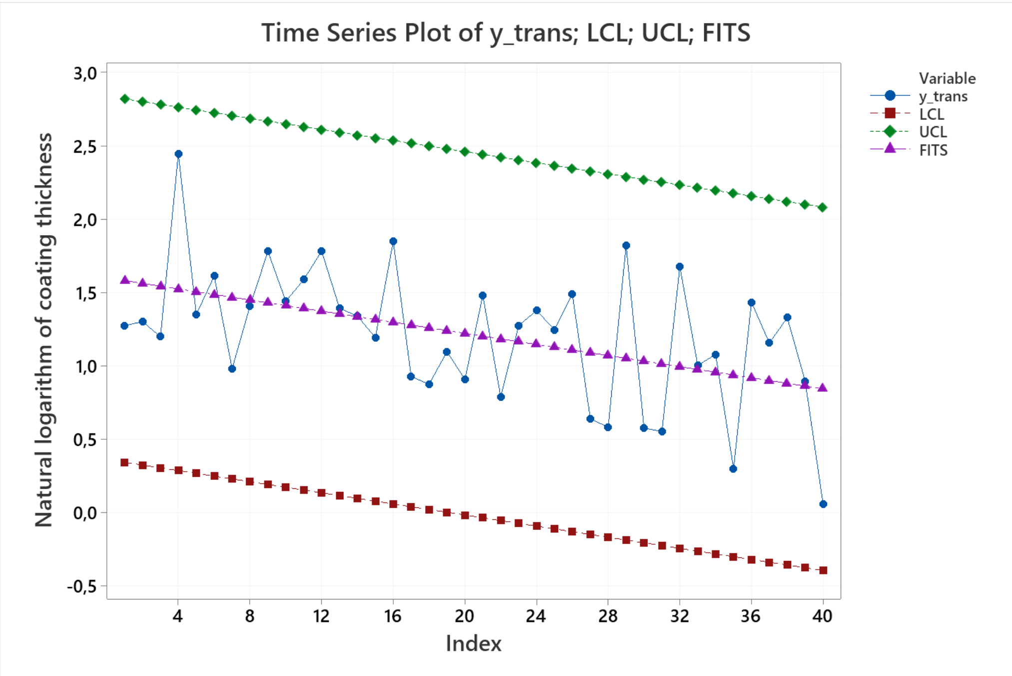
$$LCL = b_0 + b_1 t - z_{\alpha/2} \frac{\overline{MR}}{d_2(2)}$$

Where:

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

$d_2(2)$ =1,128

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

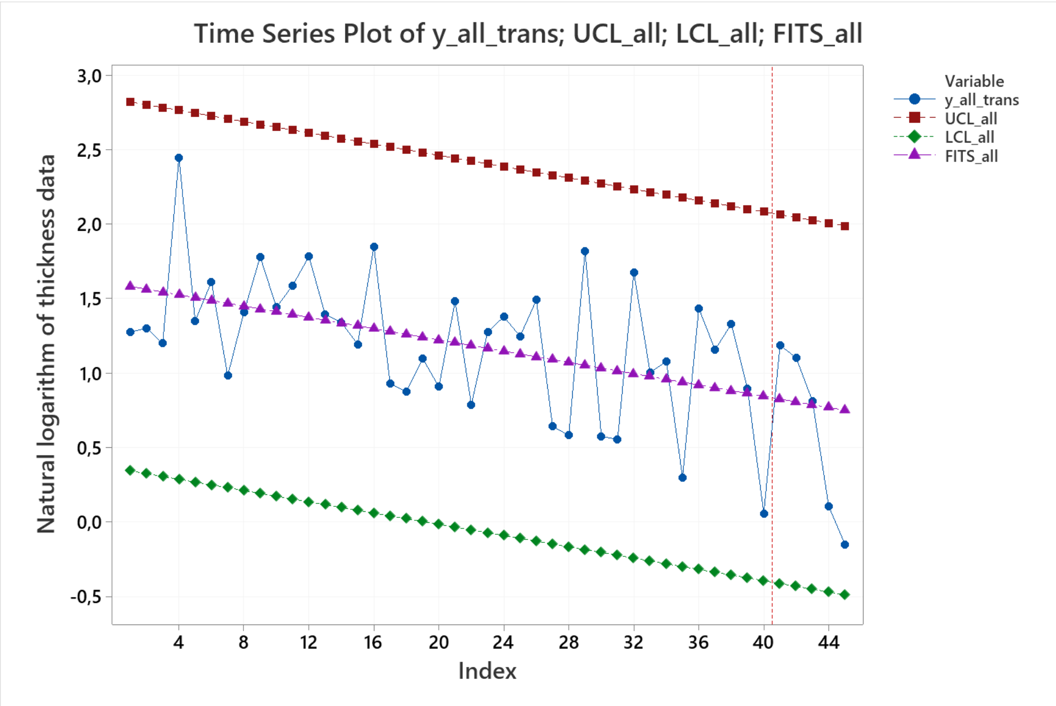


b)

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:

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

$$\sigma_{PC1}^2 = \lambda_1 = 8.42364,$$

$$\sigma_{PC2}^2 = \lambda_2 = 1.26052$$

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 $\alpha^* = 1-(1-\alpha)^{1/2} = 0{,}005013$.

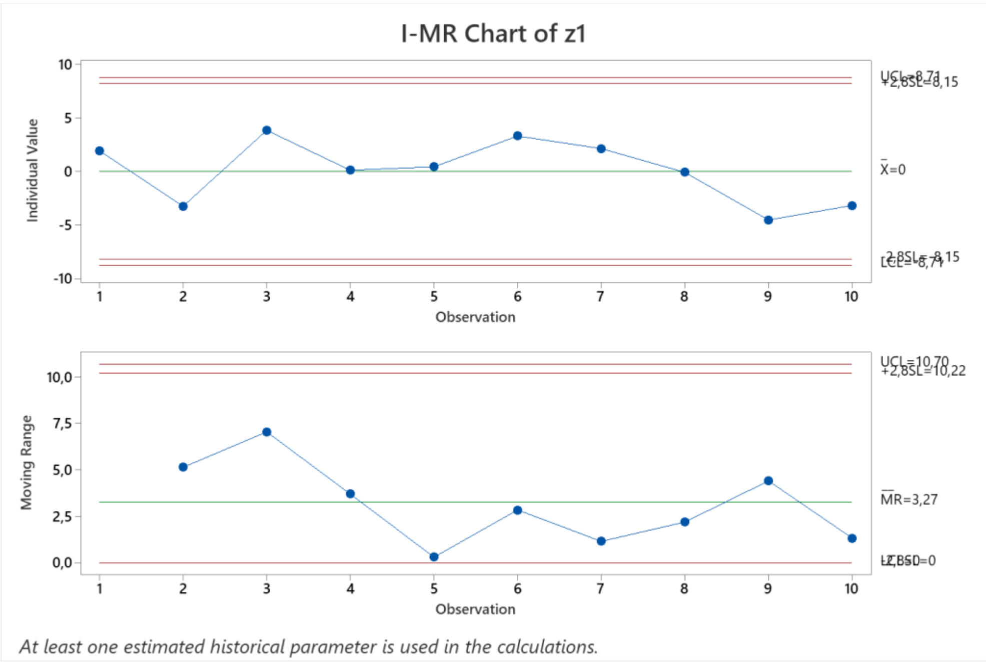
The control charts with $K = z_{\alpha^*/2} = 2.807$ have the following control limits:

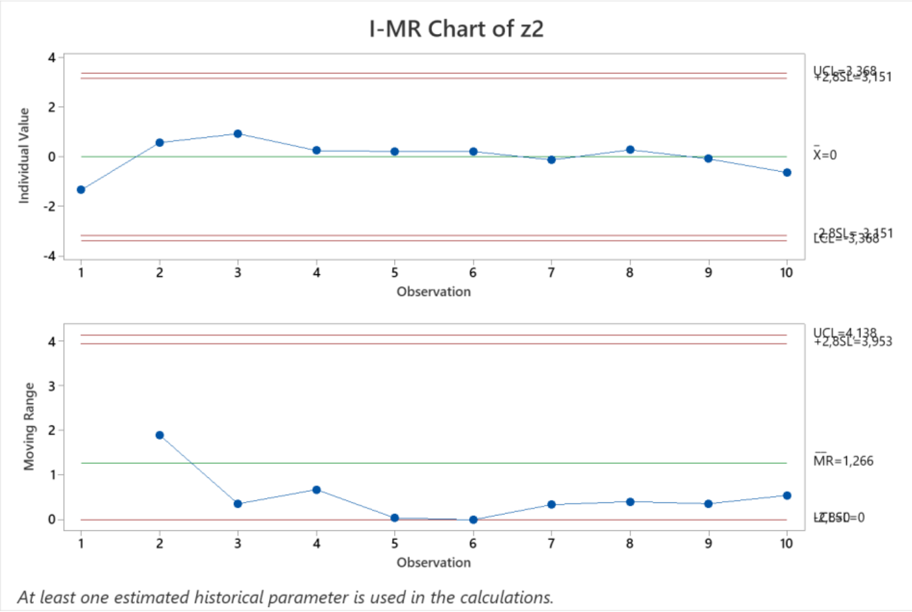
PC1		PC2	
I	MR		
LCL = -8.15, UCL = 8.15	LCL = 0, UCL = 10.22	LCL = -3.151, UCL = 3.151	LCL = 0, UCL = 3.953

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).

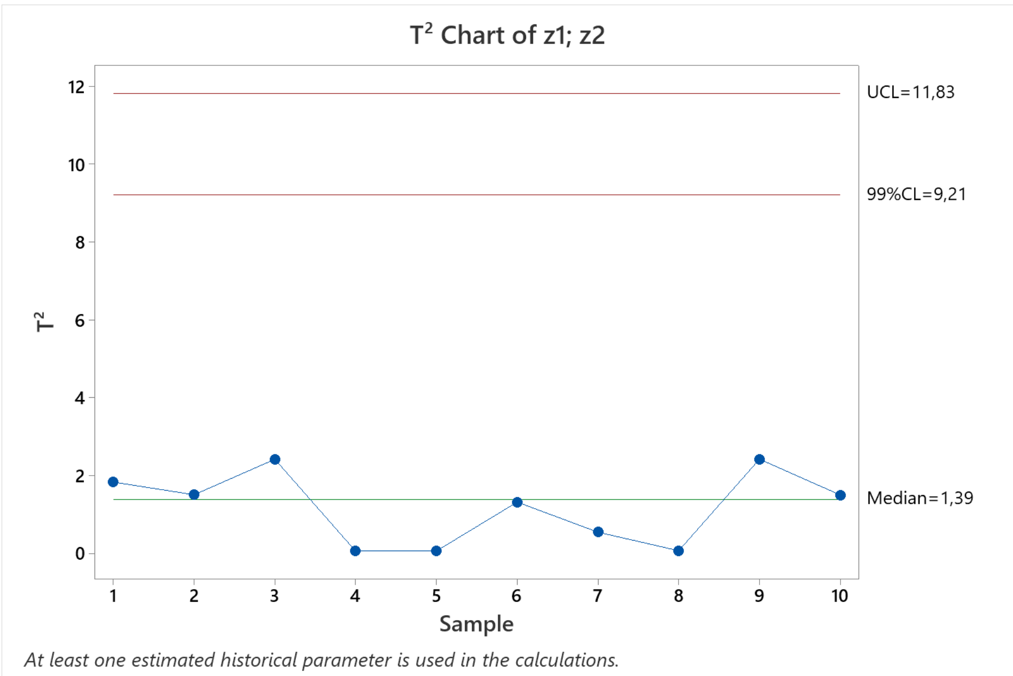




There is no violation of the control limits, although hugging is present along the second PC (which can be a symptom of a change in the process).

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 $\mu_{PC1} = 0, \sigma_{PC1}^2 = \lambda_1 = 8.42364$, 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:

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

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

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

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