QUALITY DATA ANALYSIS

24/06/2020

General recommendations:

- write the solutions in a 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)

A company installed a new health monitoring system on one machine tools running in a harsh environment. The system allows the production engineers to estimate the degradation index of a linear axis, computed during a check-up test done every day. The values of the degradation index collected during the last 60 days are reported in the table below. Every 20 days a recalibration operation of the axis is performed.

Table 1:

day	data	day	data	day	data
1	-0,293	21	5,976	41	1,68
2	-3,087	22	3,713	42	1,773
3	3,771	23	2,27	43	-0,078
4	3,659	24	-2,929	44	-0,217
5	-2,975	25	4,808	45	0,99
6	4,472	26	-0,719	46	-0,089
7	1,438	27	-2,835	47	-2,812
8	0,909	28	3,315	48	-1,815
9	0,58	29	3,322	49	3,373
10	2,788	30	5,515	50	0,391
11	8,613	31	5,181	51	5,079
12	-1,204	32	1,331	52	7,335
13	0,494	33	8,704	53	2,914
14	7,868	34	4,728	54	8,963
15	6,749	35	4,566	55	6,659
16	11,318	36	6,062	56	8,076
17	15,835	37	11,311	57	3,325
18	13,46	38	8,773	58	21,511
19	11,52	39	14,78	59	14,334
20	22,137	40	16,044	60	17,419

a) Fit a suitable model of the degradation path;
b) Design a control chart to determine if any anomaly different from the natural degradation path occurred during the monitored period (design the control chart such that the average time before a false alarm is 240 days);

c) How does the degradation model change if one single degradation curve is fitted for each period following a recalibration step? Discuss the differences between this modelling approach and an approach where a single model is used for all the degradation path.

Exercise 2 (15 points)

In a plant for the production of tomato juice, three quality characteristics are monitored through individual measurements: pH, protein content (g/100g) and vitamin C content (mg/100g). It is known that, under in-control process conditions, the three quality characteristics follow a multinormal distribution with:

$$\boldsymbol{\mu} = [5.05\ 9.69\ 172.98]'$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.02 & 0.006 & 0.016 \\ 0.006 & 0.09 & 0.032 \\ 0.016 & 0.032 & 0.38 \end{bmatrix}$$

a) Estimate the Principal Components (PCs) that explain at least 70% of the overall variability by using both the variance-covariance matrix and the correlation matrix and discuss which is the most appropriate approach; show the corresponding loadings and variances for both the solutions.
b) Design univariate control charts for the mean of the PCs retained by using the most appropriate approach in point a) such that the familywise Type I error is 0.01.
c) Table 2 shows new measurements of the three quality characteristics acquired in two consecutive weeks. By using the control charts designed in point b), verify if the process is in-control and discuss the results.

Table 2:

Week	Day	Measurements		
		pH	Protein (g/100g)	Vitamin (mg/100g)
Week 1	Monday	5,16	10,07	173,18
	Tuesday	5,35	9,91	173,48
	Wednesday	4,52	9,56	172,73
	Thursday	4,66	9,23	172,01
	Friday	5,37	9,66	173
Week 2	Monday	4,75	10,11	172,76
	Tuesday	4,59	10,18	172,73
	Wednesday	4,63	10,24	174,39
	Thursday	4,54	10,57	175,93
	Friday	4,7	10,25	172,07

Exercise 3 (3 points)

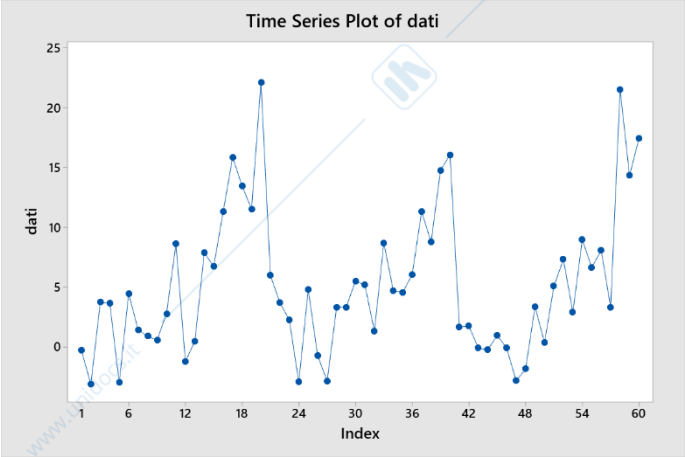
A company is interested in observing the stability of the linear drift of a tool. To this aim, for each tool, a linear model is fitted $y = \beta_0 + \beta_1 t + \varepsilon_t$ starting from the same number n of data observed in each drift curve.

Then, for each curve the slope $\hat{\beta}_1 = b_1$ is estimated using ordinary least squares. Assuming $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, and assuming known β_0 and β_1, describe the expression of the control limits for a control chart for monitoring the curve slope of each tool.

Exercise 1 Solution

a)

Data snooping: time series plot



There is a non linear degradation path. The effect of the recalibration operation is evident at time 21 and 41. Data are clearly nonrandom, as confirmed by the runs test and the ACF and PACF plots

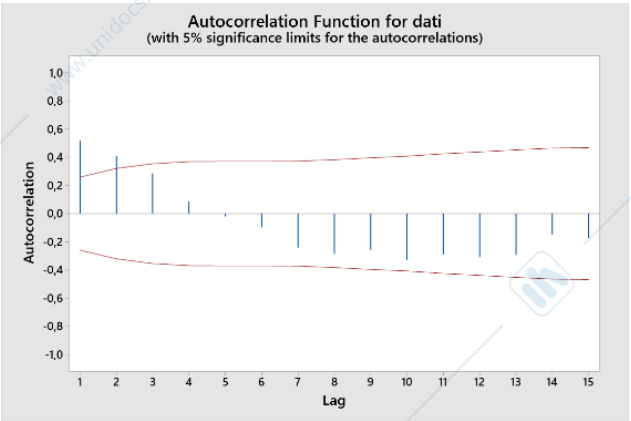
Test

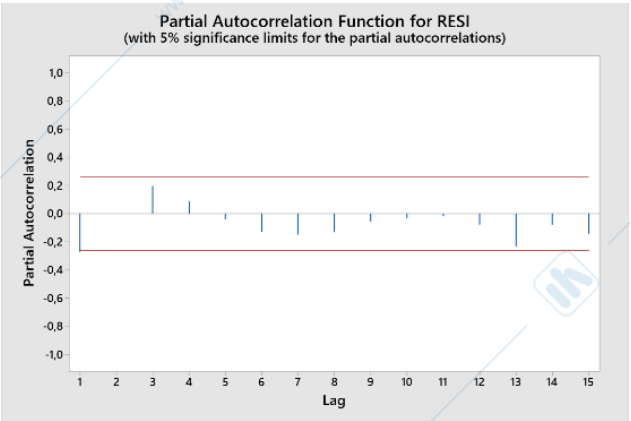
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs

Observed	Expected	P-Value
16	29,80	0,000

ACF and PACF:





In order to fit a proper model, different solutions are possible.

One possible solution consists of defining a time variable that accounts for the degradation path before the recalibration occurs, i.e. counting the “age” (i.e, ranging from 1 to 20 days before the first recalibration and then starting again from 1 to 20 days after each recalibration). Such regressor is called “minitrend”. Additional regressors, i.e., “minitrend^2” and “minitrend^3”, can be also

considered to account for quadratic and curvilinear paths. A dummy variable that is equal to 1 in days where a calibration was performed and 0 for all other days can be included as well.

By performing a step-wise regression, the following result is obtained:

ES1_B

Regression Analysis: dati versus dummy; minitrend; minitrend^2; minitrend^3

Stepwise Selection of Terms

α to enter = 0,15; α to remove = 0,15

Regression Equation

dati = 0,575
+ 0,002088 minitrend^3

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0,575	0,548	1,05	0,299	
minitrend^3	0,002088	0,000167	12,53	0,000	1,00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3,15235	73,02%	72,55%	70,96%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1559,9	1559,91	156,97	0,000
minitrend^3	1	1559,9	1559,91	156,97	0,000
Error	58	576,4	9,94		
Lack-of-Fit	18	113,1	6,28	0,54	0,918
Pure Error	40	463,3	11,58		
Total	59	2136,3			

Fits and Diagnostics for Unusual Observations

Obs	dati	Fit	Resid	Std Resid	
20	22,137	17,282	4,855	1,63	X
40	16,044	17,282	-1,238	-0,42	X
57	3,325	10,835	-7,510	-2,43	R
58	21,511	12,754	8,757	2,86	R
60	17,419	17,282	0,137	0,05	X

R Large residual
X Unusual X

The constant term has to be eliminated, as it is not significant. The model could be possibly defined in order to respect the hierarchical principle (i.e., by including the linear and quadratic trend in addition to the cubic term). In the following, the non-hierarchical model is considered, assuming the coefficients of the linear and quadratic terms are equal to zero.
After removing the non significant terms we have:

ES1_A

Regression Analysis: dati versus minitrend; minitrend^2; minitrend^3

Stepwise Selection of Terms

α to enter = 0,15; α to remove = 0,15

Regression Equation

dati = 0,002206 minitrend^3

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
minitrend^3	0,002206	0,000124	17,81	0,000	1,00

Model Summary

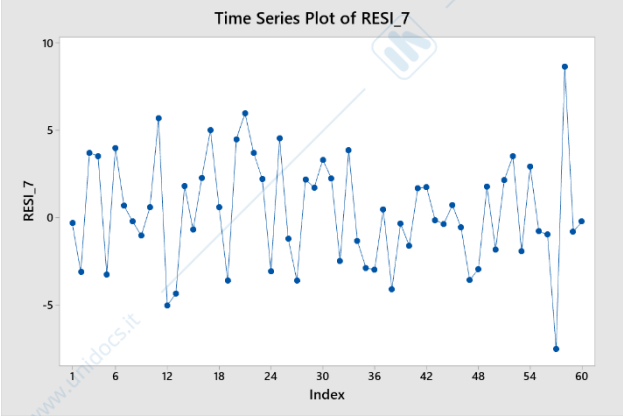
S	R-sq	R-sq(adj)	R-sq(pred)
3,15497	84,32%	84,06%	83,60%

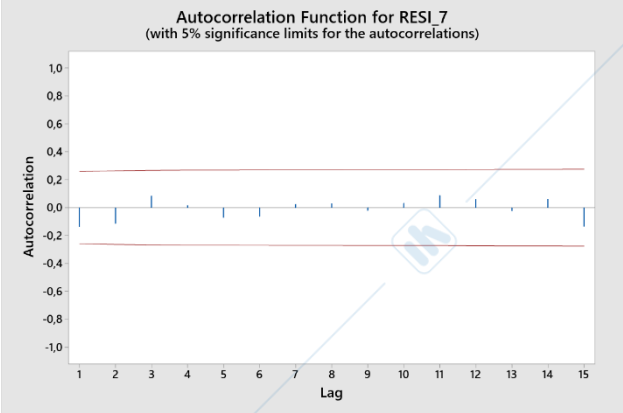
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3158,7	3158,73	317,34	0,000
minitrend^3	1	3158,7	3158,73	317,34	0,000
Error	59	587,3	9,95		

Lack-of-Fit	19	124,0	6,53	0,56	0,910
Pure Error	40	463,3	11,58		
Total	60	3746,0			

In this case the final residuals satisfy the assumptions:



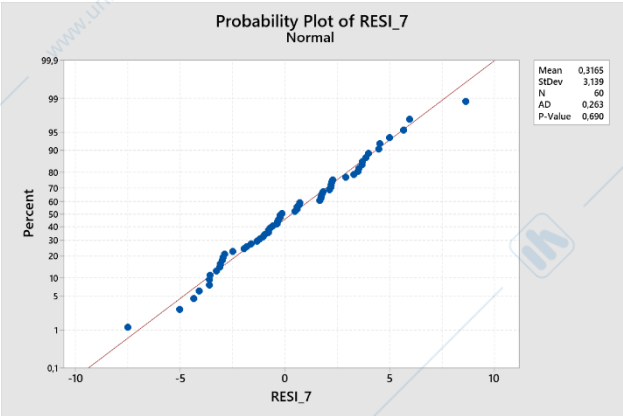


Test

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
33	30,97	0,596



b)

With reference to the fitted model we can design a control chart as follows.

The average time to signal is $ATS = \Delta T \cdot ARL(H_0) = 240$ days. Since the time interval between successive measurements is 1 day, we have:

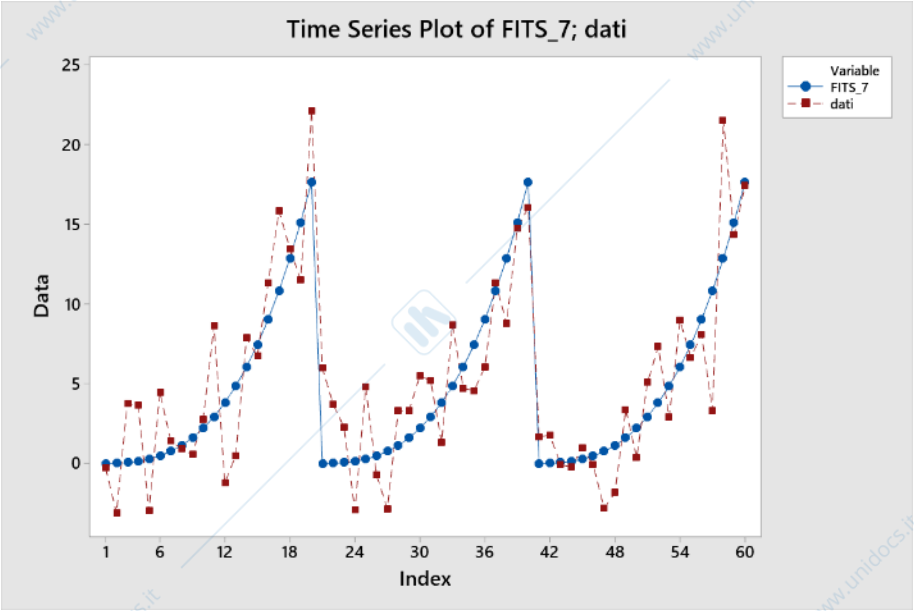
$$ARL(H_0) = \frac{ATS}{\Delta T} = \frac{240}{1} = 240$$

Thus:

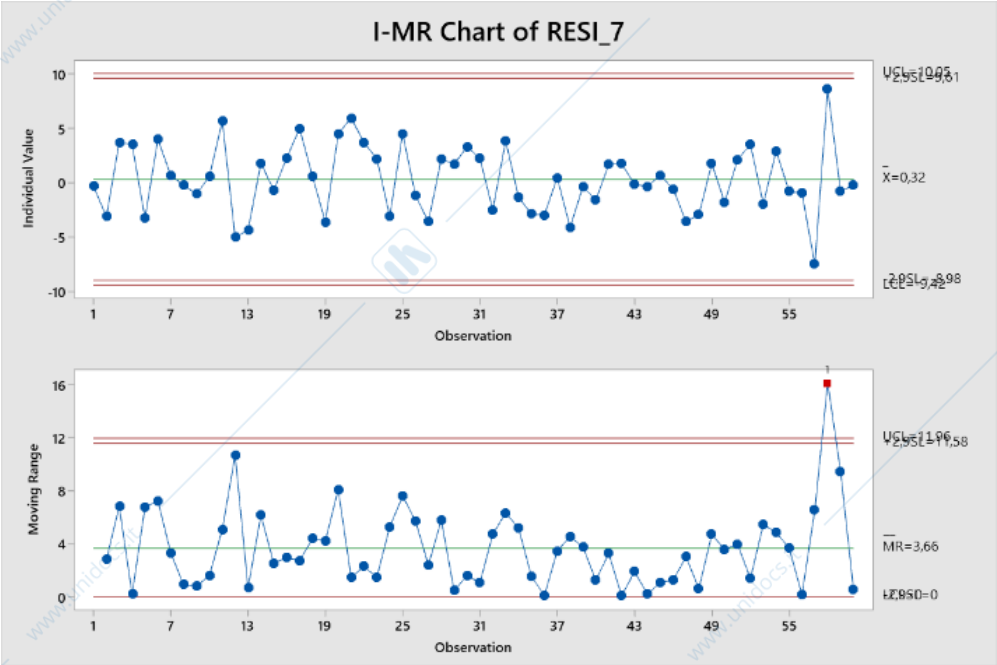
$$\alpha = \frac{1}{ARL(H_0)} = \frac{1}{240} = 0{,}0042$$

$$K = z_{\alpha/2} = 2{,}863$$

The Fitted value chart is the following:



The special cause control chart is the following:



Since a violation of the limits is present only in the MR chart (day 58), it is possible to further evaluate if it is caused by a violation of the assumptions associated to the MR control chart itself.

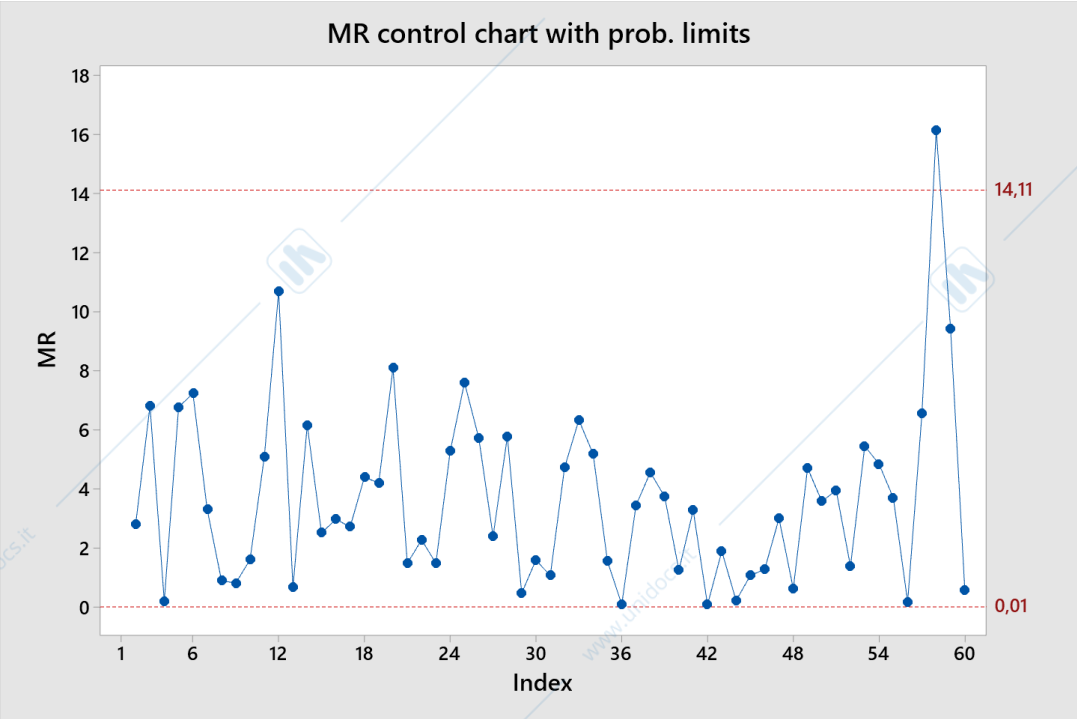
An MR control chart with probabilistic limits can be designed referring to the true half-normal distribution as follows:

$$\text{UCL} = D_{1-\alpha/2}\frac{\overline{MR}}{d_2} \qquad \text{LCL} = D_{\alpha/2}\frac{\overline{MR}}{d_2}$$

Where:

$$D_{1-\alpha/2} = \sqrt{2}z_{\alpha/4} \qquad D_{\alpha/2} = \sqrt{2}z_{1/2-\alpha/4}$$

The resulting control chart for MR is the following:



The violation of the limits on day 58 is still present. Assuming no assignable cause for the control chart is found out, the control chart design is over.

c)

By fitting a single degradation model (point a), the following results were obtained.

Stepwise Selection of Terms

α to enter = 0,15; α to remove = 0,15

Regression Equation

dati = 0,002206 minitrend^3

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
minitrend^3	0,002206	0,000124	17,81	0,000	1,00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3,15497	84,32%	84,06%	83,60%

By fitting three separate degradation models, one for each degradation path, the following results can be achieved:

<table>
<tr><th></th><th>Model</th><th>Normality</th><th>LBQ (lag=5)</th><th>R2adj</th></tr>
<tr><td>degradation curve 1</td><td>Coefficients<table><tr><th>Term</th><th>Coef</th><th>SE Coef</th><th>T-Value</th><th>P-Value</th><th>VIF</th></tr><tr><td>minitrend^3_1</td><td>0,002396</td><td>0,000221</td><td>10,86</td><td>0,000</td><td>1,00</td></tr></table></td><td>p-val: 0.548</td><td>p-val: 0.3151</td><td>85,39%</td></tr>
<tr><td>degradation curve 2</td><td>Coefficients</td><td>p-val: 0.345</td><td>p-val: 0.945</td><td>82,32%</td></tr>
</table>

	Term / **Coef** / **SE Coef** / **T-Value** / **P-Value** / **VIF** minitrend^3_1 / 0,001973 / 0,000203 / 9,70 / 0,000 / 1,00			
degradation curve 3	**Coefficients** **Term** / **Coef** / **SE Coef** / **T-Value** / **P-Value** / **VIF** minitrend^3_1 / 0,002248 / 0,000219 / 10,27 / 0,000 / 1,00	p-val: 0.207	p-val: 0.3471	83,93%

In this case, all the coefficients are estimated with less degrees of freedom leading to a higher uncertainty in the estimate of the model coefficient. The values of coefficient for the cubic term oscillate around the value that was estimated with the single model fitted in point a), which is a better estimate as all the data were used in the same model, leading to a standard deviation of the estimated coefficient that is lower than the standard deviations of estimated coefficients in separate models. The benefit of having a better estimate is especially clear in terms of residual noise, as the degrees of freedoms to estimate the error term reduces if three models are separately fitted.

Exercise 2 Solution

a)

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

$\lambda_1 = 0.384242$

$\lambda_2 = 0.086775$

$\lambda_3 = 0.018983$

The first PC explains 78.4% of the overall variability. Its loadings are the following:

Matrix EIGVET

0,045414
0,108920
0,993013

The first PC gives a large weight to the third variable (vitamin content), which has a much larger variance than other two variables.

Correlation matrix:

$$\mathbf{P} = \begin{bmatrix} 1 & 0.1414 & 0.1835 \\ 0.1414 & 1 & 0.1730 \\ 0.1835 & 0.1730 & 1 \end{bmatrix}$$

PCA on correlation matrix - eigenvalues (i.e., the variances of the PCs) are the following:

$\lambda_1 = 1.33261$

$\lambda_2 = 0.85923$

$\lambda_3 = 0.80816$

The first two PCs explain 73.1% of the overall variability. Their loadings are the following:

Matrix EIGVETR

-0,570097	0,647144
-0,556718	-0,757361
-0,604197	0,087226

The first PC is an average of all the three quality characteristics; the second PC is a contrast between the first and the second quality characteristics (pH and protein content).

Since the three quality characteristics have different scales, PCA shall be performed by using the correlation matrix.

b)

The first two PCs are such that their mean and variances are:

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

$$\sigma^2_{PC1} = \lambda_1 = 1.33261 \, , \sigma^2_{PC2} = \lambda_2 = 0.85923$$

Thus, 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 (Bonferroni's correction) is $\alpha^* = 0.01/2 = 0.005$

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

PC1

$$UCL = 3.240$$
$$CL = 0$$
$$LCL = -3.240$$

PC2

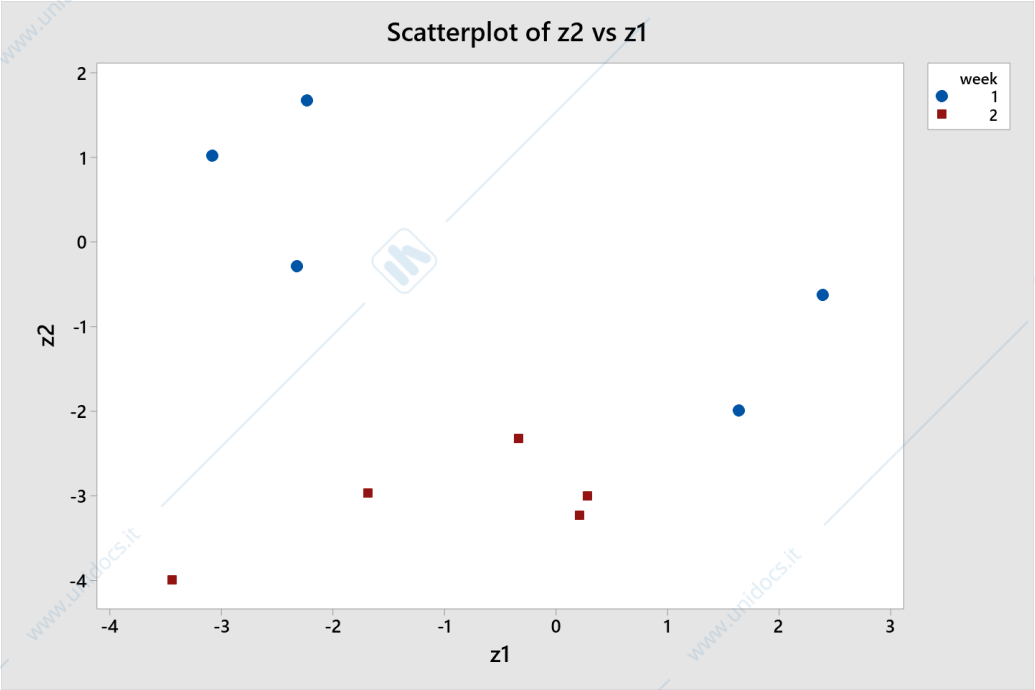
$$UCL = 2.602$$
$$CL = 0$$
$$LCL = -2.602$$

c)

Since we are using the correlation matrix for the PCA, we need to standardize the data before projecting them onto the space spanned by the first two PCs. The resulting projections (scores) are the following:

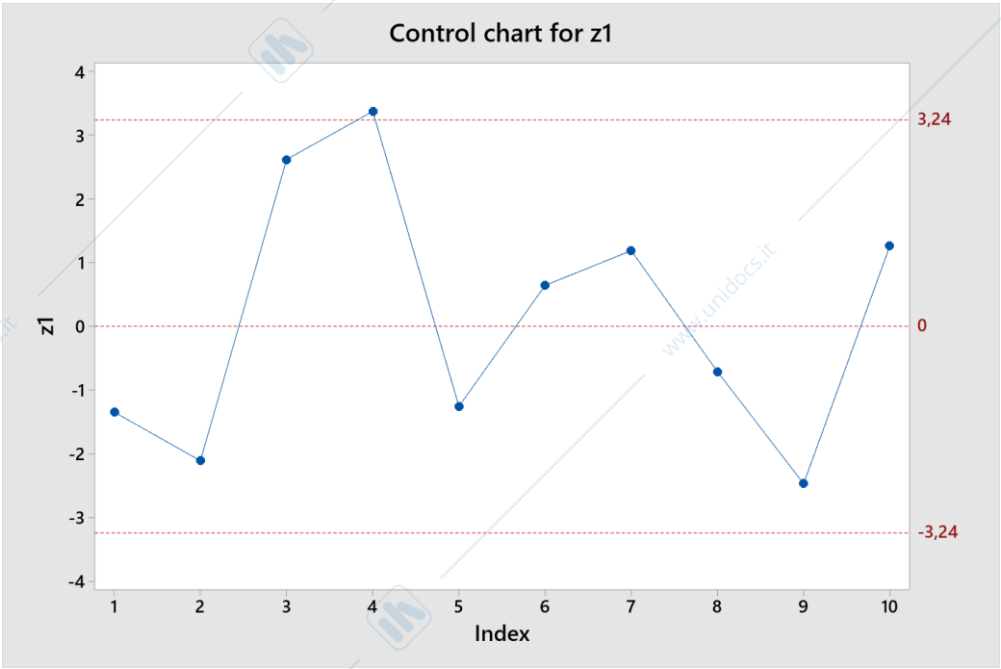
Week	Z1	Z2
1	-1,34463	-0,42766
1	-2,10769	0,88815
1	2,62281	-2,13246
1	3,37653	-0,76061
1	-1,25391	1,54289
2	0,64558	-2,46423
2	1,19008	-3,37736
2	-0,70954	-3,11090
2	-2,46853	-4,13793
2	1,26364	-3,14410

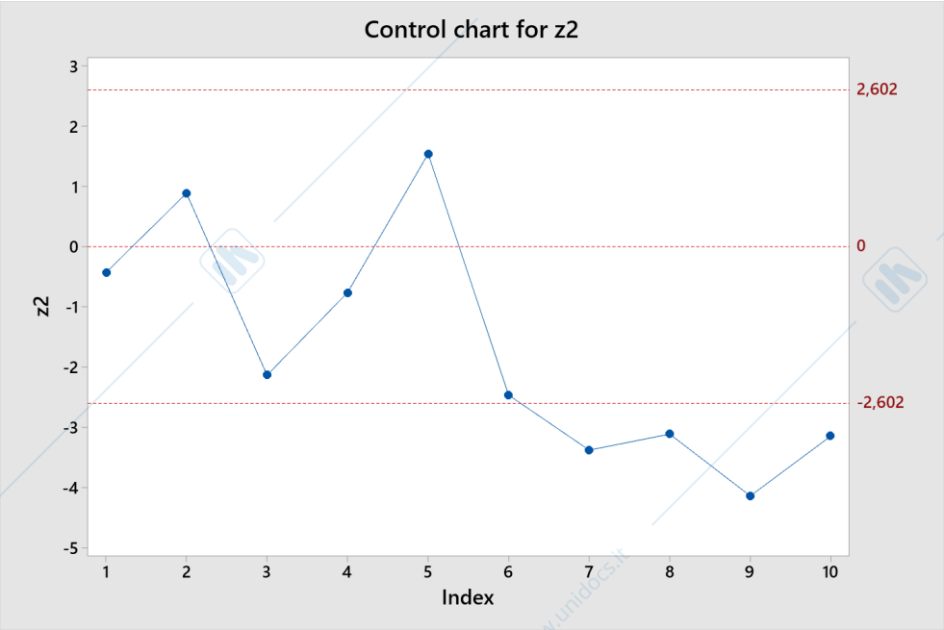
Scatter plot of the scores:



The data acquired during week 2 exhibit a lower mean than those acquired in week 1 when projected along the second PC.

The two control charts are the following:





The control charts show that during the second week the process exhibited a shift with measurements from day 6 (Monday of week 2) and day 10 (Friday of week 2) signalled as out of control. The control chart on the first PC signals an out-of-control only on Thursday of week 1 (day 4).

Since the second PC is a contrast between pH and protein content, the data observed in week 2 exhibit a contrast between these two variables that is statistically different from the one observed in week 1 and during the control chart design phase.

Exercise 3 solution

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
- $S_{xx} = \sum_{i=1}^{n}(t_i - \bar{t})^2$

By using the Shewart'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 drift curve slopes for different tools. It can be possibly combined with a control chart on $\hat{\sigma}_\varepsilon^2$, to monitor the model residuals as well.