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	4,493	21	0,084	41	1,122
2	-1,378	22	1,091	42	3,786
3	1,761	23	-3,934	43	-1,206
4	0,313	24	2,508	44	4,625
5	-1,503	25	0,724	45	5,035
6	3,894	26	2,091	46	1,528
7	2,796	27	6,610	47	5,170
8	5,913	28	2,028	48	0,946
9	2,969	29	1,551	49	3,528
10	3,264	30	-0,970	50	-1,571
11	2,956	31	5,817	51	5,044
12	3,540	32	8,498	52	-0,582
13	1,738	33	6,000	53	4,123
14	2,629	34	0,882	54	9,169
15	11,267	35	10,780	55	8,543
16	12,114	36	3,505	56	7,463
17	8,626	37	3,600	57	7,912
18	15,198	38	14,143	58	6,094
19	16,545	39	14,318	59	18,129
20	14,186	40	20,353	60	9,044

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:

$$\boldsymbol{\mu} = [4.59\ 8.81\ 157.25]'$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.02 & 0.005 & 0.018 \\ 0.005 & 0.09 & 0.035 \\ 0.018 & 0.035 & 0.38 \end{bmatrix}$$

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	4,69	9,15	157,44
	Tuesday	4,48	8,9	157,26
	Wednesday	4,11	8,69	157,03
	Thursday	4,98	9,24	157,71
	Friday	4,88	8,78	157,27
Week 2	Monday	4,32	9,19	157,05
	Tuesday	4,17	9,25	157,03
	Wednesday	4,21	9,31	158,54
	Thursday	4,13	9,61	159,94
	Friday	4,27	9,32	156,43

Exercise 3 (3 points)

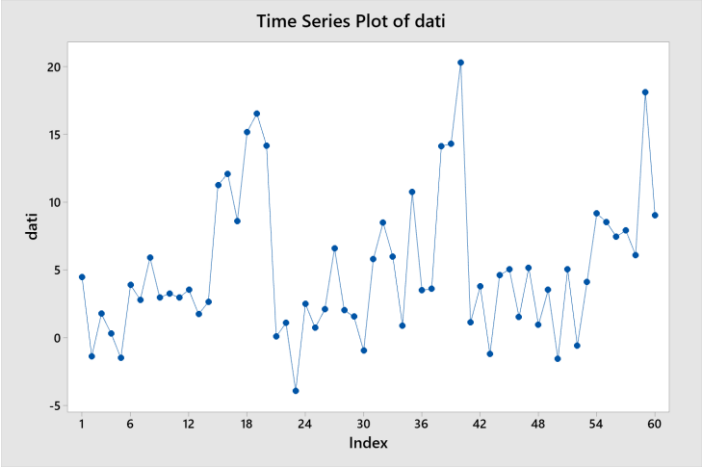
A company is interested in observing the stability of the linear drift of a tool. To this aim, for each tool, a linear model is fitted $y = \beta_0 + \beta_1 t + \varepsilon_t$ starting from the same number n of data observed in each drift curve.

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

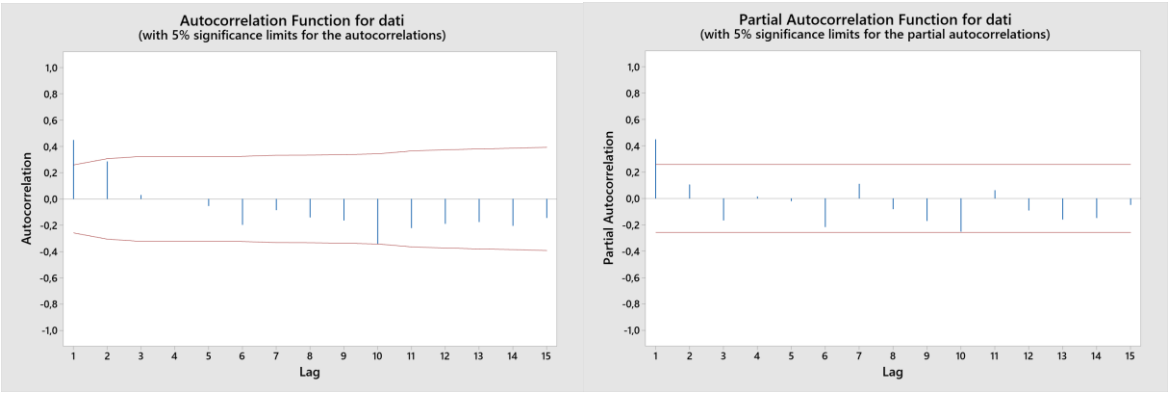
Test

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	29,37	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_A

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 = 1,297
+ 0,001747 minitrend^3

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1,297	0,543	2,39	0,020	
minitrend^3	0,001747	0,000165	10,58	0,000	1,00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3,12293	65,86%	65,27%	62,79%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1091,2	1091,18	111,88	0,000
minitrend^3	1	1091,2	1091,18	111,88	0,000
Error	58	565,7	9,75		
Lack-of-Fit	18	183,8	10,21	1,07	0,414
Pure Error	40	381,9	9,55		
Total	59	1656,8			

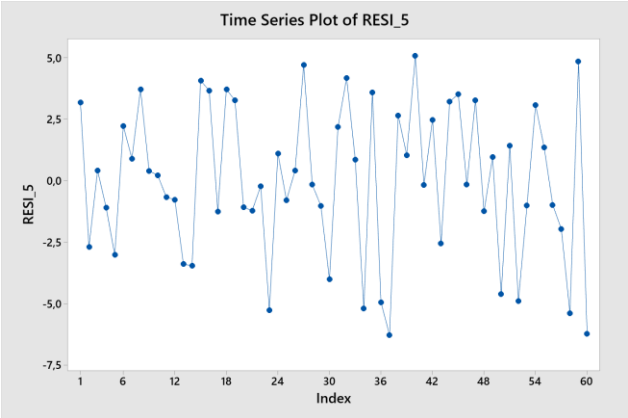
Fits and Diagnostics for Unusual Observations

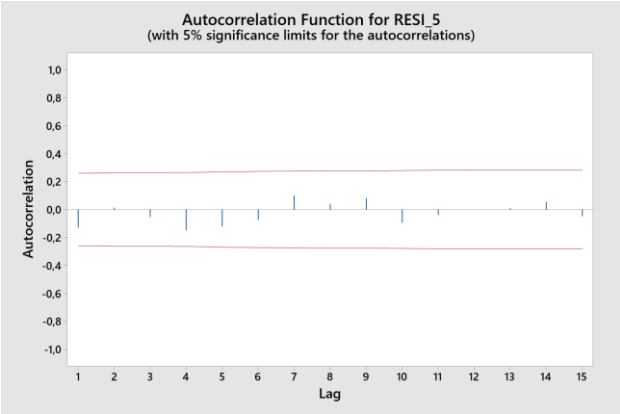
Obs	dati	Fit	Resid	Std Resid	
20	14,186	15,271	-1,085	-0,37	X
37	3,600	9,878	-6,278	-2,05	R
40	20,353	15,271	5,082	1,73	X
60	9,044	15,271	-6,227	-2,11	R X

R Large residual
X Unusual X

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.

In this case the final residuals satisfy the assumptions:



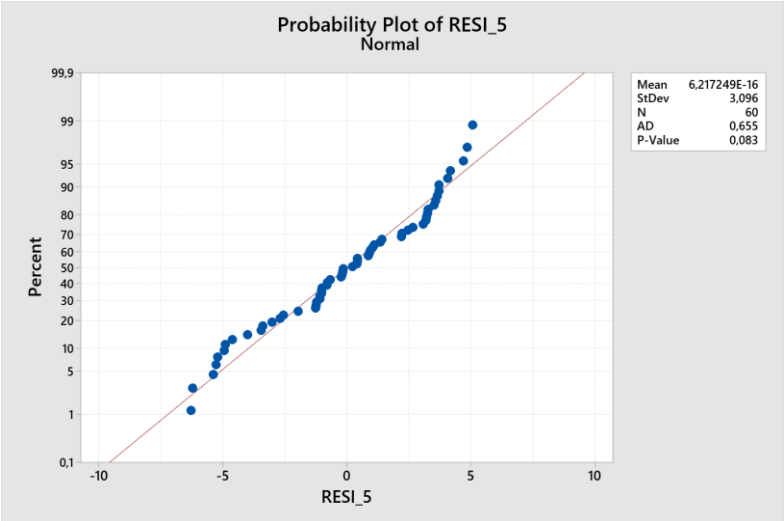


Test

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
34	31,00	0,435



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:

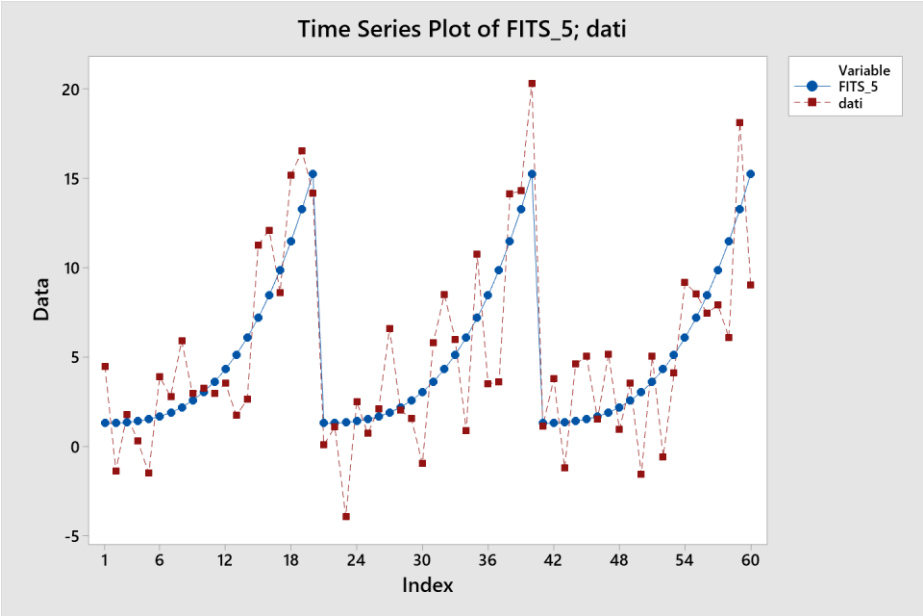
$$ARL(H_0) = \frac{ATS}{\Delta T} = \frac{240}{1} = 240$$

Thus:

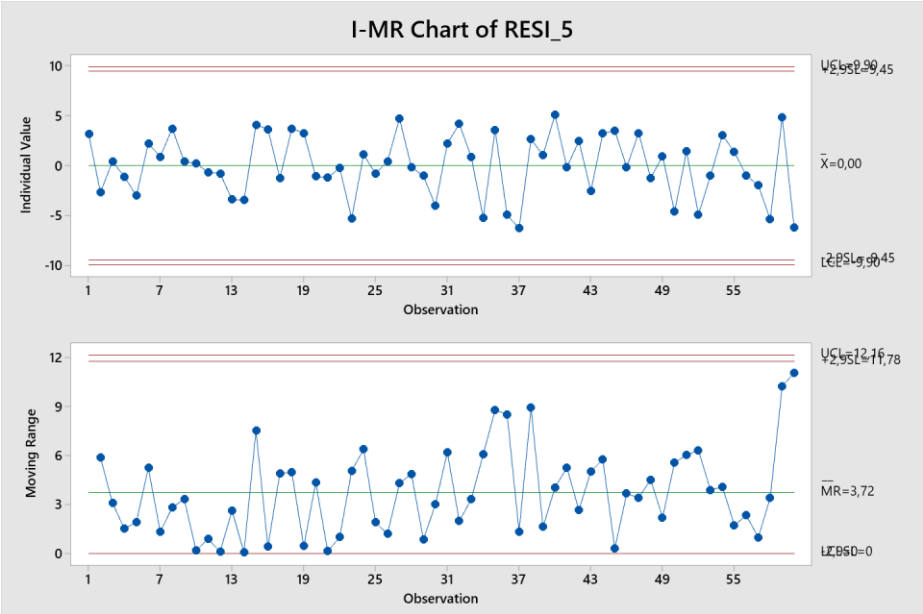
$$\alpha = \frac{1}{ARL(H_0)} = \frac{1}{240} = 0{,}0042$$

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

The Fitted value chart is the following:



The special cause control chart is the following:



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 = 1,297
+ 0,001747 minitrend^3

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1,297	0,543	2,39	0,020	
minitrend^3	0,001747	0,000165	10,58	0,000	1,00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3,12293	65,86%	65,27%	62,79%

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

	Model	Normality	LBQ (lag=5)	R2adj
degradation curve 1	**Coefficients** **Term** **Coef** **SE Coef** **T-Value** **P-Value** **VIF** minitrend^3 0,002203 0,000184 11,95 0,000 1,00	p-val: 0.635	p-val: 0.5029	87,64%
degradation curve 2	**Coefficients** **Term** **Coef** **SE Coef** **T-Value** **P-Value** **VIF** minitrend^3 0,002098 0,000237 8,87 0,000 1,00	p-val: 0.144	p-val: 0.9265	79,51%
degradation curve 3	**Coefficients** **Term** **Coef** **SE Coef** **T-Value** **P-Value** **VIF** minitrend^3 0,001732 0,000234 7,41 0,000 1,00	p-val: 0.456	p-val: 0.5480	72,93%

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 constant term, which was weekly significant in the single model, is not statistically significant with a confidence of 95% in the three distinct models.

The values of coefficient for the cubic term are close to 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:

$\lambda_1 = 0.385098$

$\lambda_2 = 0.085956$

$\lambda_3 = 0.018946$

The first PC explains 78.6% of the overall variability. Its loadings are the following:

Matrix EIGVET

0,050514
0,118473
0,991672

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:

$$\mathbf{P} = \begin{bmatrix} 1 & 0.1179 & 0.2065 \\ 0.1179 & 1 & 0.1893 \\ 0.2065 & 0.1893 & 1 \end{bmatrix}$$

The eigenvalues (i.e., the variances of the PCs) are:

$\lambda_1 = 1.34501$ (explained variance: 44.8%)

$\lambda_2 = 0.88283$ (explained variance: 29.4%)

$\lambda_3 = 0.77215$

The first two PCs explain 74.2% of the overall variability. Their loadings are the following:

Matrix EIGVETR

-0,560626 0,657142
-0,537187 -0,751673
-0,630181 0,056140

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:

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

$$\sigma^2_{PC1} = \lambda_1 = 1.34501,$$

$$\sigma^2_{PC2} = \lambda_2 = 0.88283$$

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

$$UCL = 3.255$$
$$CL = 0$$
$$LCL = -3.255$$

PC2

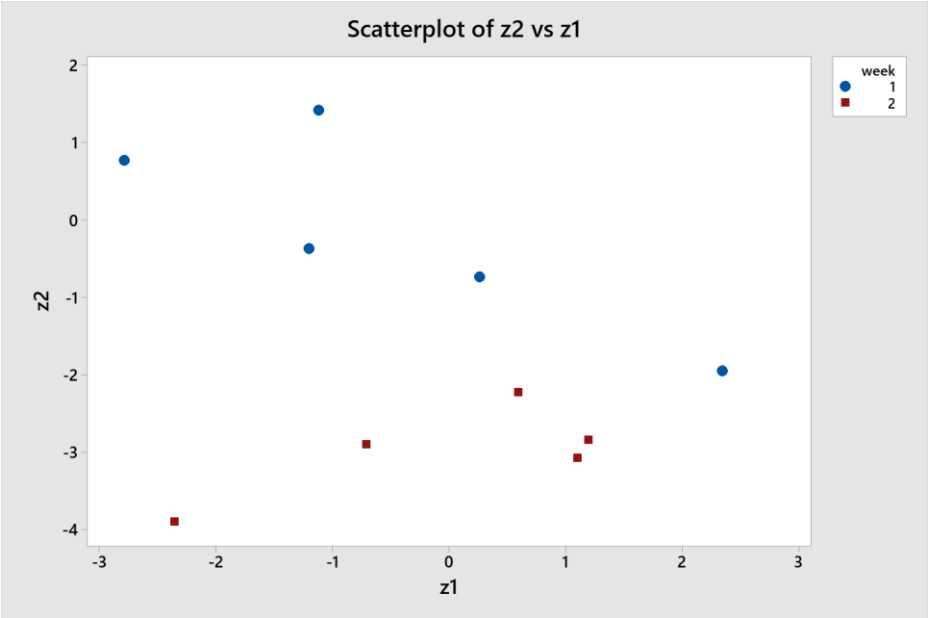
$$UCL = 2.637$$
$$CL = 0$$
$$LCL = -2.637$$

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:

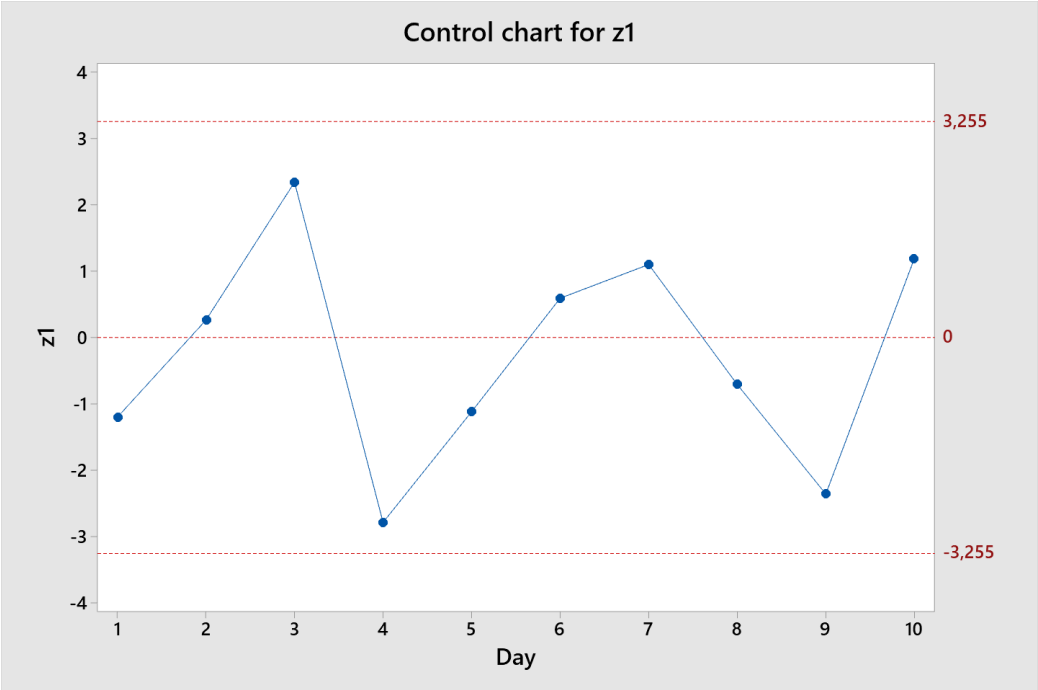
Week	Z1	Z2
1	-1,19947	-0,36992
1	0,26469	-0,73573
1	2,34261	-1,94978
1	-2,78627	0,77671
1	-1,11635	1,42453
2	0,59436	-2,22494
2	1,10200	-3,07410
2	-0,70766	-2,90105
2	-2,35891	-3,89696
2	1,19361	-2,83947

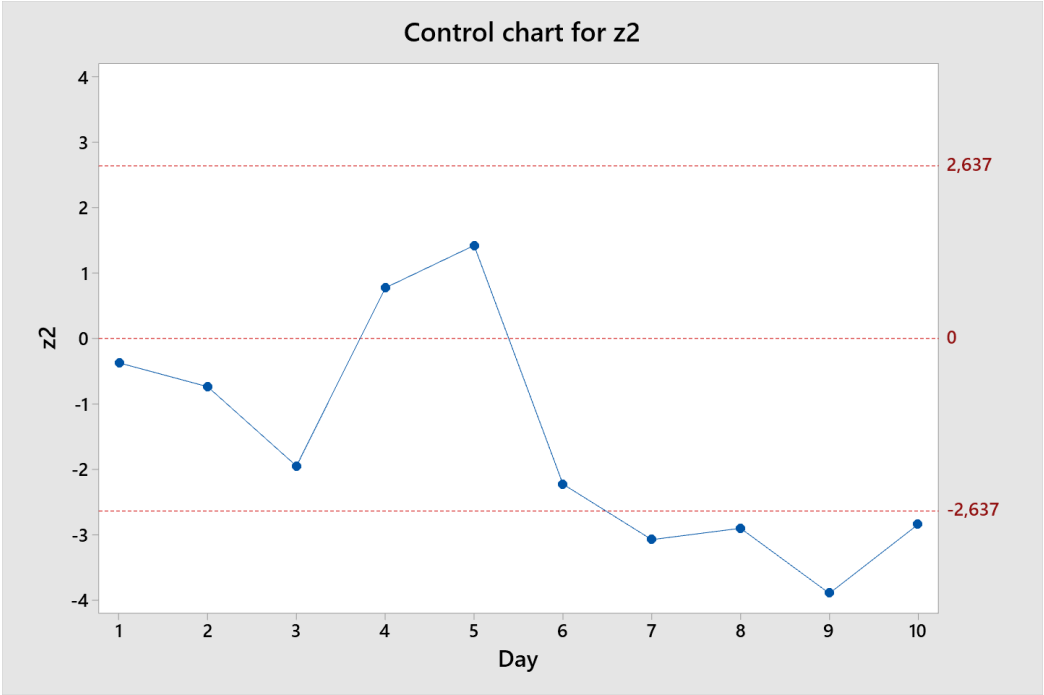
Scatter plot of the scores:



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. No out-of-control pattern is signalled by control chart on the first PC, instead.

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.