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)

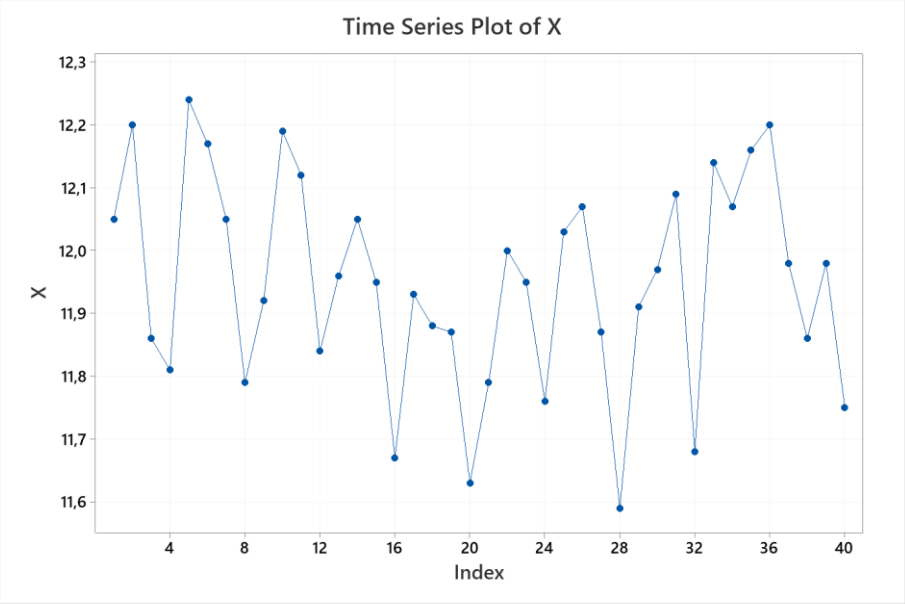
Using the sample statistics defined in point e) of Exercise 2, report the eigenvalue and eigenvector corresponding to the first PC. Show that the variance explained by this first PC is higher than the variance explained by a simple linear combination of the four variables where equal weight is given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

Solutions

Exercise 1

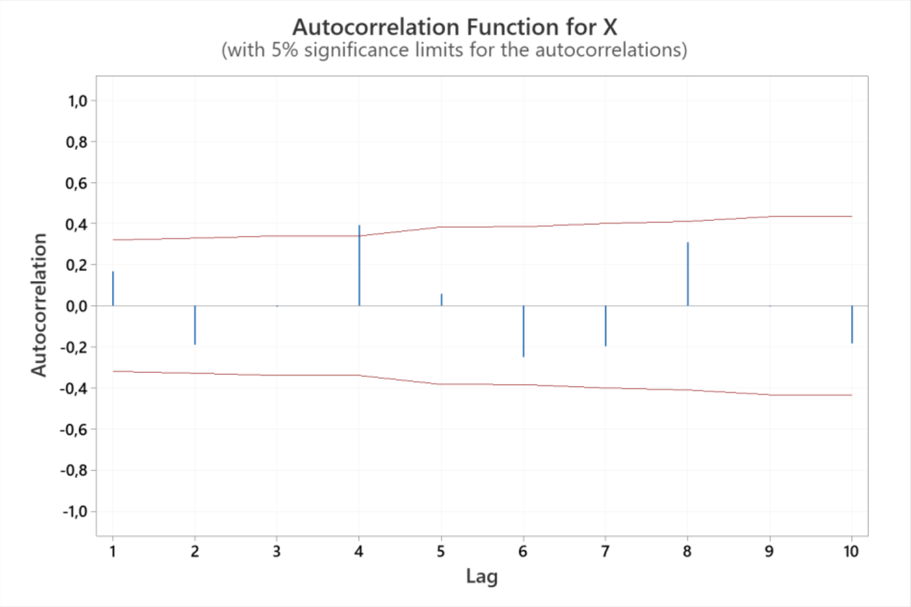
a)

Time series plot of the spindle power data:

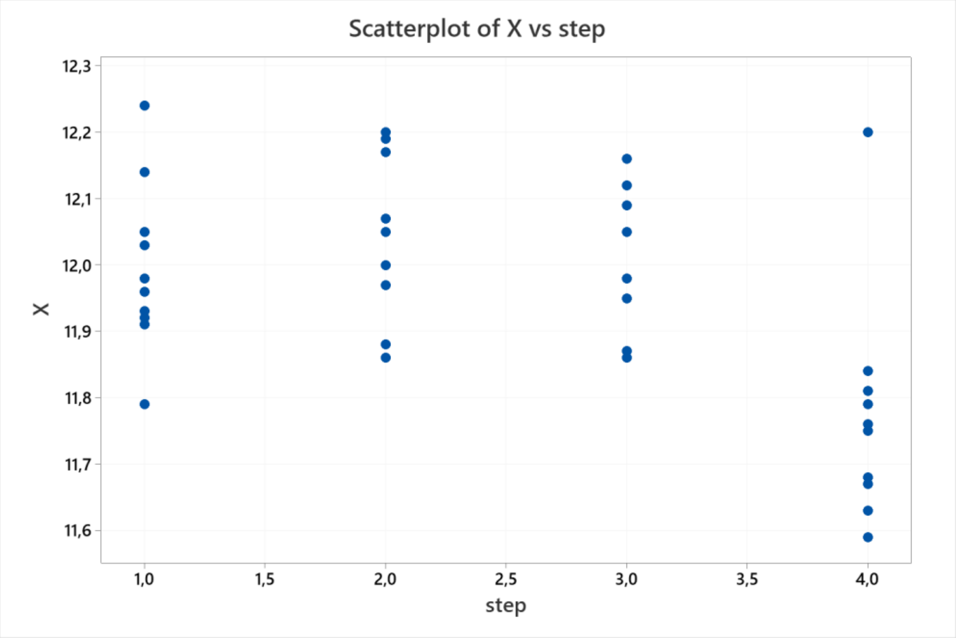


There is a meandering pattern with a seasonal drop of the spindle power in correspondence of step 4.

This lag 4 effect is also visible from the SACF:

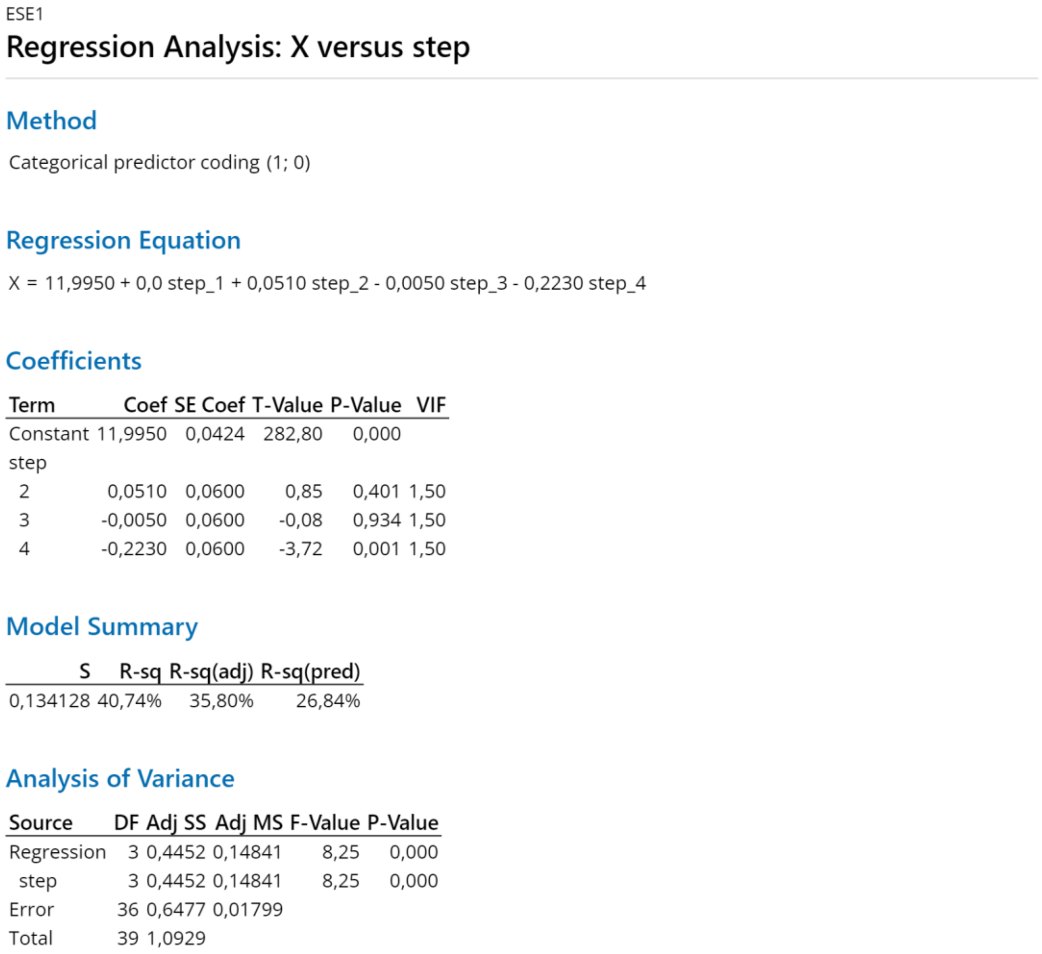


The way in which the power varies along the different process steps is shown in the following scatter plot:



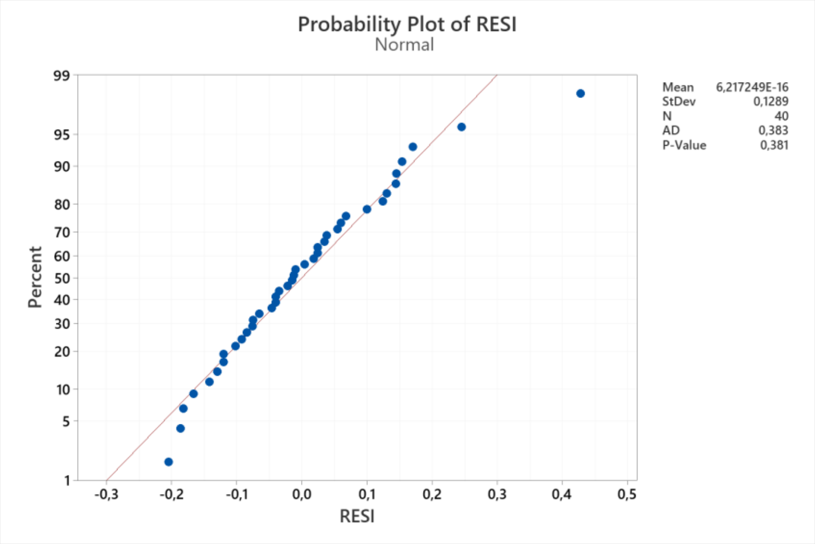
A part from one apparent outlying value, step 4 (finishing) yields a lower power consumption, as expected.

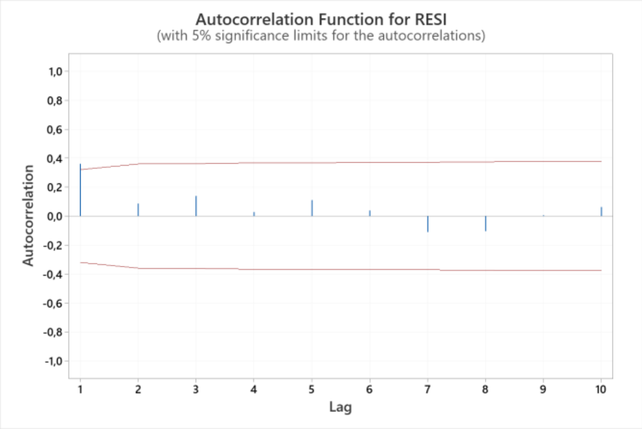
Based on this, it would be possible to fit a model of the spindle power using as regressor the categorical variable “step” as follows:

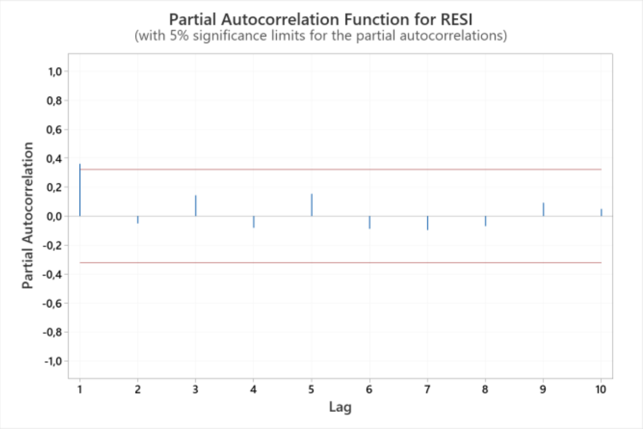


```
ESE1
Regression Analysis: X versus step

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:





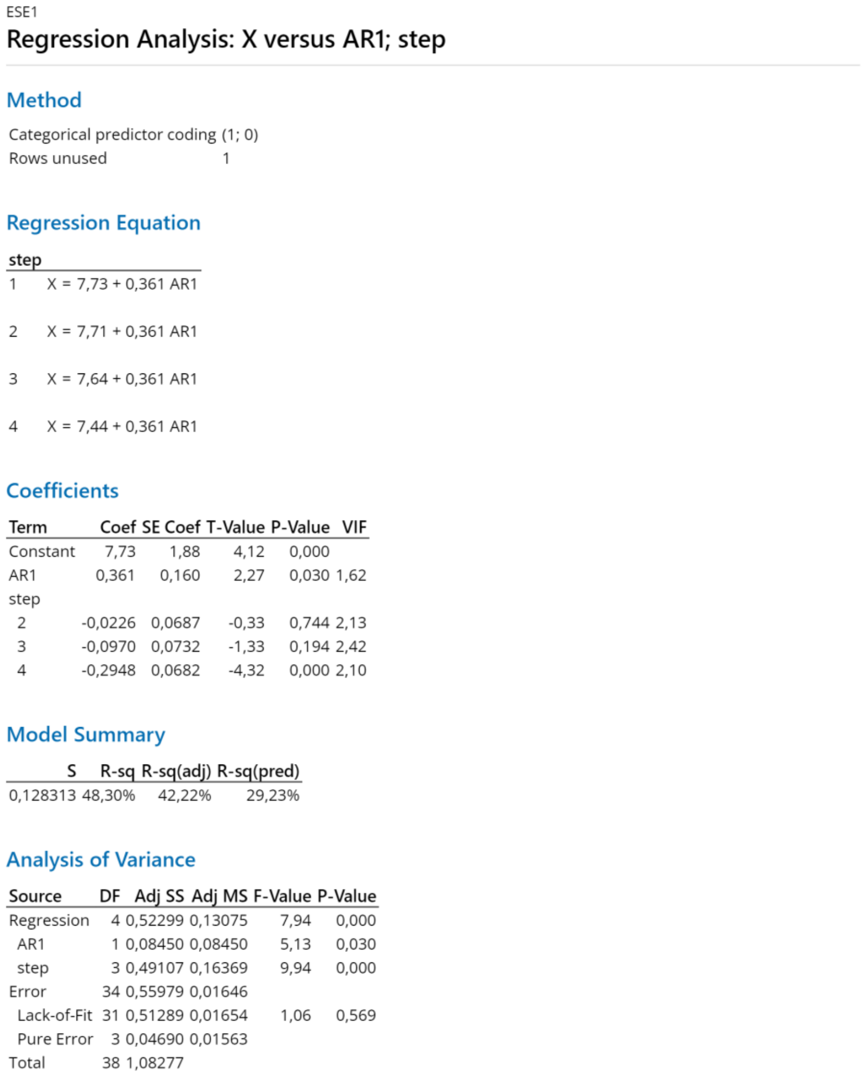


Bartlett’s test at lag = 1 (95% confidence):

$$|r_k| = 0.361$$

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

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

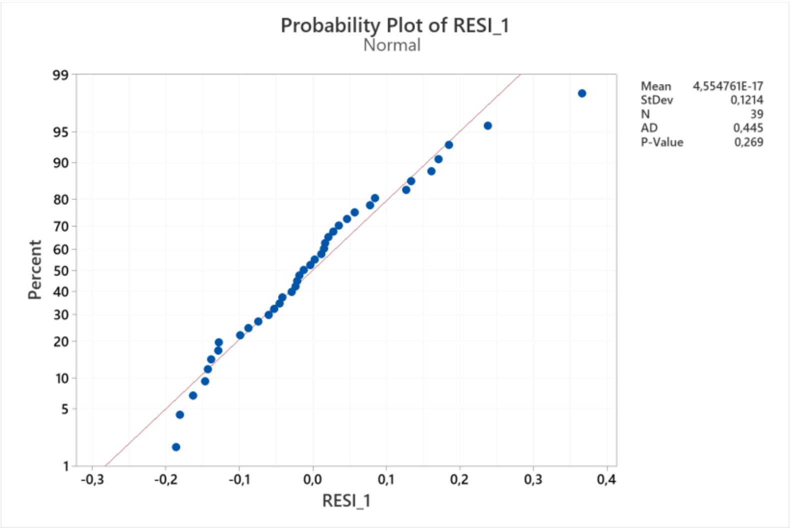
Model Summary

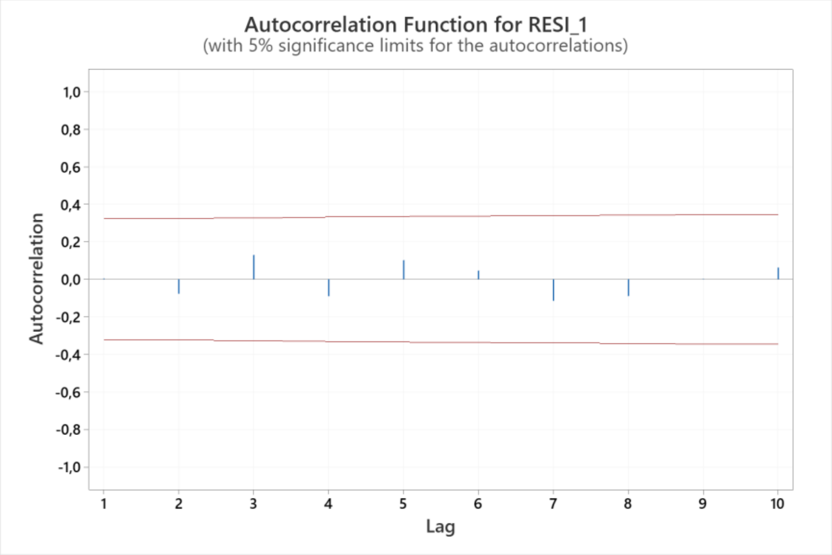
S	R-sq	R-sq(adj)	R-sq(pred)
0,128313	48,30%	42,22%	29,23%

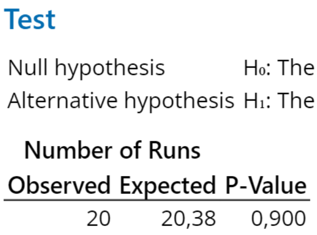
Analysis of Variance

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.







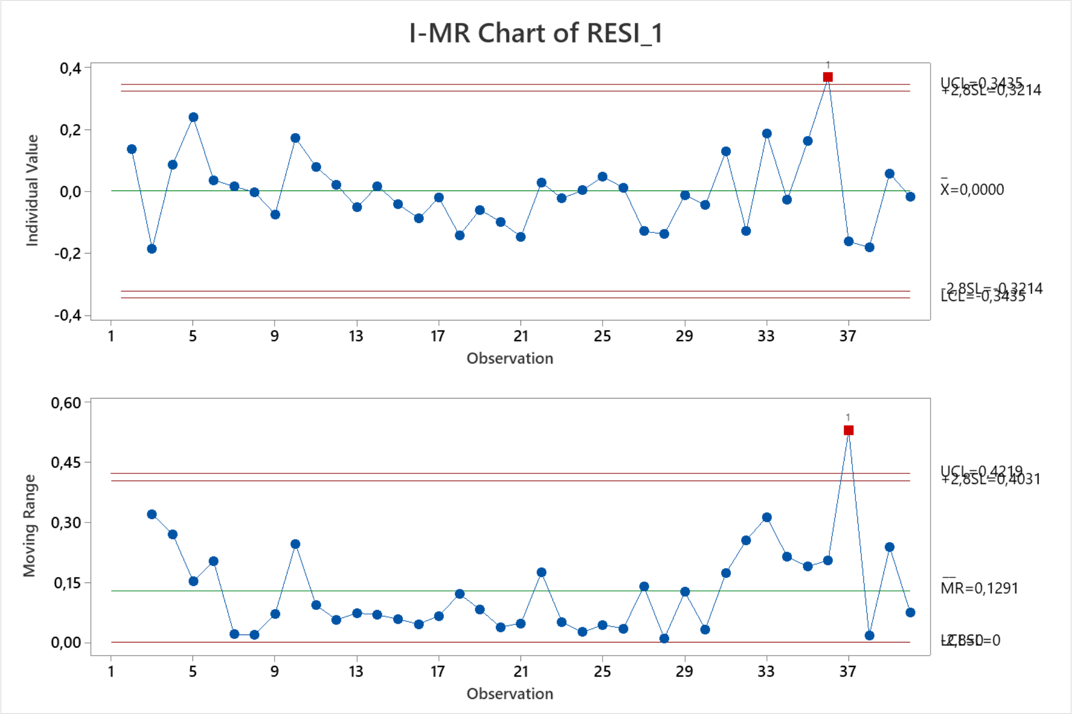
Test

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

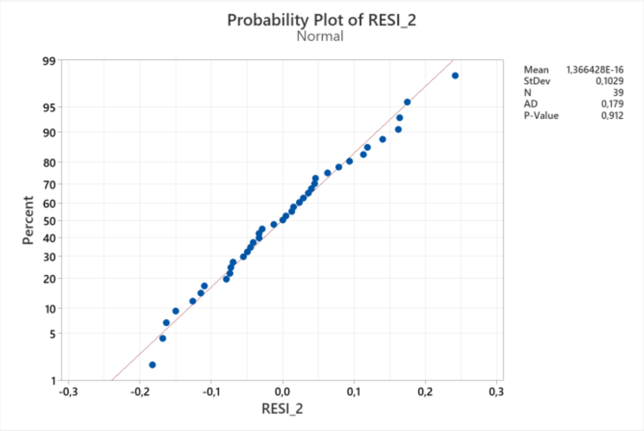
Number of Runs		
Observed	Expected	P-Value
20	20,38	0,900

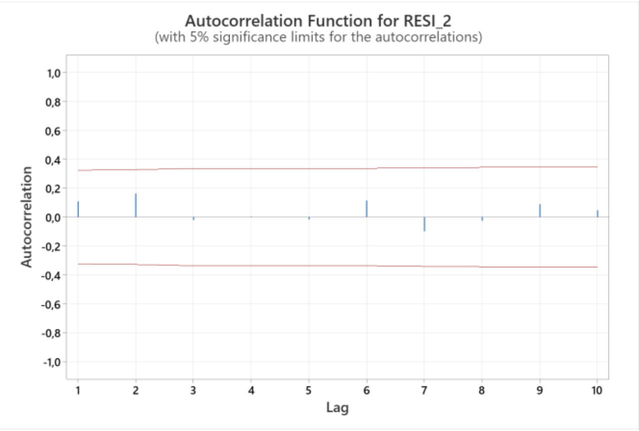
b)

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).

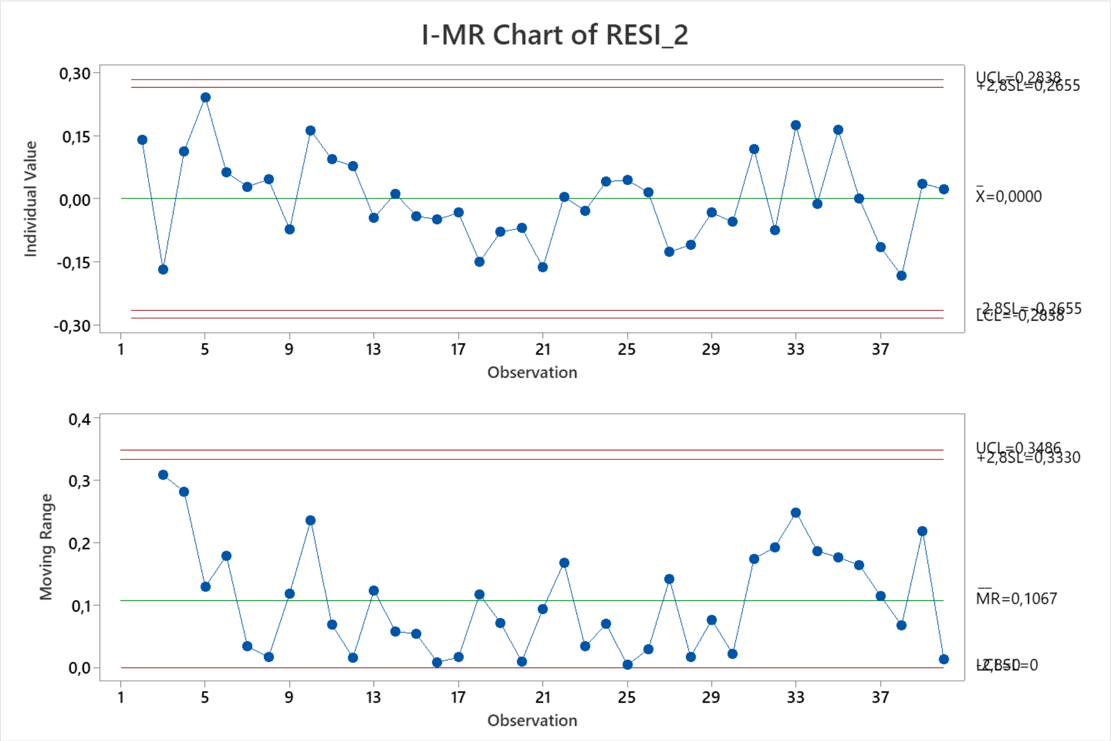


Observation 36 violates the control limits of both charts. Assuming the existence of an assignable cause, it is possible to introduce a dummy variable that is equal to 1 for this observation and 0 elsewhere. The new model is still appropriate, with normal and independent residuals:





The resulting control chart is the following:

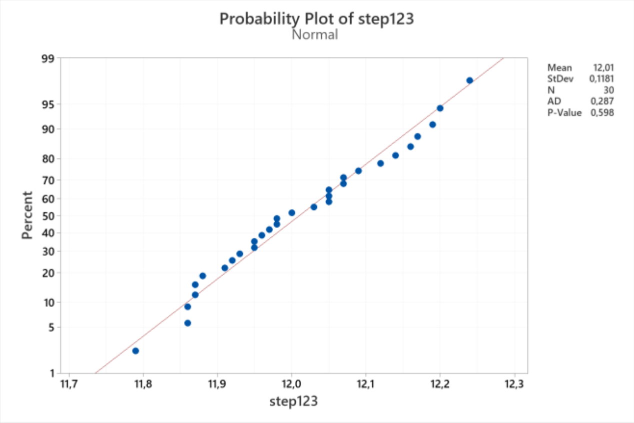


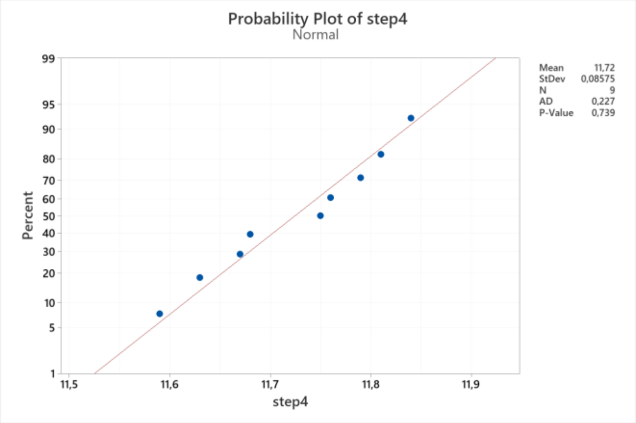
No further violation is present. The design phase is over.

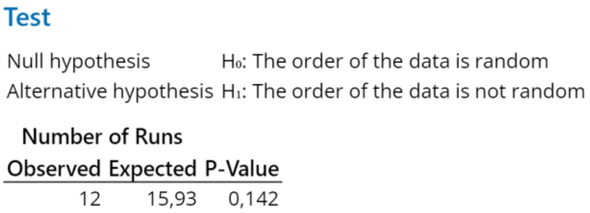
<u>c)</u>

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).

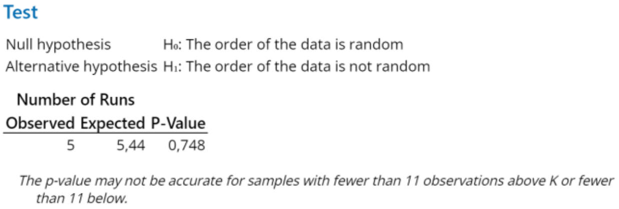






Test

```
Null hypothesis          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
      12    15,93   0,142
```



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
       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)

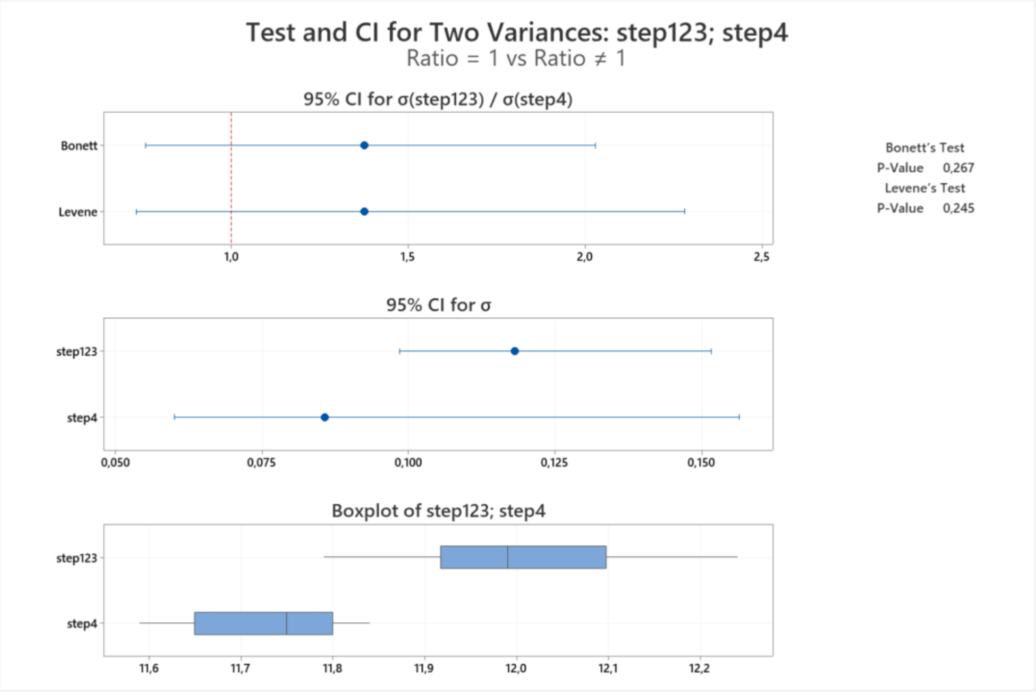
Ratio of Standard Deviations

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

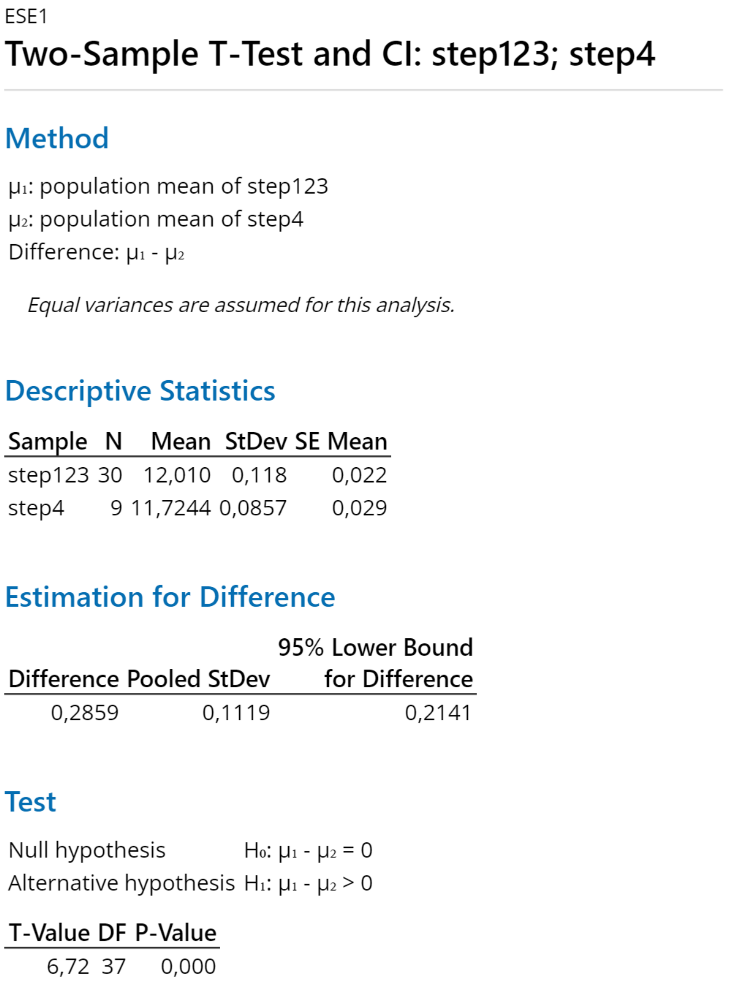
Test

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:



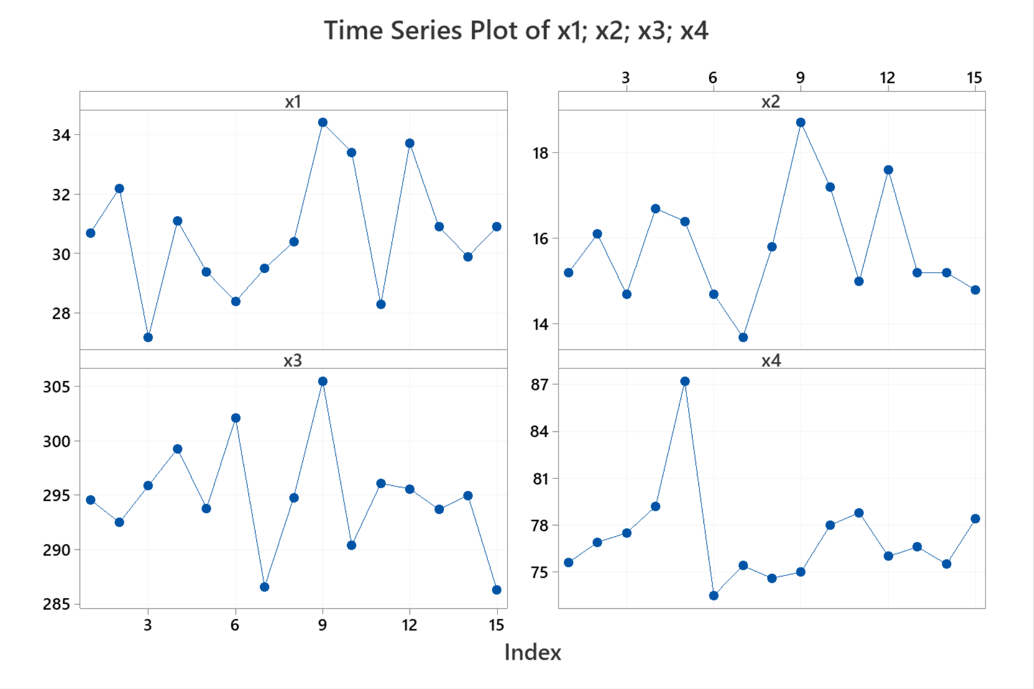
```
ESE1
Two-Sample T-Test and CI: step123; step4

Method

μ₁: population mean of step123
μ₂: population mean of step4
Difference: μ₁ - μ₂

   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

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

Test

Null hypothesis        H₀: μ₁ - μ₂ = 0
Alternative hypothesis H₁: μ₁ - μ₂ > 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.

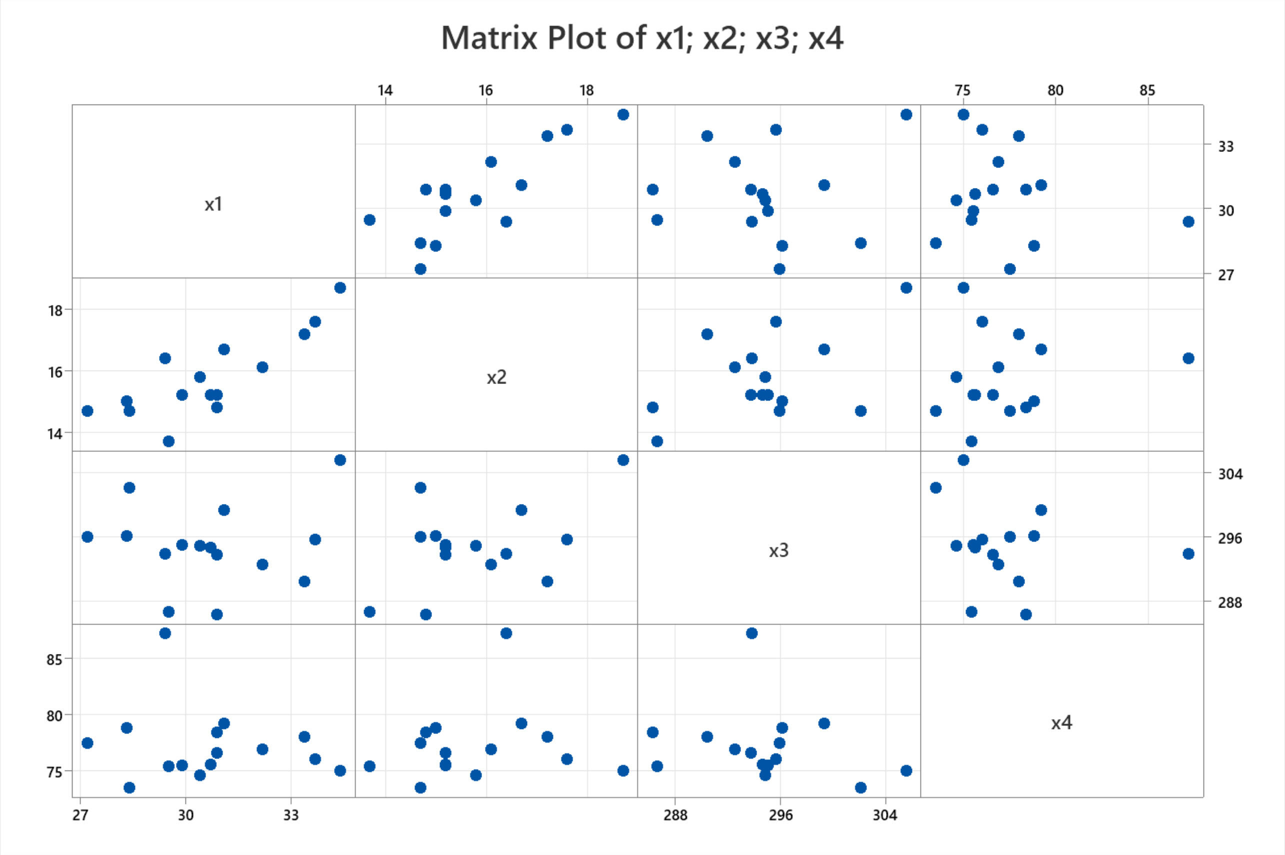
Exercise 2

a)

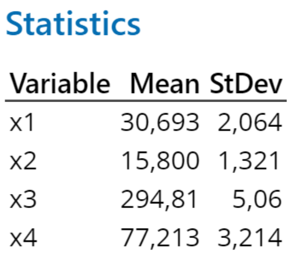
The time series plot of the four variables:



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.

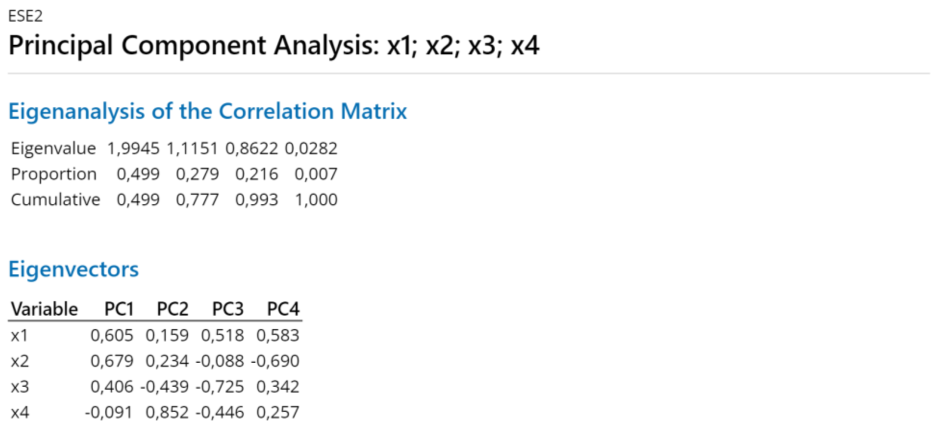


Statistics

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

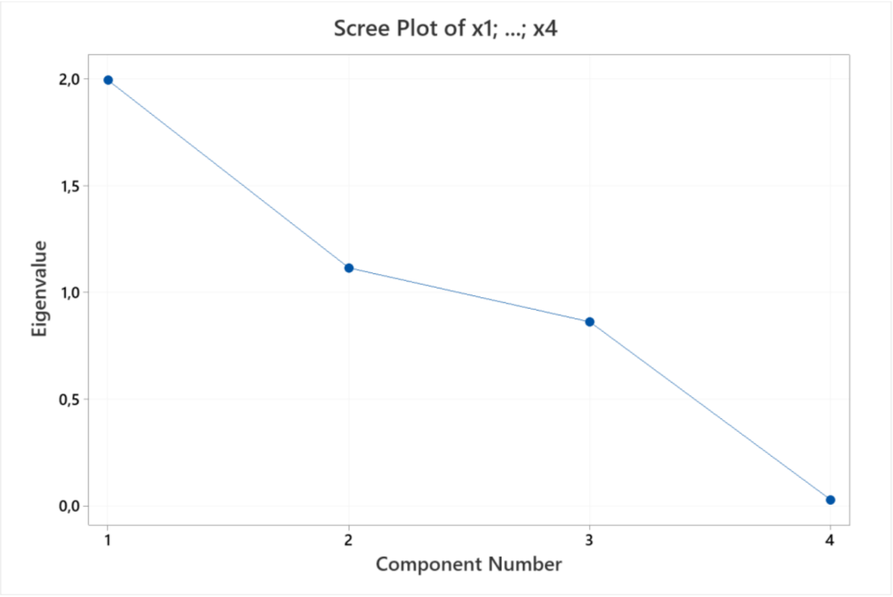
Principal Component Analysis: x1; x2; x3; x4

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

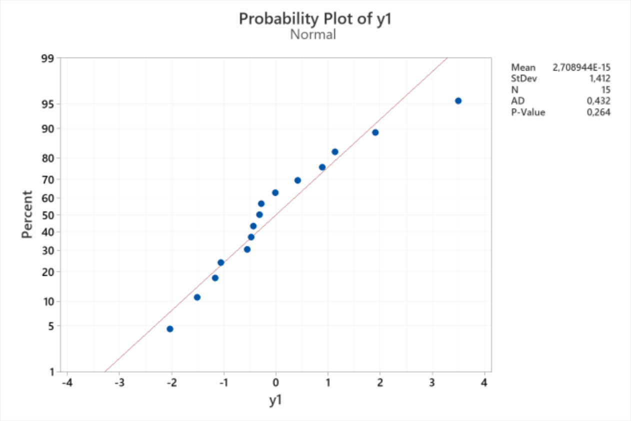


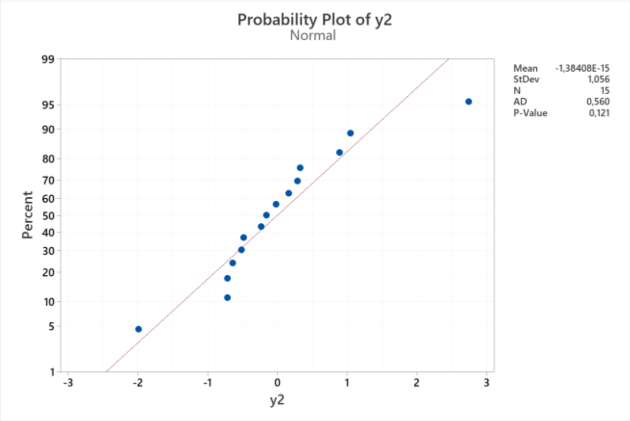
The number of PCs to retain to explain at least 75% of the overall data variability is 2.

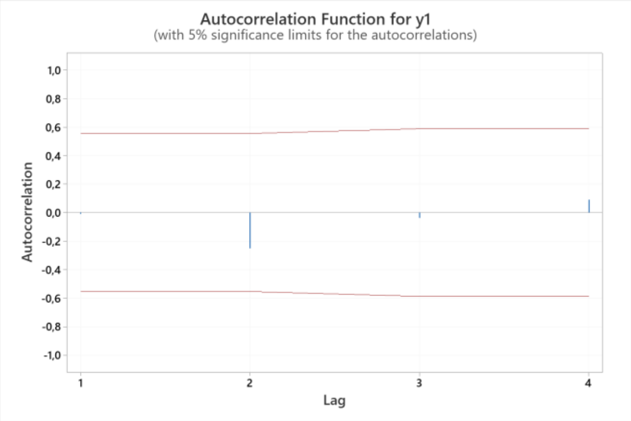
c)

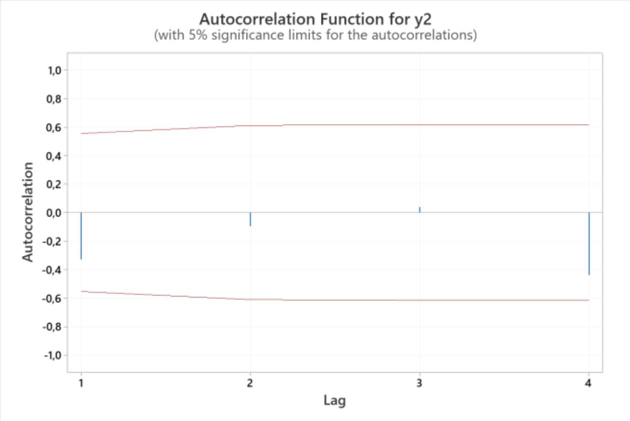
Before designing the T^2 control chart on the scores of the first 2 PCs, assumptions shall checked (marginal normality is assumed as a sufficient condition for multivariate normality).

The scores of the two PCs are normal and independent.


Probability Plot of y1
Normal

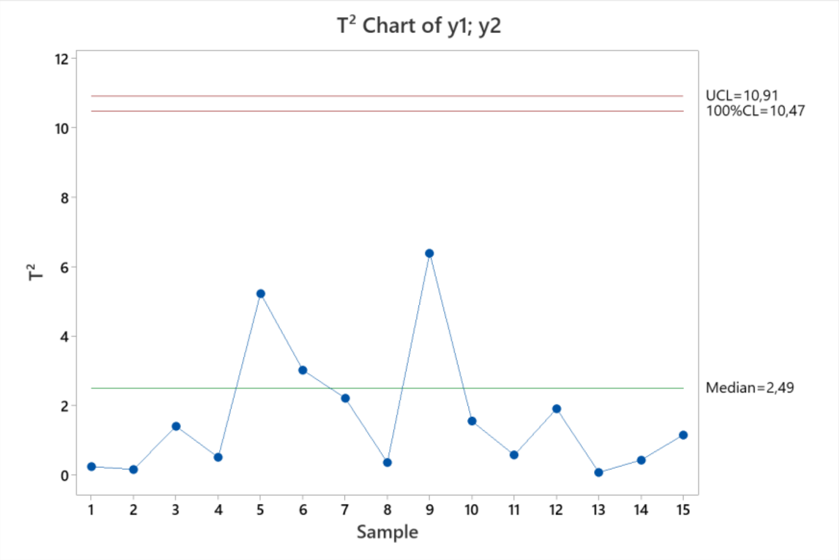

Probability Plot of y2
Normal


Autocorrelation Function for y1
(with 5% significance limits for the autocorrelations)


Autocorrelation Function for y2
(with 5% significance limits for the autocorrelations)

```
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
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).

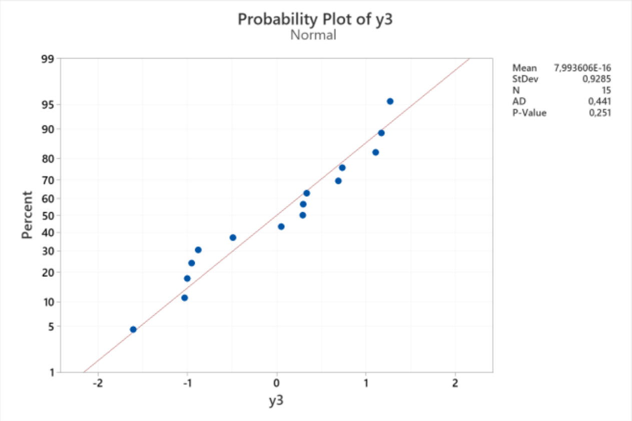

T^2 Chart of y1; y2

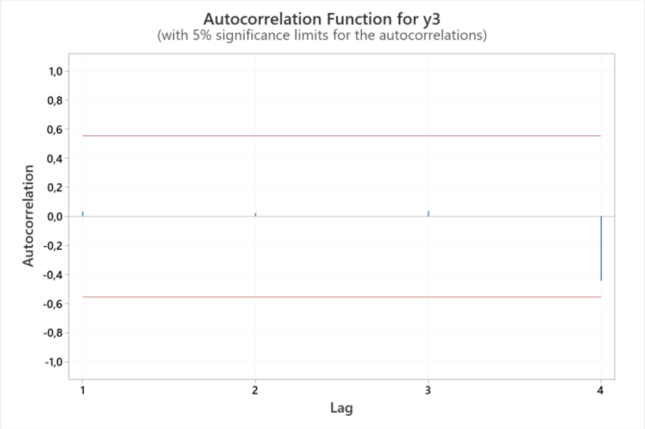
The process is in-control.

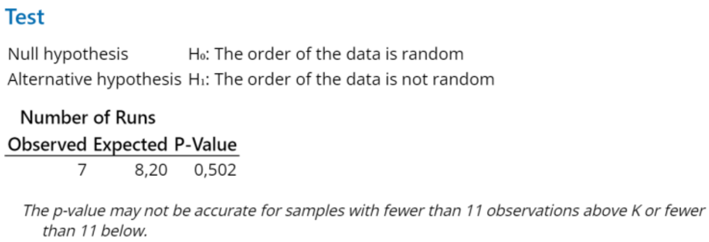
d)

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.

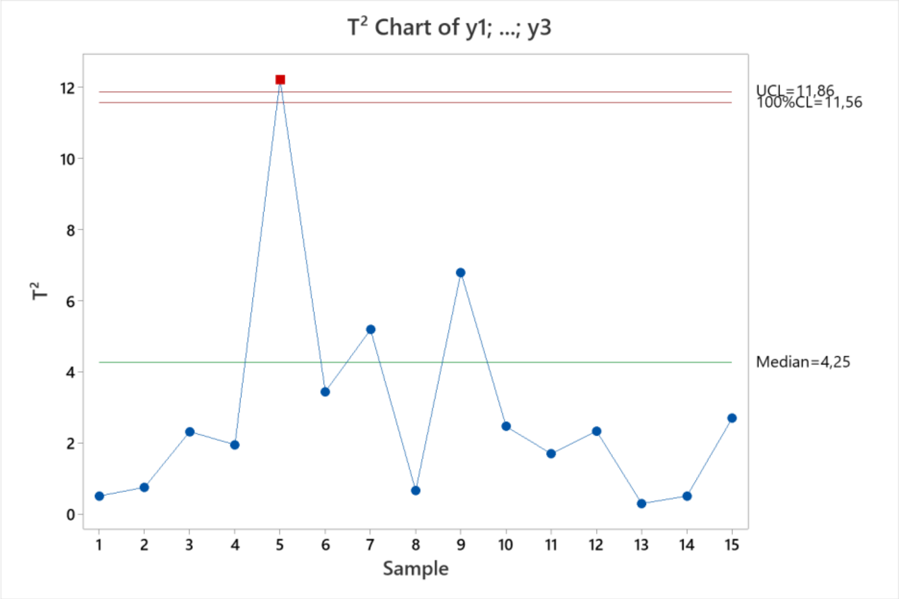




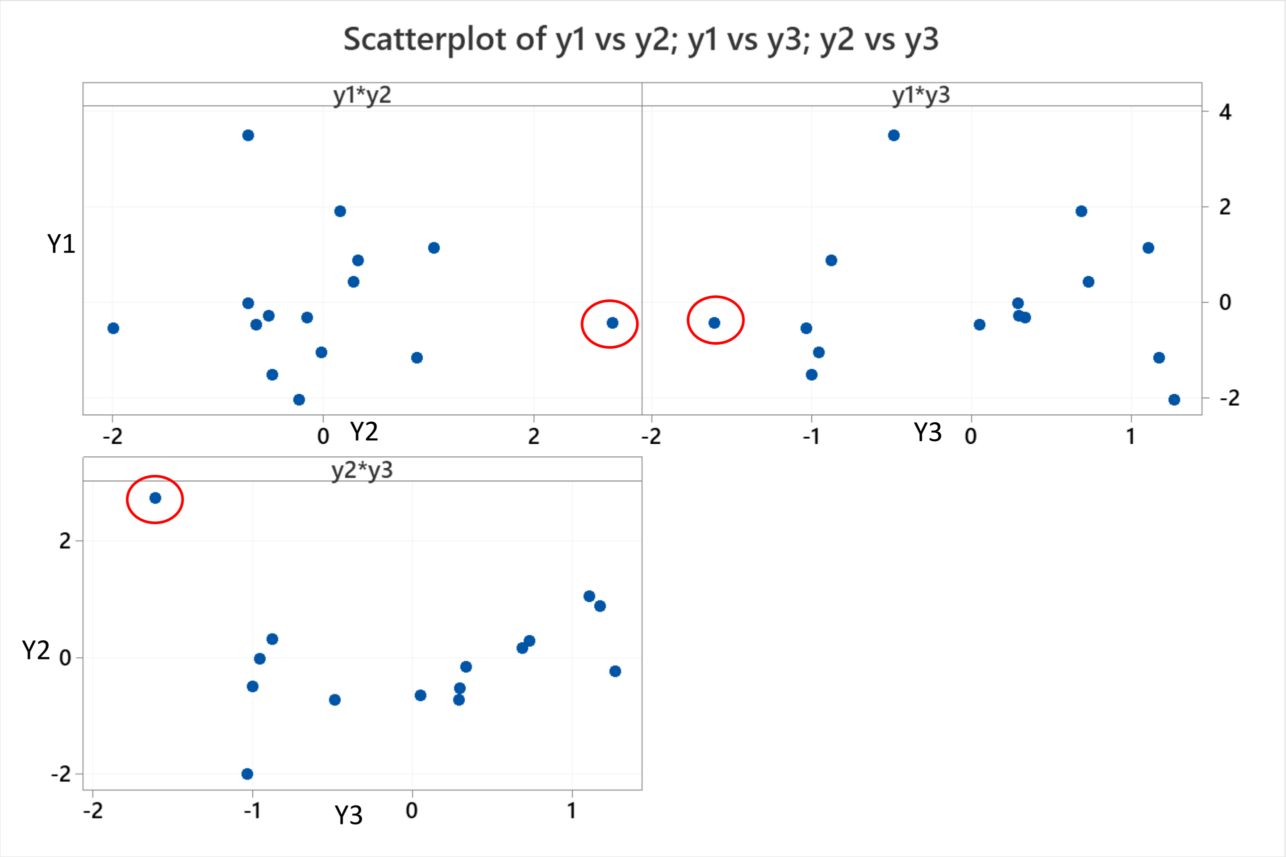


```
Test

Null hypothesis        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
       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.
```

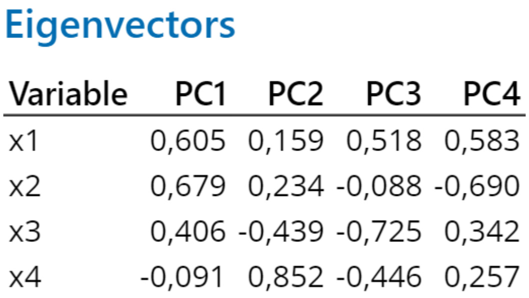
The resulting control chart is the following:



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:

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

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

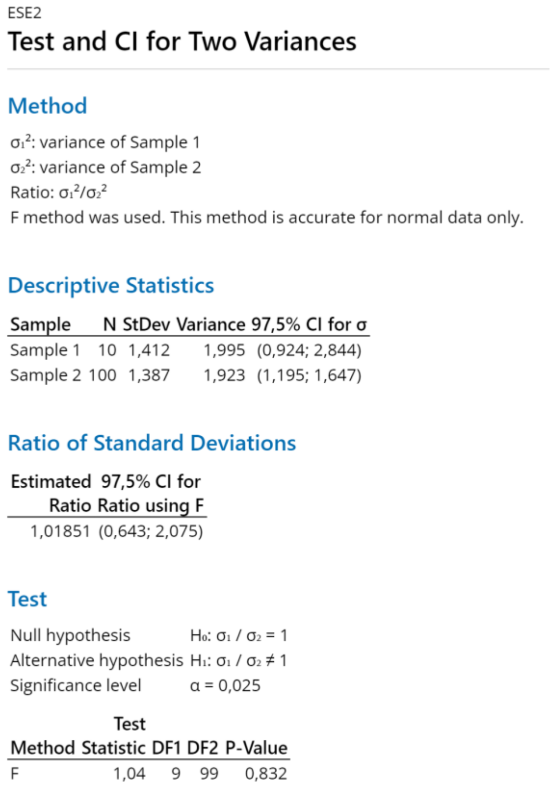
λ2 = 1,11505

λ3 = 0,86217

λ4 = 0,02823

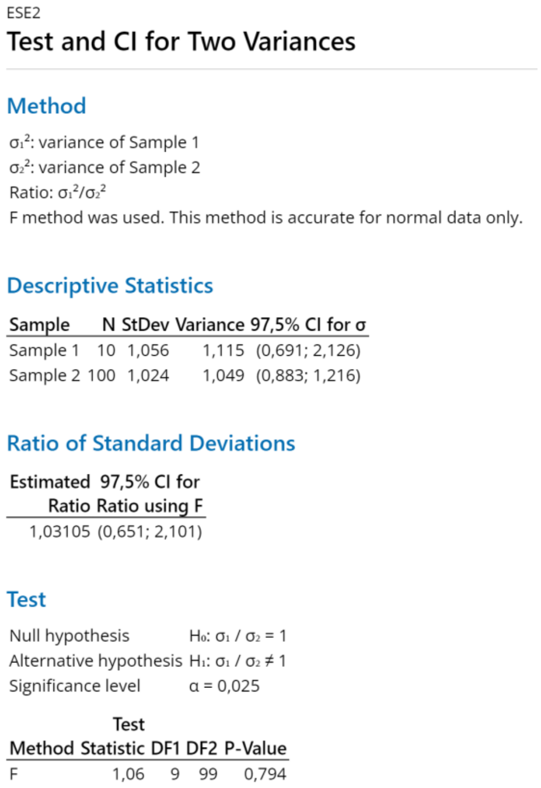
Considering a familywise $\alpha = 0.05$ and assuming that normality and randomness hold for the new data as well, and that the two samples are independent, the following two tests can be designed to compare the explained variances of PC1 and PC2:

PC1:



```
ESE2
Test and CI for Two Variances

Method
σ₁²: variance of Sample 1
σ₂²: variance of Sample 2
Ratio: σ₁²/σ₂²
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  97,5% CI for
    Ratio Ratio using F
  1,01851  (0,643; 2,075)

Test
Null hypothesis         H₀: σ₁ / σ₂ = 1
Alternative hypothesis  H₁: σ₁ / σ₂ ≠ 1
Significance level      α = 0,025

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

PC2:



```
ESE2
Test and CI for Two Variances

Method
σ₁²: variance of Sample 1
σ₂²: variance of Sample 2
Ratio: σ₁²/σ₂²
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  97,5% CI for
    Ratio Ratio using F
  1,03105 (0,651; 2,101)

Test
Null hypothesis         H₀: σ₁ / σ₂ = 1
Alternative hypothesis  H₁: σ₁ / σ₂ ≠ 1
Significance level      α = 0,025

              Test
Method 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:

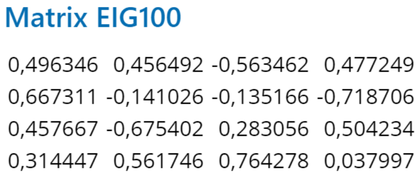
$\lambda 1 = 1{,}92269$

$\lambda 2 = 1{,}04890$

$\lambda 3 = 0{,}81935$

$\lambda 4 = 0{,}20906$

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 $\mathbf{R}$ 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.