QUALITY DATA ANALYSIS

19/06/2023

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.
- Exam duration: 2h
- **For multichance students only: skip exercise 1 point 4) and exercise 3 question 1).**

Exercise 1 (15 points)

In a plant for the production of hydraulic actuators, the head of the quality control department is aiming to implement a new statistical process monitoring tool. The quality characteristic of interest is the holding force measured in kN during acceptance testing. The quality department received holding force measurements that consist of five measurements for each tested actuator. The data for control chart design are stored in "exe1_phase1.csv".

1) Assume that actuators were manufactured and tested with the same sequential order displayed in the dataset, but no information about the time order of individual measurements for each actuator is available. Design an $\bar{X} - R$ control chart such that the average number of samples before a false alarm is 400 and discuss the result.
2) Based on the result of point 1), design a more appropriate control chart for the available data (with the same average number of samples before a false alarm used in point 1). Discuss the results.
3) After some investigations, the data analysts found that the five individual measurements reported for each actuator were carried out always with the same time order, and the order is the one displayed in the dataset. Based on this new information, identify and fit a model for these data and use it to design an appropriate control chart (same average number of samples before a false alarm used in point 1) and 2)). *Note: in case of violations of control limits, assume no assignable cause was found.*
4) Using the control chart designed in point 3) determine if the new samples stored in "exe1_phase2.csv" are in-control or not.

Exercise 2 (14 points)

A manufacturing company that produces electronic components is interested in monitoring the quality of the produced components using statistical techniques. The company has collected data on several variables related to the components, including dimensions, electrical characteristics, and performance indicators. The data are stored in "exe2_phase1.csv".

1) Apply PCA to the data and determine the number of principal components that should be retained to capture at least 80% of the total variance (report the eigenvectors and the eigenvalues of the retained components). Discuss and motivate the choice of using either the variance covariance matrix or the correlation matrix of the data.
2) Based on the results of point 1), design a suitable control charting approach for these data, such that the familywise Type I error is at most 1%. Motivate the choice of the proposed control chart and discuss the result. *Note: in case of violations of the control limits, assume that no assignable cause has been found.*

3) You are now in phase 2 (usage stage of the CC). Test the control chart designed in point 2) on the new observations stored in the file “exe2_phase2.csv” and determine if the process is in-control or not. Report the index of the out-of-control observations (if any).

Exercise 3 (4 points)

In the following questions select one of the four possible choices as your answer and provide a short justification of your choice. Answers **without** justification will **not** receive any credit.

Question 1 (2 points):
In a data set, we fit the simple linear regression model of Y on X, and the Ordinary Least Squares (OLS) line is given by: $\hat{Y} = -0.138 - 1.33 * X$. In the ANOVA table the overall F statistic has the value: $F = 16.81$. Which of the following will be the T-test statistic value (T) that examines the hypothesis testing: $H_0\colon \beta_1 = 0$ versus $H_1\colon \beta_1 \neq 0$? (β_1 refers to the slope of the regression line).

a) $T = -4.1$
b) $T = +4.1$
c) $T = 16.81$
d) We cannot tell from the above output only.

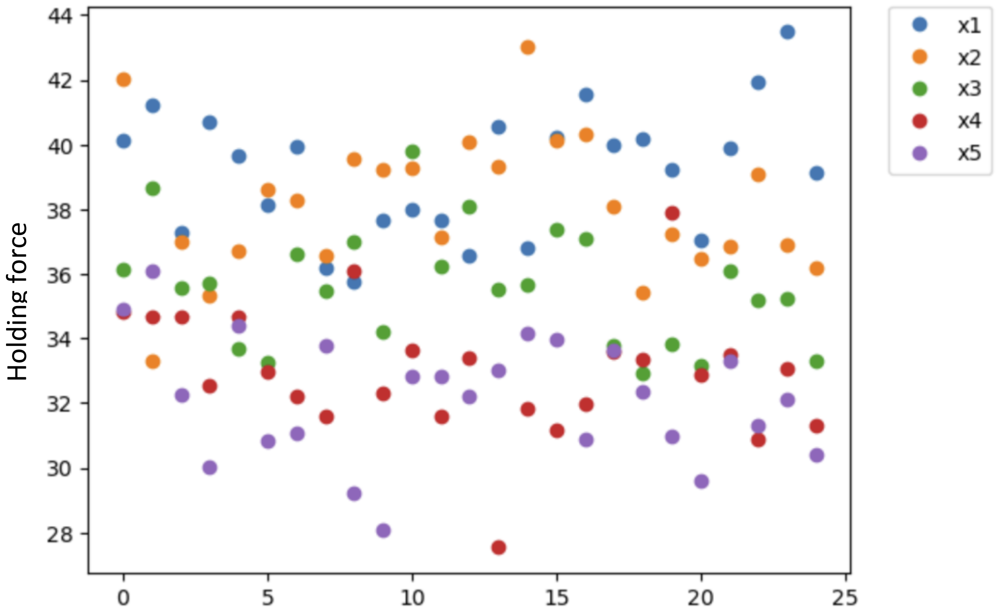
Question 2 (2 points):
In a data set, we fit the simple linear regression model of Y on X, and the Ordinary Least Squares (OLS) line is given by: $\hat{Y} = 1.67 - 2.84 * X$. If the coefficient of determination was $R^2 = 0.81$, then which of the following is true for the (Pearson) correlation coefficient between X and Y, i.e. $r(X, Y)$?

a) $\mathrm{r(X, Y)} = +0.9$
b) $\mathrm{r(X, Y)} = -0.9$
c) $\mathrm{r(X, Y)} = 0$
d) None of the above

Exercise 1 solution

1)

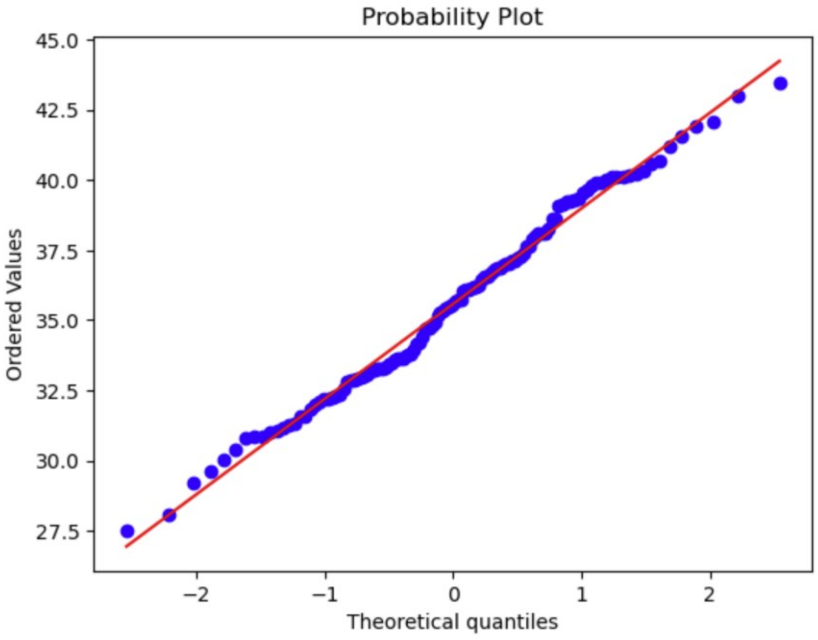
Since the time order of available measurements is known only between samples, and not within samples, we can plot the data as follows, where different colors were used for individual observations in the samples :



Sample

There seems to be no systematic pattern over time (between samples), whereas a systematic pattern may be present within the sample, as moving from observation x1 to x5 a decreasing mean may be present. This is a potential violation of randomness assumption one shall take care of.

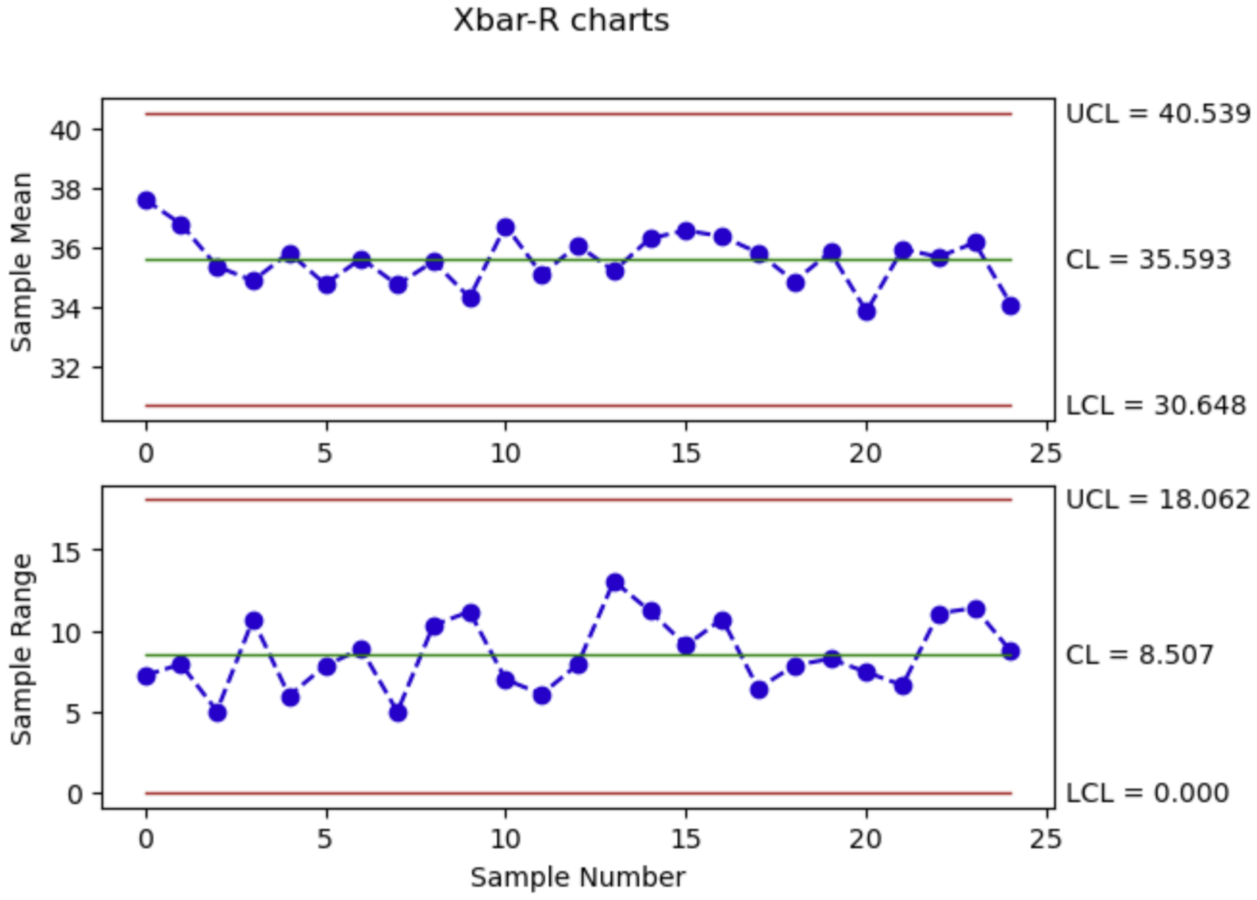
The normality assumption is met (Shapiro-Wilk’s test p-value = 0.351).



Without knowing the time order within the sample it is not possible to make additional tests to check the randomness of the data. Let's assume data to be normal and independent, design the Xbar-R control chart and discuss the results.

Being $ARL_0 = 350$, then $\alpha = \frac{1}{350} = 0.0025$ and K = 3.023.

The resulting control chart with n = 5 is:

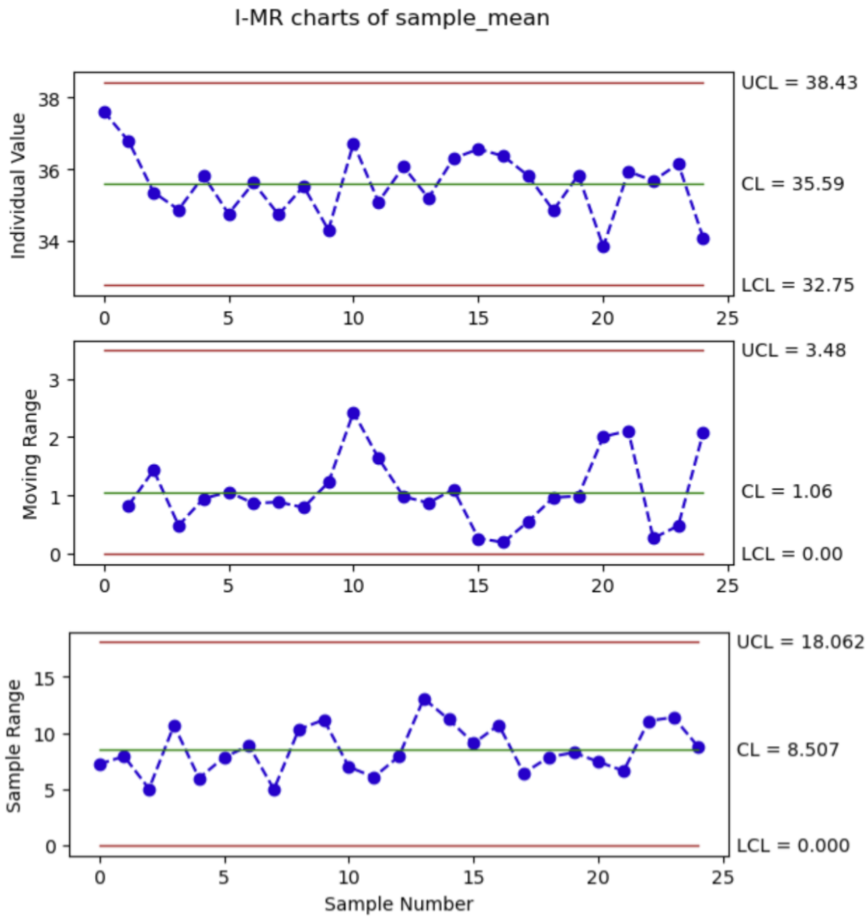


Hugging is evident in the X-bar chart. This may be the consequence of a violation of randomness assumptions *within* the sample. Thus, the X-bar – R control chart is not an appropriate statistical monitoring method for these data.

2)

A more appropriate approach would consist of designing an I-MR-R control chart, to monitor the within and between sample variability.

By using the same ARL_0 used in point 1), the I-MR-R control chart is the following:

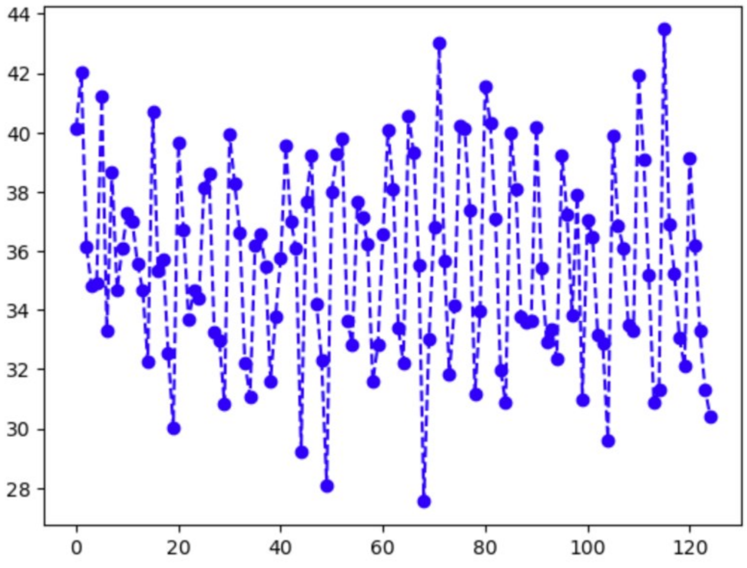


No violation of the control limit is present. The control chart design phase is over.

3)

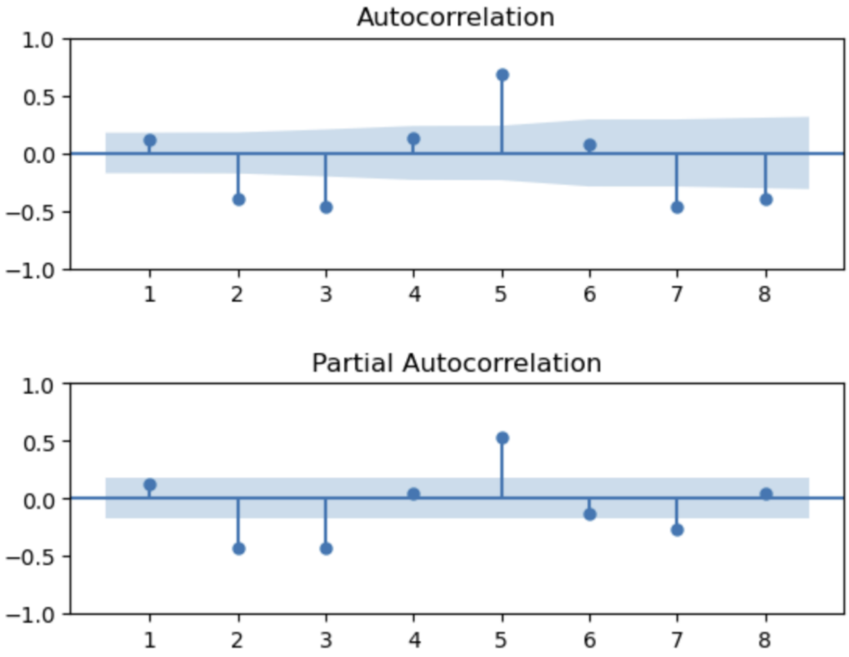
An even more effective control charting scheme would consist of modelling the systematic source of non-random variability within the sample. Being known the time order of individual measurements within the samples, it is possible to make additional analysis and tests.

The time series pattern of the individual measurements is the following:

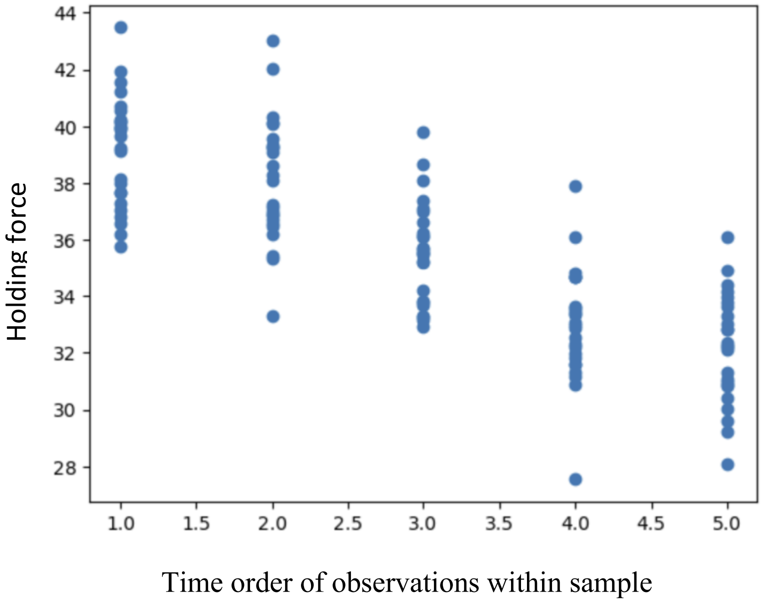


The p-value of the runs-test is p-val = 0.178.

Regardless of the runs-test result, a systematic and non-random pattern in the data is evident, as confirmed by the sample ACF and PACF functions.

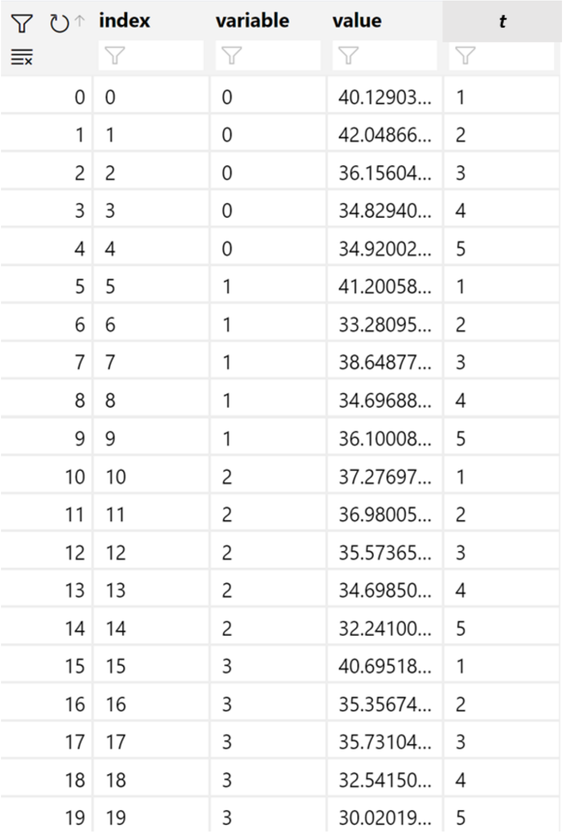


A positive autocorrelation is present at lag 5, corresponding to the sample size n = 5. To further investigate this dependence, it is possible to plot the individual measurements for each actuator with respect to the time-order of within-sample measurements:



The figure shows that there is a decreasing trend of the holding force values moving from the first observation in the sample to the last one.

One way to model this pattern is to use as a regressor the time order within the sample. This can be done by using a “within sample trend” regressor, referred to as *t*, defined as follows:



	index	variable	value	*t*
0	0	0	40.12903...	1
1	1	0	42.04866...	2
2	2	0	36.15604...	3
3	3	0	34.82940...	4
4	4	0	34.92002...	5
5	5	1	41.20058...	1
6	6	1	33.28095...	2
7	7	1	38.64877...	3
8	8	1	34.69688...	4
9	9	1	36.10008...	5
10	10	2	37.27697...	1
11	11	2	36.98005...	2
12	12	2	35.57365...	3
13	13	2	34.69850...	4
14	14	2	32.24100...	5
15	15	3	40.69518...	1
16	16	3	35.35674...	2
17	17	3	35.73104...	3
18	18	3	32.54150...	4
19	19	3	30.02019...	5

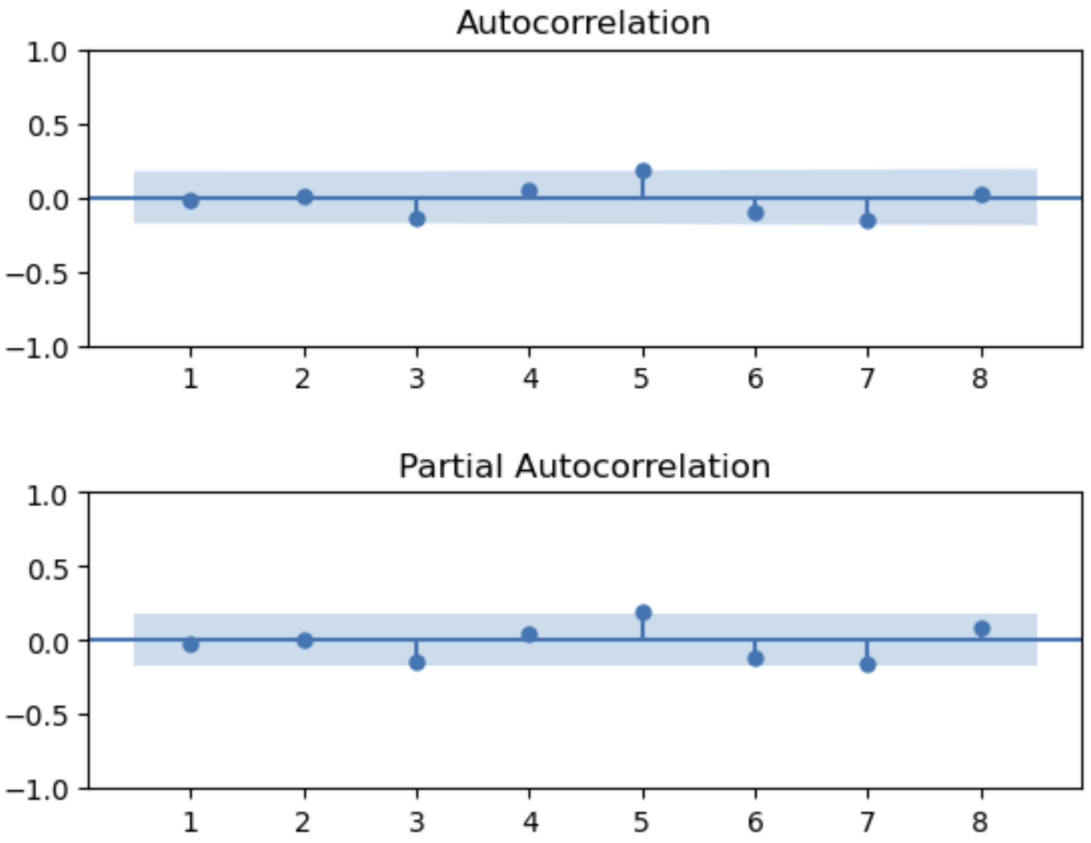
Let's fit a linear model of the holding force against the variable t.

```
REGRESSION EQUATION
-------------------
Holding force =  41.328 - 1.911 t

COEFFICIENTS
------------
 Term     Coef  SE Coef  T-Value     P-Value
const 41.3275   0.4224  97.8480 1.6187e-118
t     -1.9114   0.1273 -15.0094  1.4097e-29

MODEL SUMMARY
-------------
     S   R-sq  R-sq(adj)
2.0135 0.6468      0.644

ANALYSIS OF VARIANCE
---------------------
    Source    DF      Adj SS      Adj MS   F-Value      P-Value
Regression   1.0    913.3796    913.3796  225.2833   1.4097e-29
     const   1.0  38817.3565  38817.3565 9574.2273  1.6187e-118
     t       1.0    913.3796    913.3796  225.2833   1.4097e-29
     Error 123.0    498.6862      4.0544       NaN          NaN
```

```
Total 124.0  1412.0657        NaN        NaN          NaN
```

Let’s check model residuals.

Sample ACF and PACF



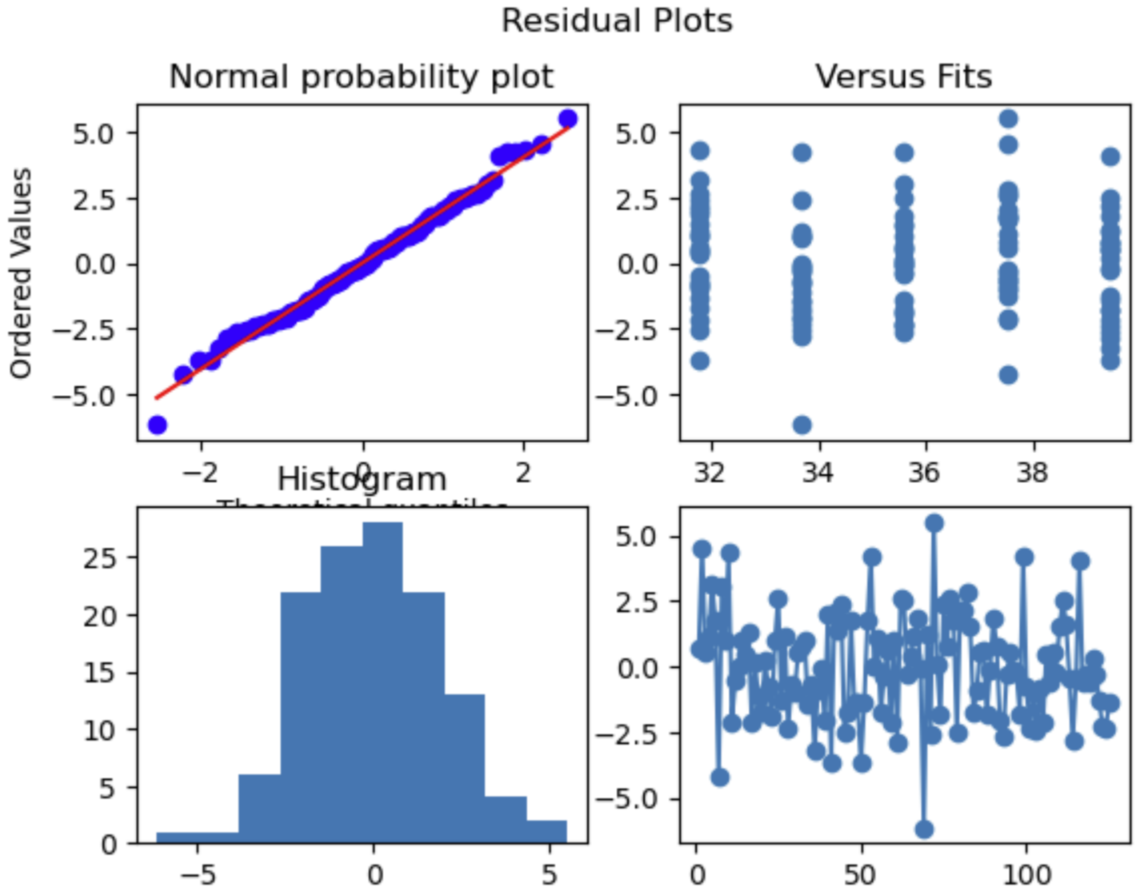
Runs-test of residuals: p-val 0.552

Bartlett test at lag 5 (with alpha = 0.05):

```
Test statistic rk = 0.185666
Rejection region starts at 0.391993
```

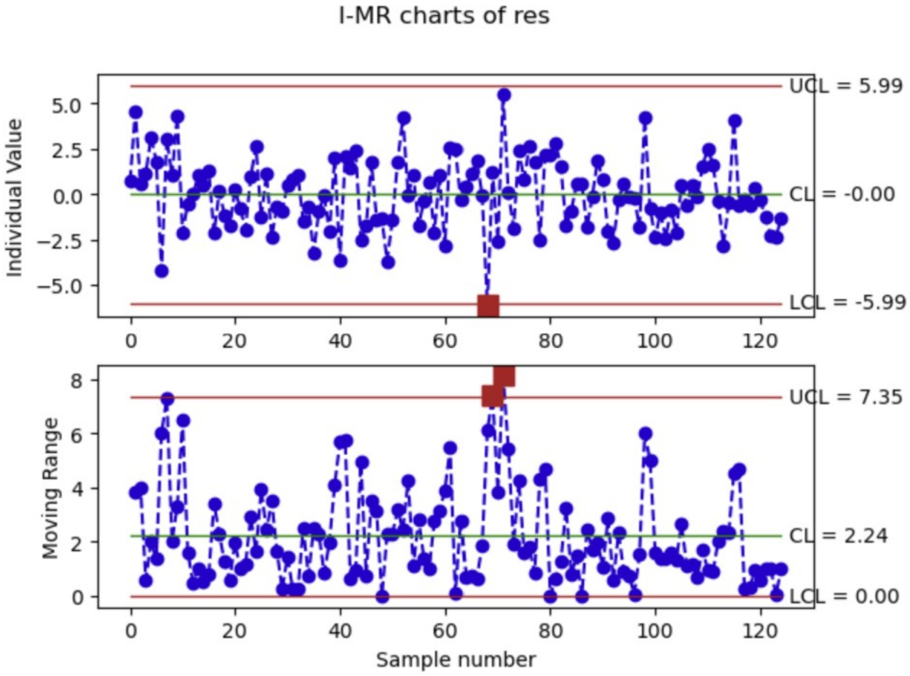
According to the runs-test and Bartlett test, residuals are random.

Shapiro-Wilk’s test on the normality of residuals: p-val = 0.757



The residuals are normal and independent. The model is appropriate. All terms are significant.

The special cause control chart with $ARL_0 = 400$ for the model residuals is the following:



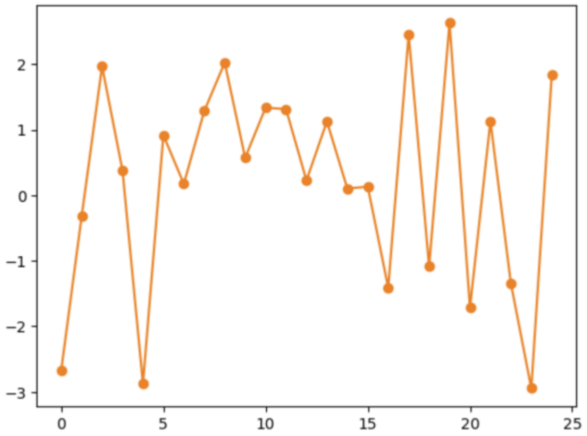
Two violations of the limits are present, but according to the exercise text, no assignable cause is found for them. Therefore, they can be labelled as false alarms. The control chart design is over.

4)

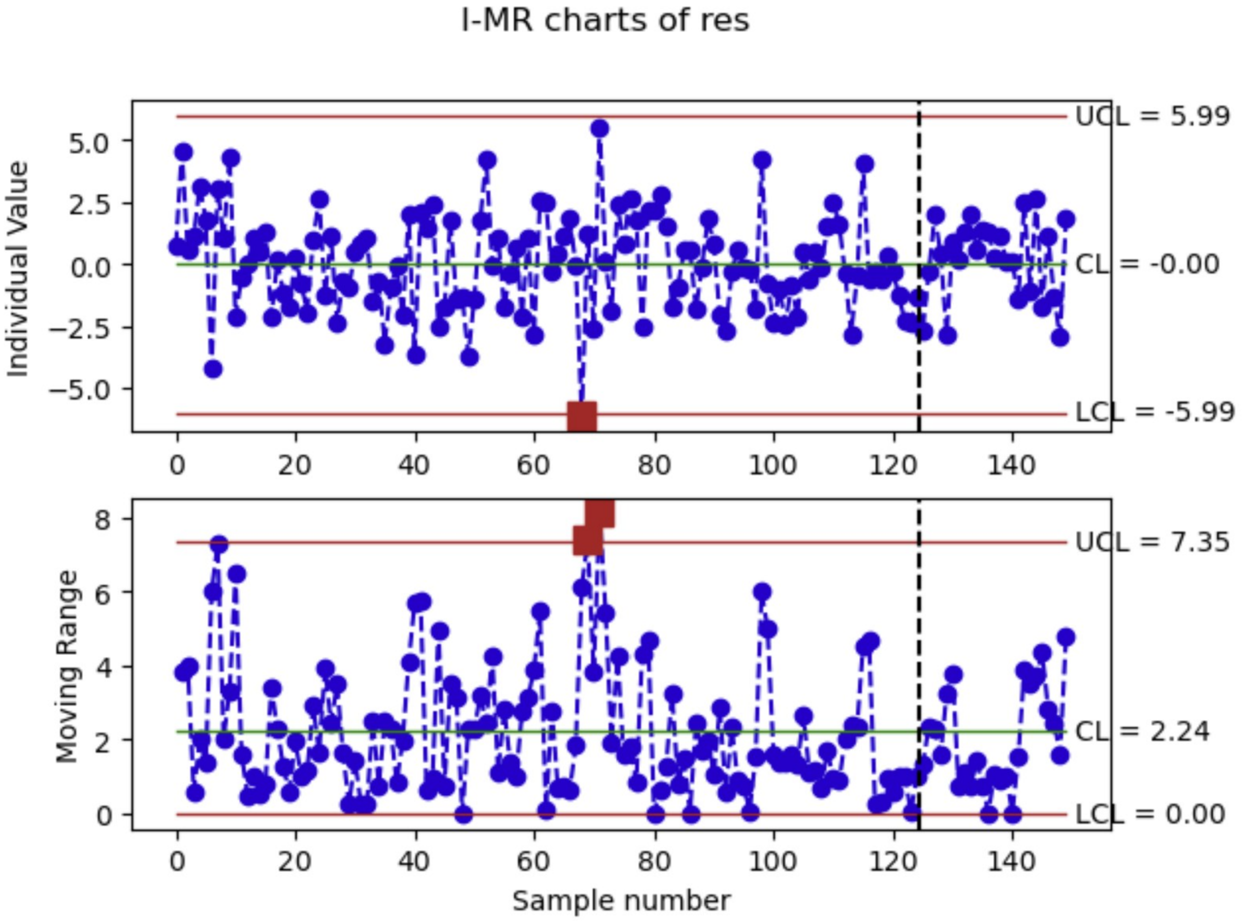
To determine whether new data are in-control or not, the same model used in point 3 shall be fit to them.

```
Holding force =  41.328 - 1.911 t
```

The resulting residuals for the new 5 samples are:



By plotting the new residuals on the previously design special cause control chart, we get:

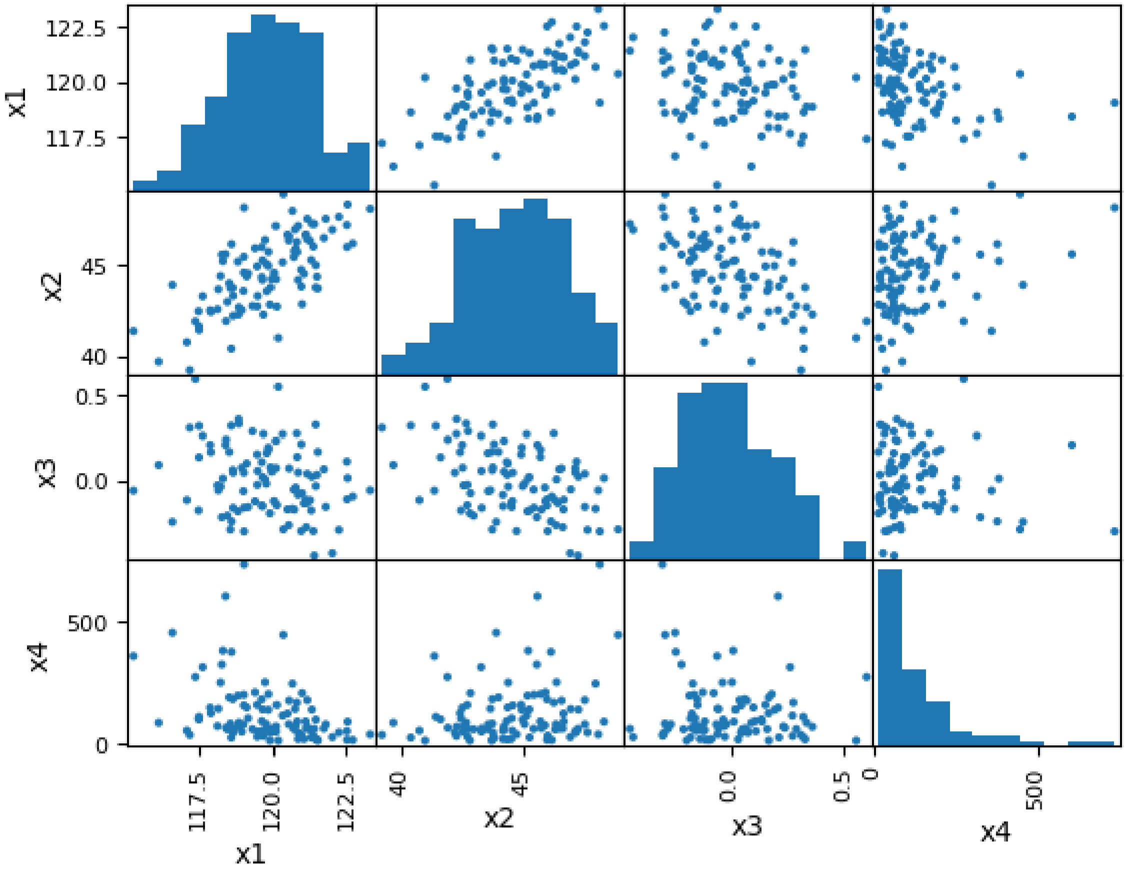


The new observations are in control.

Exercise 2 solutions

1)

Inspect the data and estimate the mean and the variance-covariance matrix.



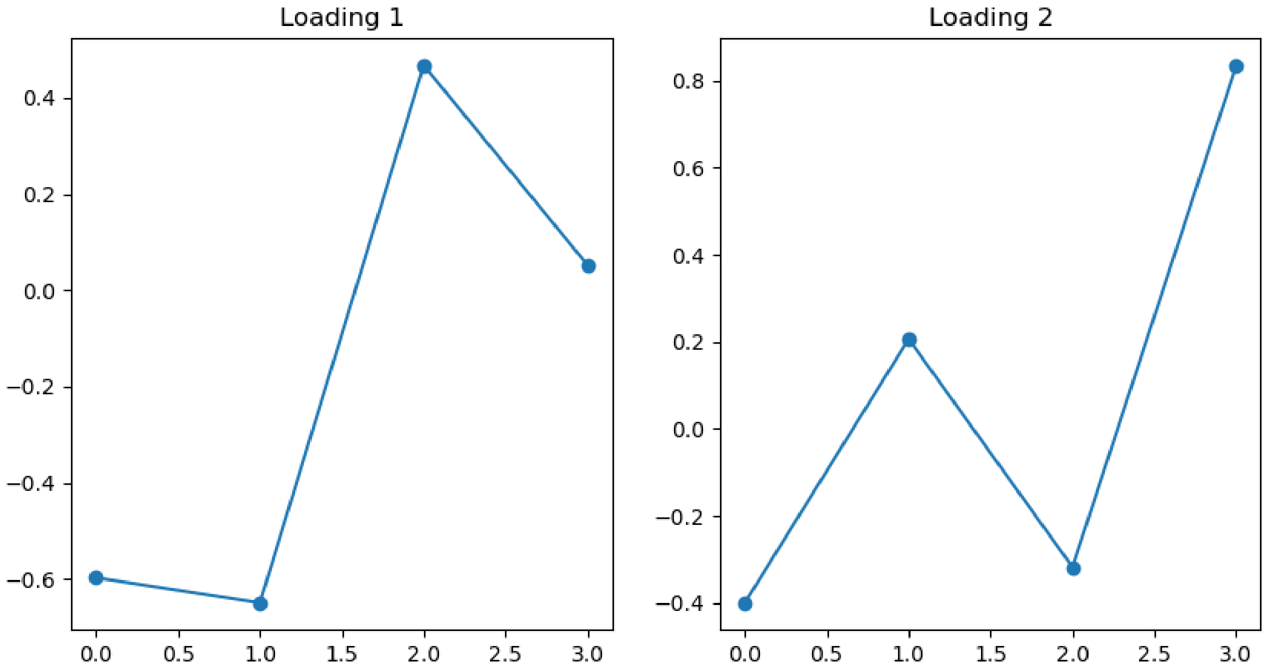
```
Sample Mean:
x1    119.78458
x2     44.62588
x3      0.00111
x4    127.80437
dtype: float64

Sample Variance-Covariance Matrix:
           x1         x2        x3            x4
x1   2.411731   2.201534 -0.074078    -75.914402
x2   2.201534   4.235564 -0.203526     54.559614
x3  -0.074078  -0.203526  0.041152     -2.478084
x4 -75.914402  54.559614 -2.478084  15482.351080
```

Since there is a large difference in the scale and variance of the individual variables, we shall apply PCA to the standardized data (which is equivalent to apply PCA using the correlation matrix of the original data).

```
Explained variance ratio

 [0.49179434 0.3195449  0.16620453 0.02245623]

Cumulative explained variance ratio
 [0.49179434 0.81133924 0.97754377 1]
```

We need to retain the first 2 PCs to capture at least 80% of the total variance.

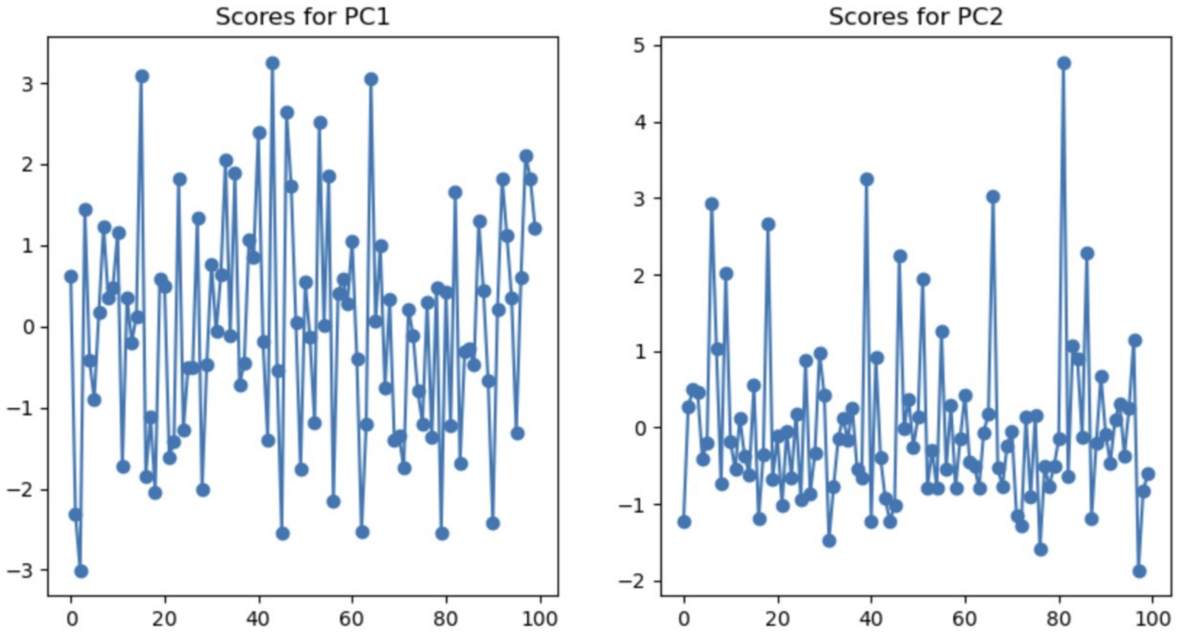
```
Eigenvalues

 [1.96717736 1.2781796 ]

Eigenvectors
 [[-0.59742388 -0.64956733  0.46736347  0.05213804]
 [-0.39983512  0.20609723 -0.31777229  0.83467154]]
```



2)

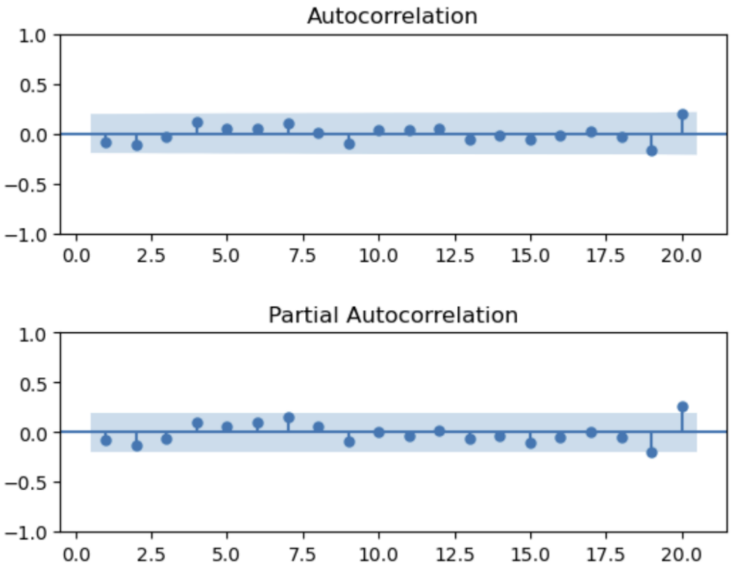
We can design two univariate control charts for the first two PCs (uncorrelated) or one multivariate control chart, but first compute the scores and check the normality and independence assumptions.

Randomness is met as shown below (although the p-value of the runs-test is borderline for PC2):



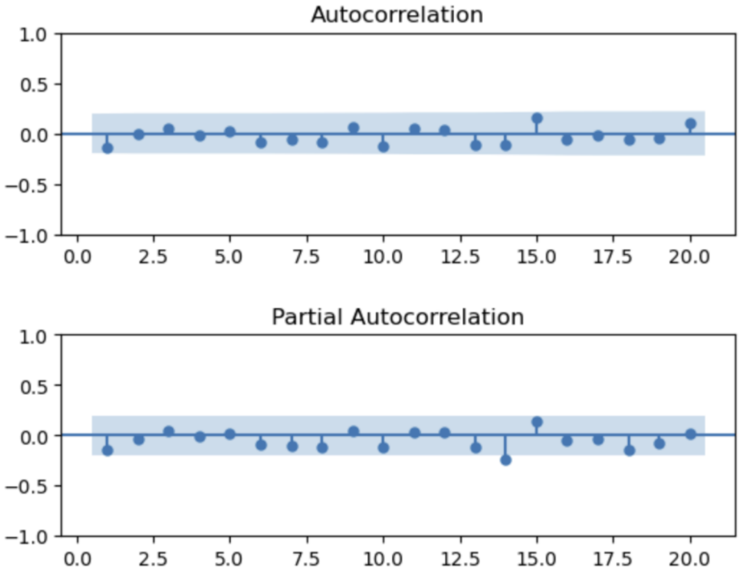
PC1: Runs test p-value = 0.675

PC1: SACF and SPACF:

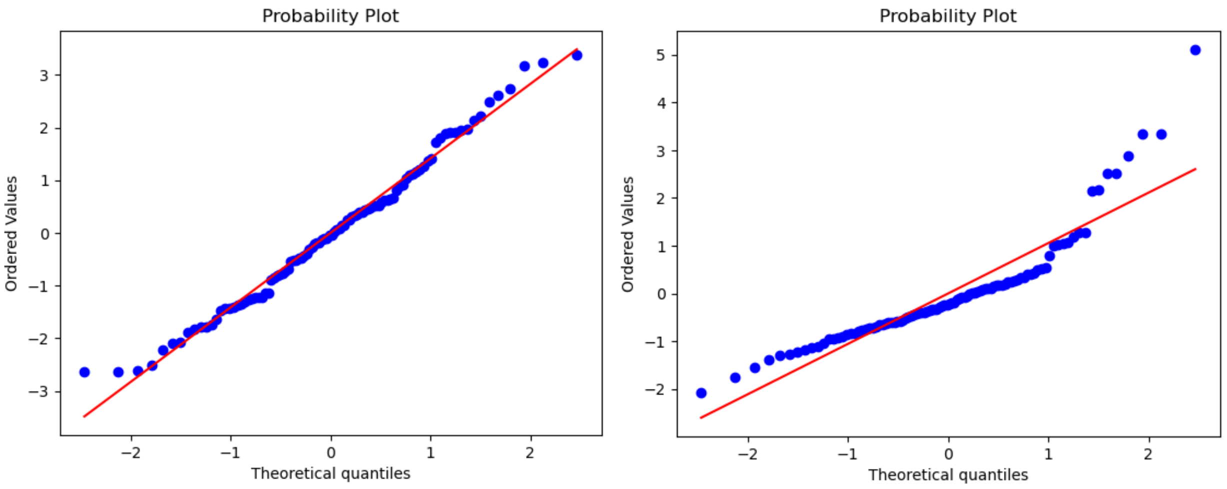


PC1: Runs test p-value = 0.043

PC1: SACF and SPACF:

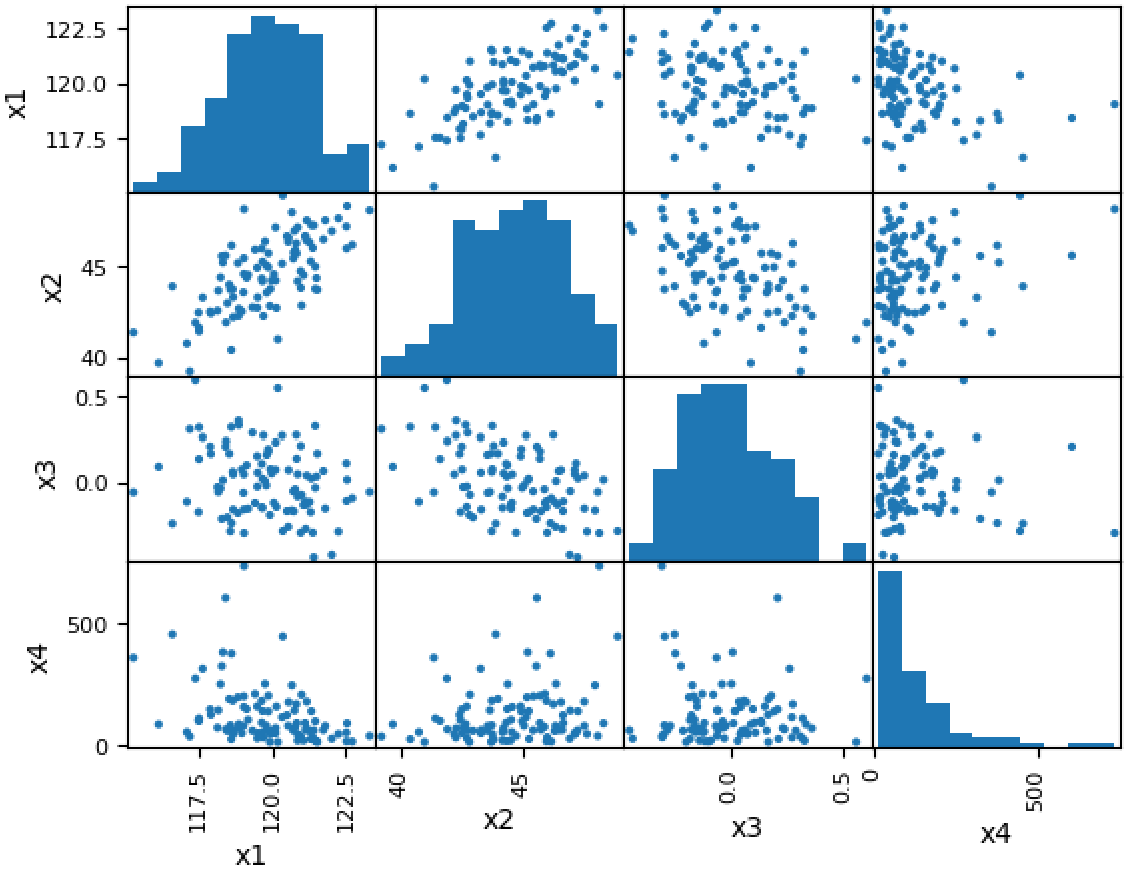


However, normality is violated for PC2:



The p-values of the Shapiro-Wilk's test are p-val = 0.244 and p-val = 0.000 for PC1 and PC2 respectively.

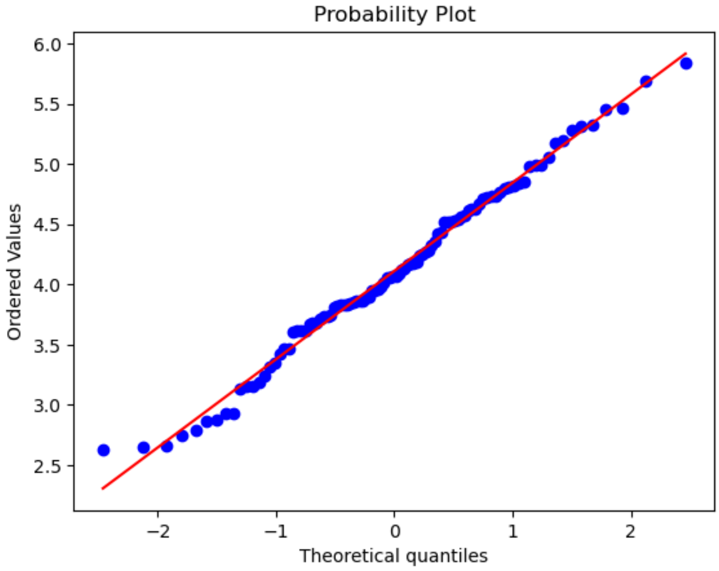
Let's have a look at the original dataset to check if there is some variable that may be responsible for the non-normality of the second PC.



We can see that the distribution of x4 is very skewed, and the weight associated to x4 in the second PC is very high. This is the reason why the second PC is not normal.

The test for normality on x4 gives a p-value = 0.000.

Apply Box-Cox to x4 to try to recover normality. After Box-Cox transformation (lambda ≈ 0.0), this is the distribution of x4 (SW p-value = 0.518).

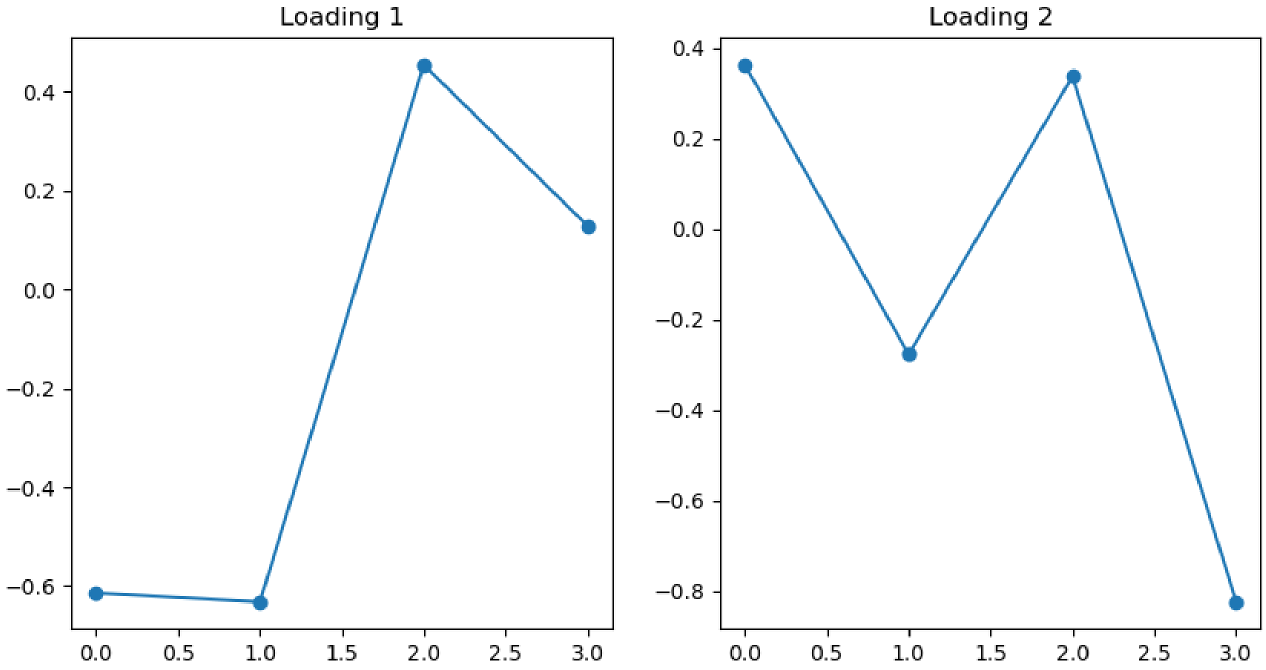


Let’s standardize the new data (with the transformed variable) and re-estimate the PCA.

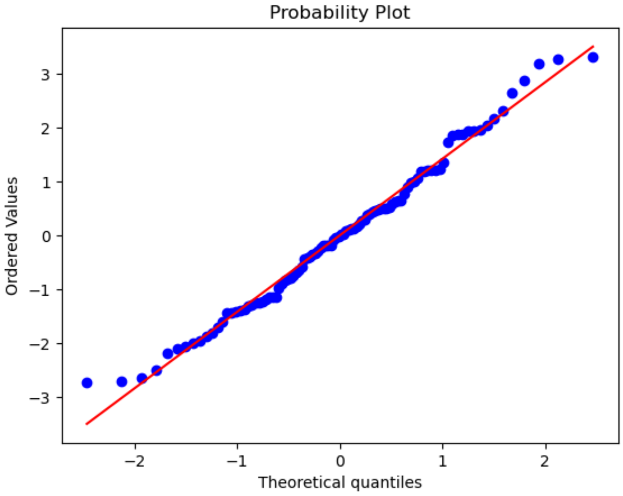
```
Explained variance ratio
 [0.494686   0.32310045 0.16783834 0.01437521]

Cumulative explained variance ratio
 [0.494686   0.81778645 0.98562479 1.        ]
```

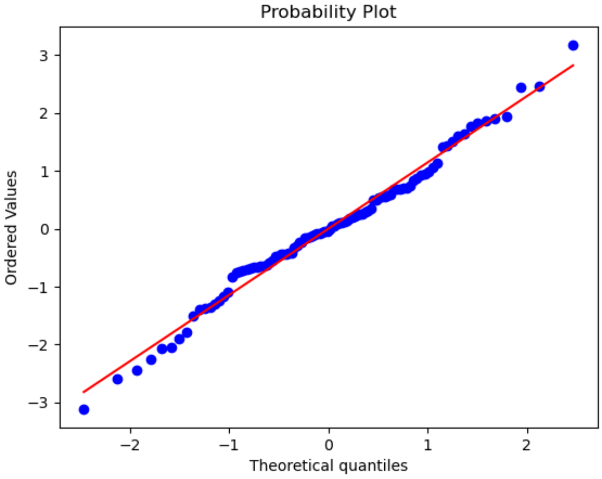
We still need to keep the first 2 PCs to retain at least 80% of the variability.

```
Eigenvalues
 [1.97874401 1.2924018 ]

Eigenvectors
 [[-0.61411239 -0.63211932  0.45467491  0.12869292]
 [ 0.36195886 -0.2773057   0.33655629 -0.82390363]]
```

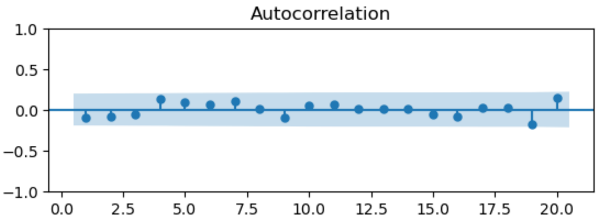


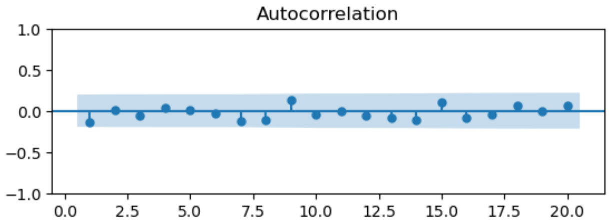
Now check again the normality and the randomness of the new PC1 and PC2.

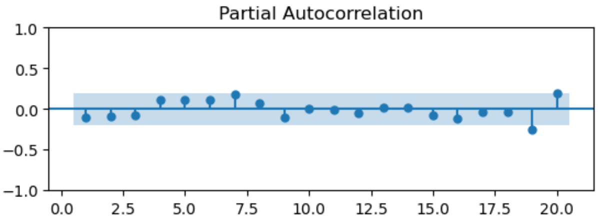


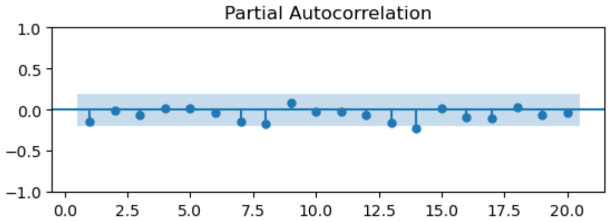


No violation of normality hypothesis (SW p-value 0.298 and 0.543).





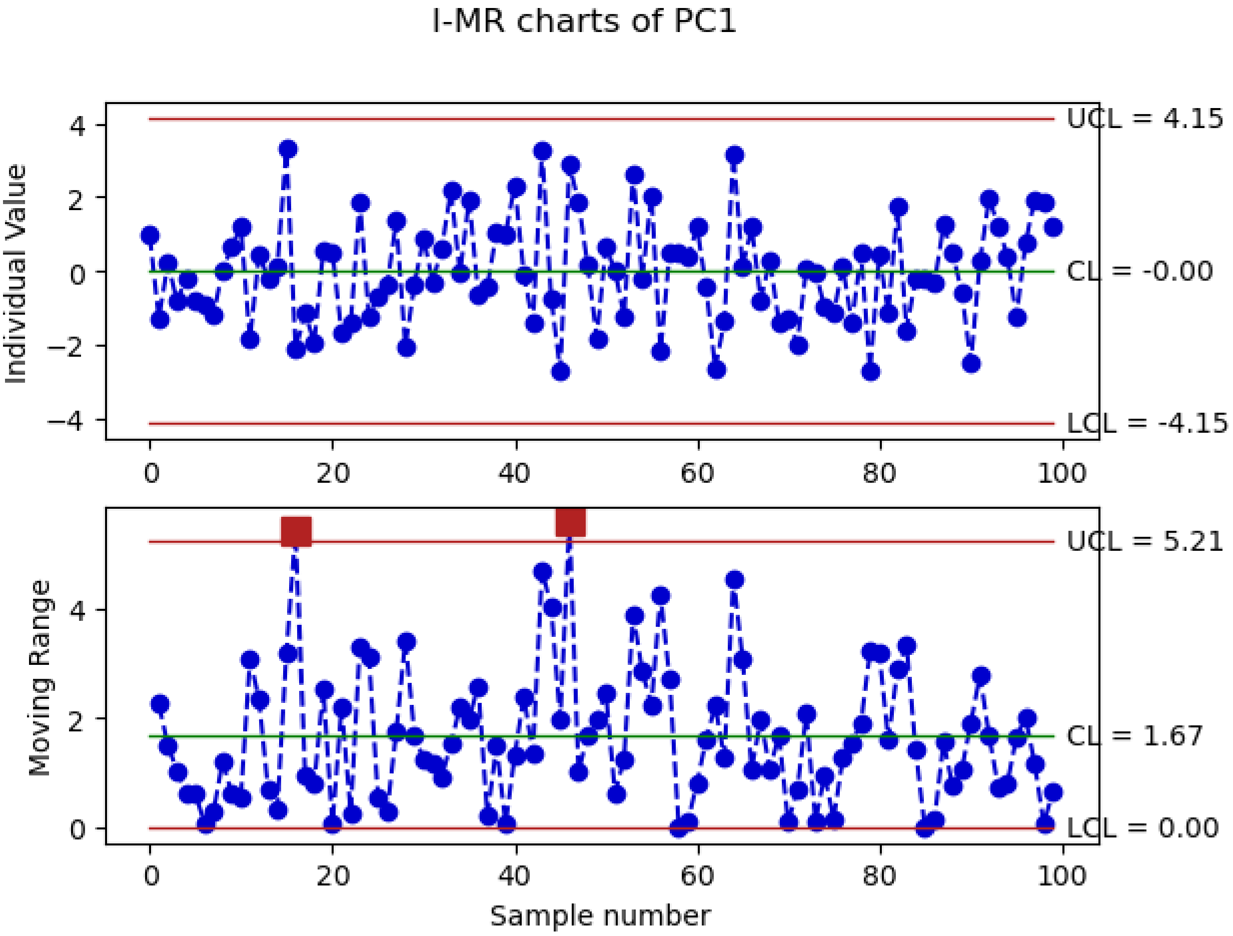


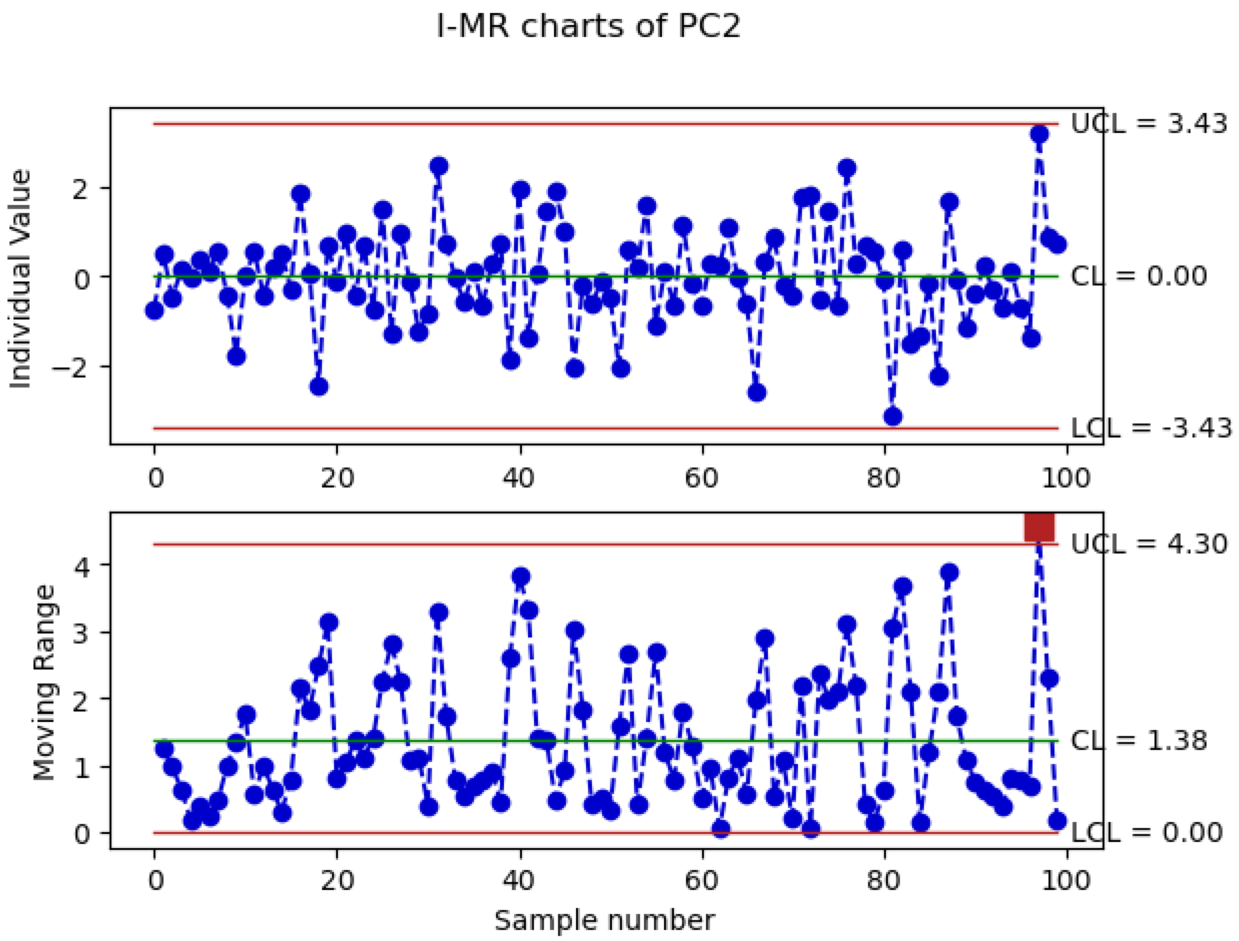


No violation of randomness (runs test p-value 0.421 0.158) or autocorrelation in the time series.

We can now design the two I-MR control charts on the two variables, but first we need to determine the value of K. Being $\alpha_{fam} = 0.01$ the family-wise type I error, and being the two PCs independent, we shall use the following correction: $\alpha = 1 - (1 - \alpha_{fam})^{1/2}$. Thus K = 2.806.

The control charts are the following:


I-MR charts of PC1
Individual Value
4
2
0
−2
−4
UCL = 4.15
CL = -0.00
LCL = -4.15
0
20
40
60
80
100
Moving Range
4
2
0
UCL = 5.21
CL = 1.67
LCL = 0.00
0
20
40
60
80
100
Sample number

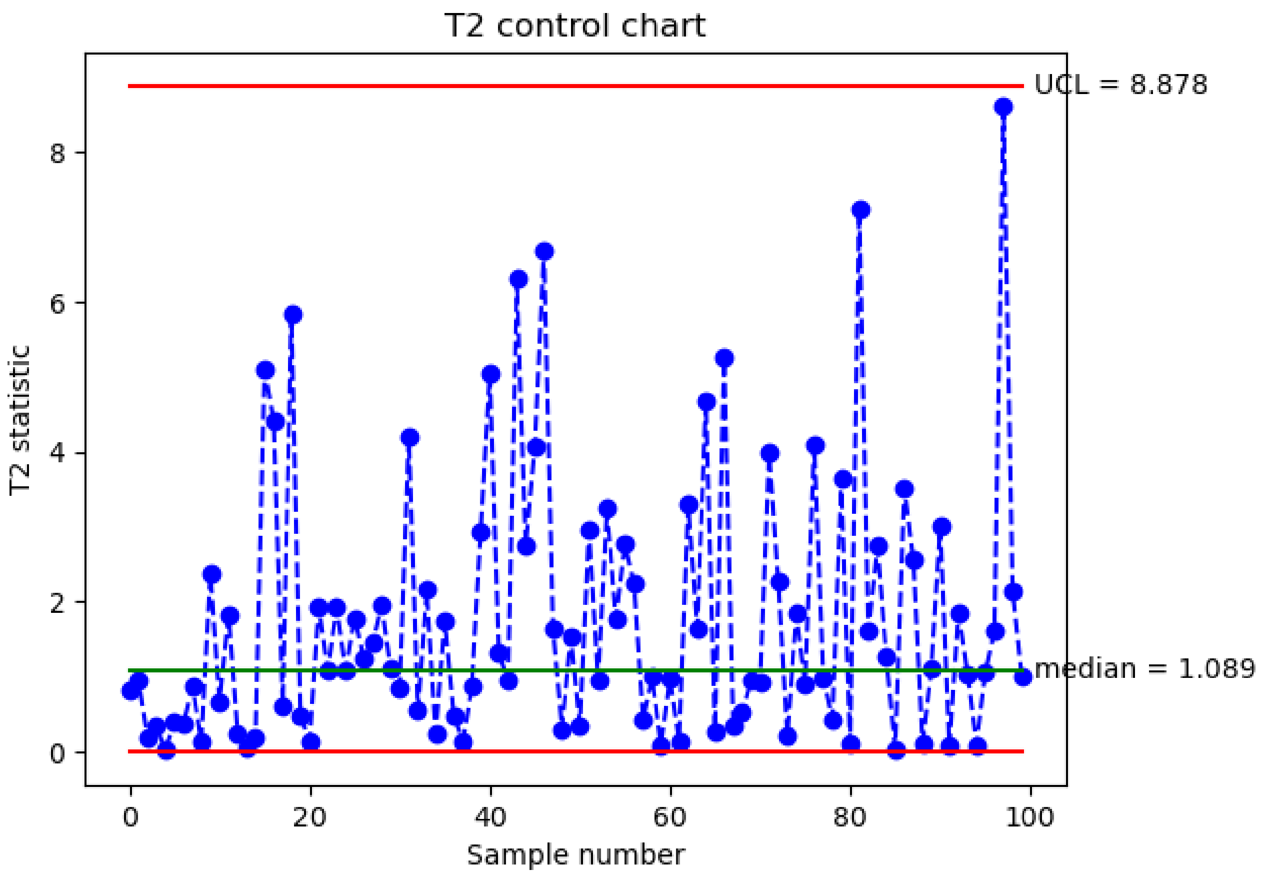


There are violations in the MR charts of both PC1 and PC2. Assuming there is NOT an assignable cause we shall keep them. If the violation is only in the MR chart, we can design the MR charts with prob. limits to check if the violation is due to the non-normality of the MR statistic.

Alternatively to designing 2 I-MR CC, we can design one multivariate (T2) control chart after estimating the mean and the variance-covariance of the scores using the short range estimator S2:

```
The short range estimator is:

           PC1       PC2
PC1  2.161556  0.005111
PC2  0.005111  1.465010
```

In this case, since we are designing only one chart, we don’t need to apply any correction to alpha which is set at 1%.



The T2 control chart does not indicate any OOC.

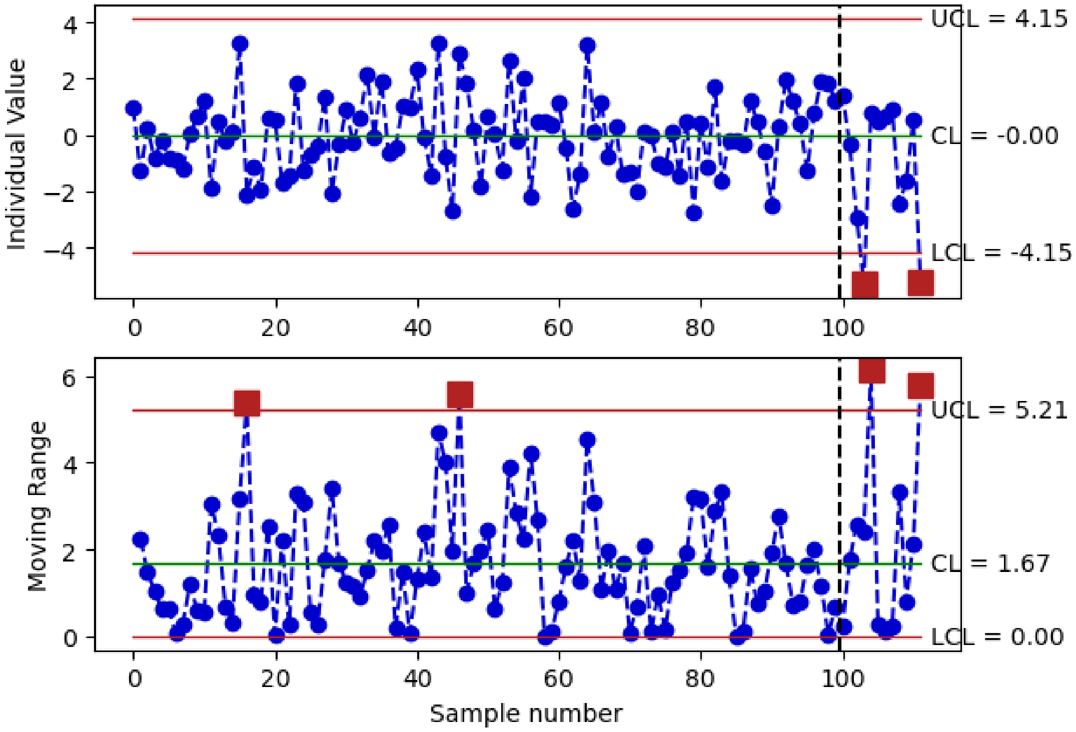
3)

We need to apply the same transformations to the new dataset, i.e.:

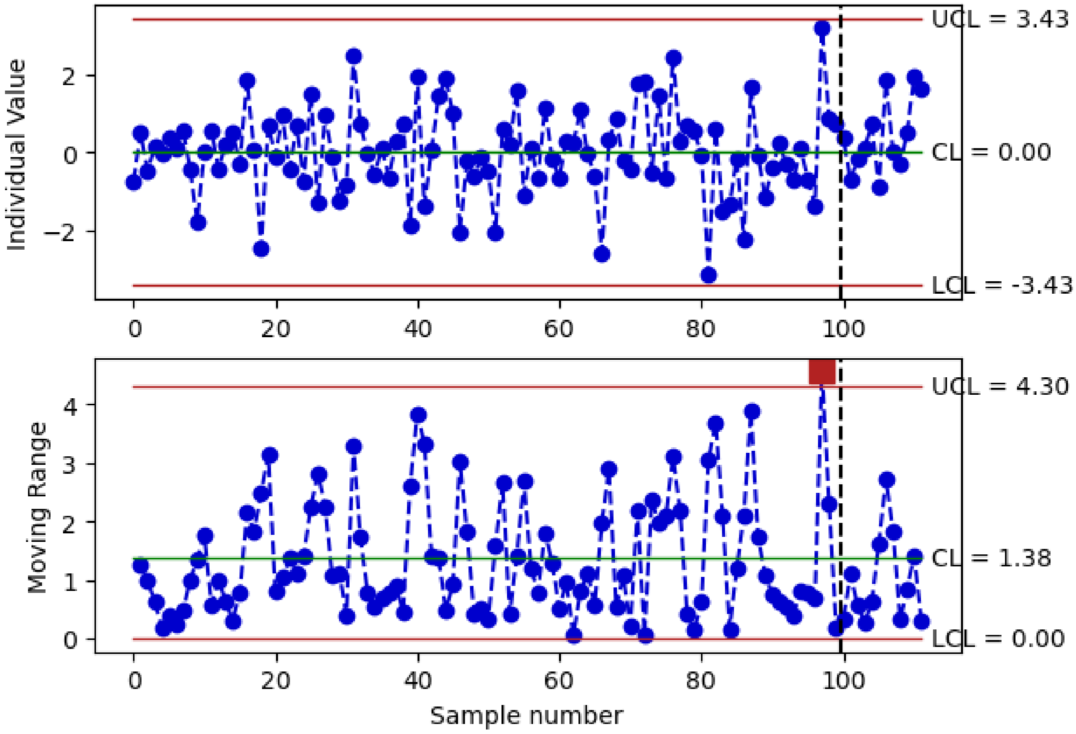
1. Box-Cox on x4 using the previously estimated lambda.
2. Standardize using the previously estimated sample mean and standard deviation.
3. Transform the new data in the PC space spanned by PCs identified in Phase 1.

To test the new observations we can use the two univariate control charts or the T2 cc.

I-MR charts of PC1

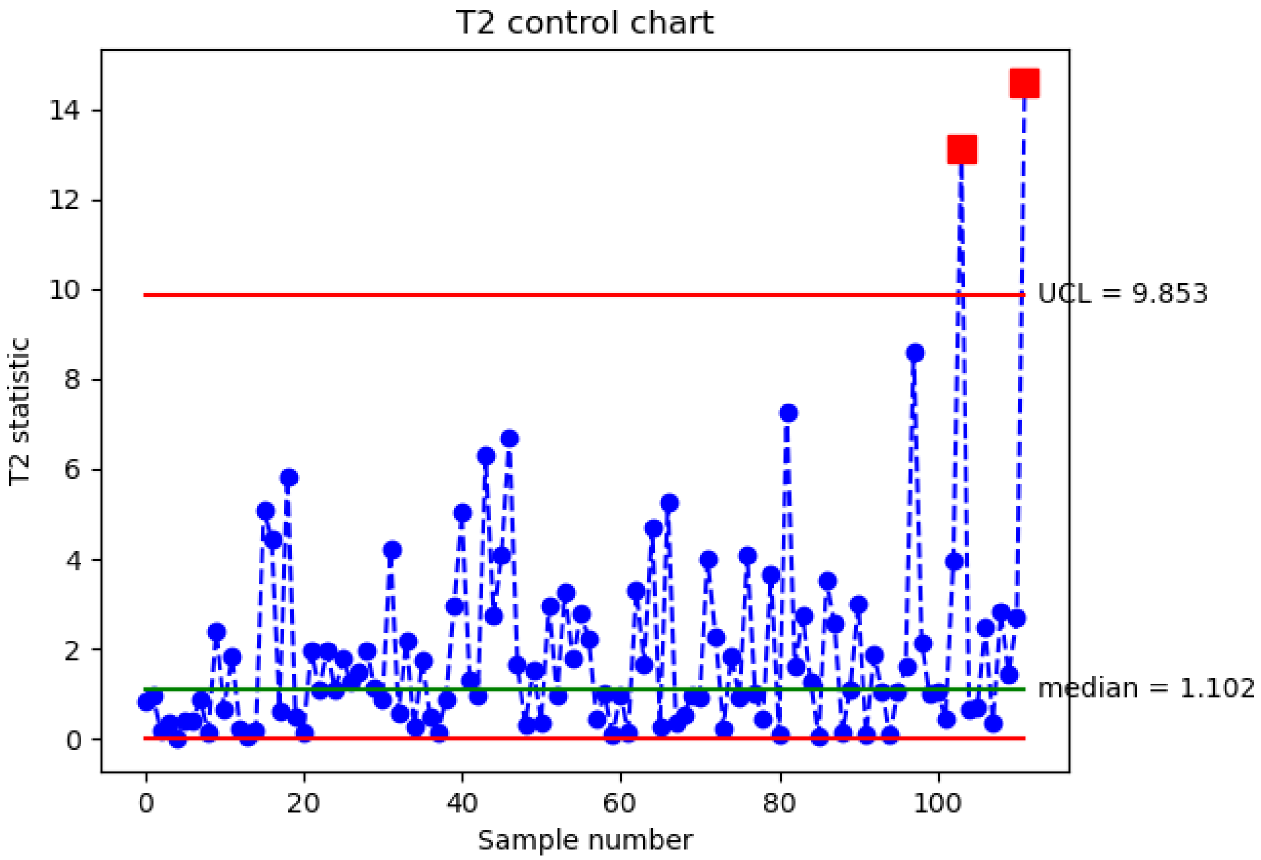


I-MR charts of PC2



```
Out of control points for the I chart on PC1: [103 111]
Out of control points for the MR chart on PC1: [ 16  46 104 111]

Out of control points for the I chart on PC2: []
Out of control points for the MR chart on PC2: [97]
```

Or use the T2 control chart.



```
Out of control points for the T2 chart: [103 111]
```

Exercise 3 solution

Question 1

Answer: a
Explanation: In the simple linear regression for the T-test statistic (T) of the slope, we have that $T^2 = F$ (where F is the overall F-test). Thus $T = \pm\sqrt{F} = \pm\sqrt{16.81} = \pm 4.1$
For the T-test we also have:

$$T = \frac{\hat{\beta}_1 - 0}{s.e.\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1}{s.e.\left(\hat{\beta}_1\right)}$$

and from the fitted model $\hat{\beta}_1 = -1.33$, therefore the T-test statistic will have a negative sign and so we have $T = -4.1$.

Question 2

Answer: b
Explanation: In the simple linear regression we have that $R^2 = \left(r(X,Y)\right)^2$. Thus $r(X,Y) = \pm\sqrt{R^2} = \pm\sqrt{0.81} = \pm 0.9$.
Furthermore, we know that:

$$r(X,Y) = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \quad \text{and} \quad b_1 = \frac{S_{XY}}{S_{XX}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}\sqrt{\frac{S_{YY}}{S_{XX}}} = r(X,Y)\sqrt{\frac{S_{YY}}{S_{XX}}}$$

thus, the estimated slope and the correlation coefficient will have the same sign. Since in this problem $b_1 = -2.84$, the correlation will be negative and therefore: $r(X,Y) = -0.9$ is the correct answer.