# QUALITY DATA ANALYSIS

**09/07/2021**

## 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

### Exercise 1 (14 points)

In a plant for the production of submarine valves, the hardness of input material provided by a given vendor is measured and monitored over time. Vickers hardness measurements are carried out periodically following a predefined procedure that involves measurements in five different locations. The data collected in ten consecutive samples are shown in Table1.

Table 1

| Sample | Measurements (VH) | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 99,94 | 104,84 | 105,55 | 104,27 | 103,41 |
| 2 | 100,28 | 104,12 | 104,07 | 103,15 | 100,05 |
| 3 | 100,48 | 100,94 | 106,19 | 105,25 | 100,58 |
| 4 | 102,66 | 103,41 | 105,49 | 102,93 | 97,16 |
| 5 | 97,91 | 105,68 | 110,62 | 102,8 | 100,95 |
| 6 | 101,93 | 105,67 | 106,38 | 101,89 | 101,11 |
| 7 | 99,56 | 100,43 | 105,65 | 101,77 | 98,02 |
| 8 | 101,83 | 105,24 | 105,97 | 105,21 | 100,9 |
| 9 | 101,15 | 104,32 | 105,77 | 104,54 | 100,7 |
| 10 | 98,81 | 103,45 | 110,14 | 103,17 | 98,26 |

1) Design an $\bar{X} - R$ control chart such that the average number of samples before a false alarm is 250 and discuss the result.
2) Based on the result of point 1), design a more appropriate control chart for the available data. Discuss the results.
3) Assuming that the five measurements within each sample are performed following the same order shown in Table 1, identify and fit a model for these data and use it to design an appropriate control chart. Discuss the results.

### Exercise 2 (14 points)

In a chemical process, the release of zinc oxides is measured over time (one measure every minute). Table 2 shows the measured values (in $\mathrm{mg/m^3}$) during 20 minutes of reaction.

Table 2

| t (min) | X $(\mathrm{mg/m^3})$ |
|---|---|
| 1 | 148,54 |
| 2 | 154,9 |
| 3 | 153,09 |
| 4 | 158,05 |
| 5 | 151,74 |
| 6 | 156,72 |
| 7 | 153,86 |
| 8 | 159,86 |
| 9 | 162,92 |
| 10 | 163,07 |
| 11 | 154,65 |
| 12 | 164,05 |
| 13 | 163,03 |
| 14 | 165 |
| 15 | 163,36 |
| 16 | 161,05 |
| 17 | 169,68 |
| 18 | 166,48 |
| 19 | 168,01 |
| 20 | 170,51 |

1) Design a trend control chart, i.e., a chart where the centre line is set at the fitted trend and the limits are set accordingly, for data in Table 2 with $ARL_0 = 250$, being known that an increasing trend is a natural signature of the process.
2) Estimate the operating characteristic curve for the trend control chart when the intercept (i.e., the $\beta_0$ term) exhibits a shift, $\delta_0$, expressed in units of residuals' standard deviation, from the in-control condition. Show the values of the Type II error for $\delta_0 = 2$ and $\delta_0 = 4$ standard deviation units.
3) Estimate the operating characteristic curve for the trend control chart when the slope (i.e., the $\beta_1$ term) exhibits a shift, $\delta_1$, expressed in units of residuals' standard deviation, from the in-control condition: estimate the curve at time t = 5 min and t = 10 min. For both the curves show the values of the Type II error for $\delta_1 = 0{,}25$ and $\delta_1 = 0{,}5$ standard deviation units.

**Exercise 3 (5 points)**

A quality characteristic is measured by sampling products from two different production lines. The two lines are independent and the quality characteristic from both of them is known to be normal with means $\mu_1$ (line 1) and $\mu_2$ (line 2), and variances $\sigma_1^2$ and $\sigma_2^2$. The head of the quality control department is interested in keeping under control the process by monitoring the ratio between the variances of the quality characteristic measured in the two lines. Being known that under in-control conditions, $\sigma_1^2 = 3{,}5$ and $\sigma_2^2 = 2{,}5$:

1) Estimate the probabilistic control limits for a control chart on the ratio between $\sigma_1^2$ and $\sigma_2^2$ (assume $\alpha = 0.01$ and the same sample size, $n = 8$, from both production lines).
2) The head of the quality department is interested in changing the sample size in order to tune the performances of the control chart. To this aim, determine the minimum sample size that allows signalling, with a probability at least equal to 70%, that the ratio between the variances is four times the one under in-control conditions.

---

**Multichance students can skip:**

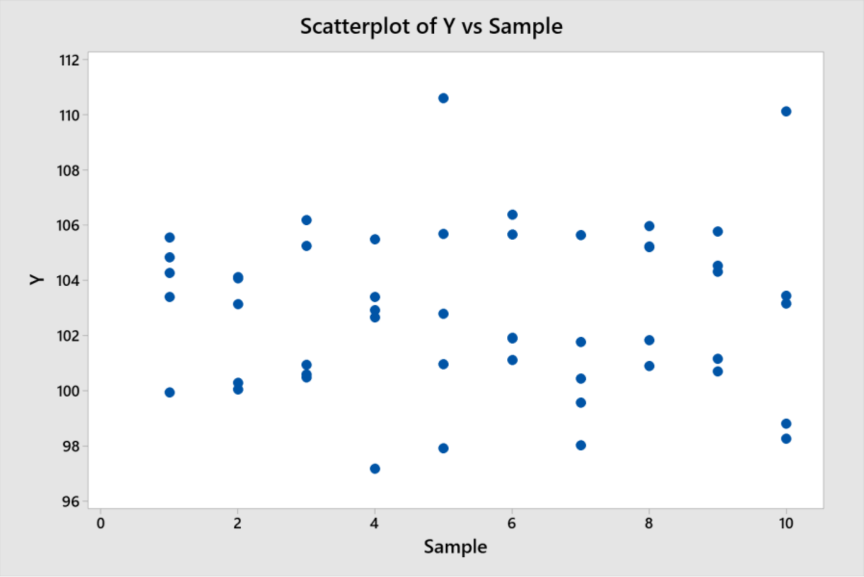
- Exercise 1 : point 2)
- Exercise 2 : point 2)
- Exercise 3 : point 1)

**Solutions**

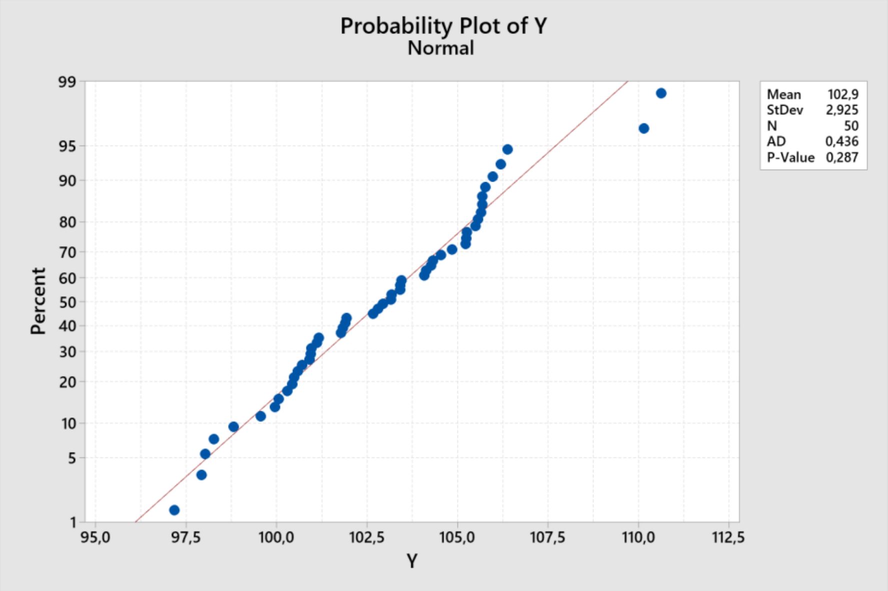
**Exercise 1**

1)

Data snooping :

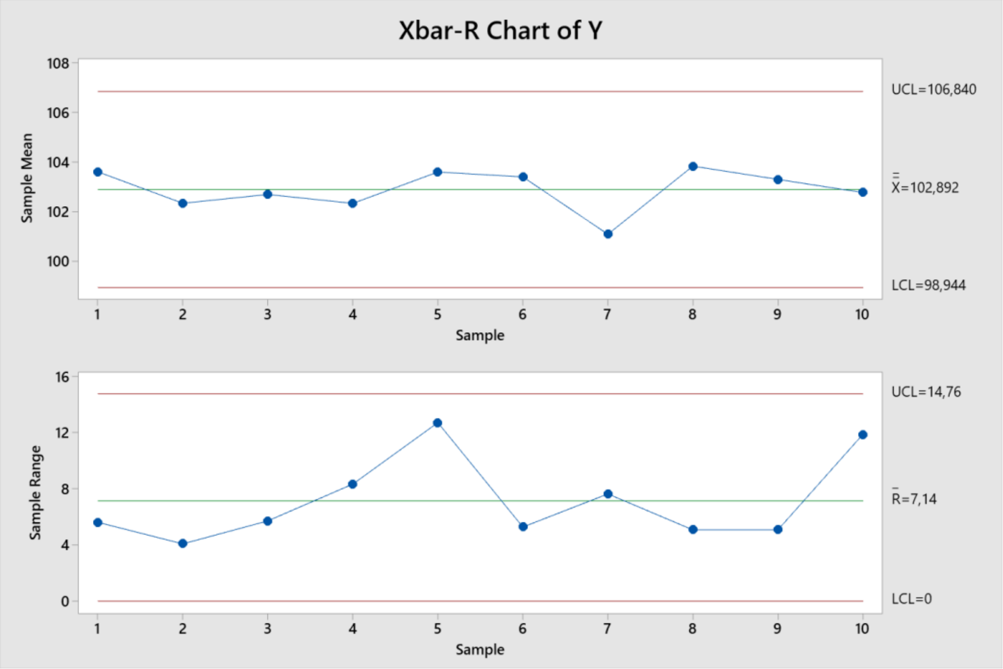


No evident pattern is present in the data. Since the time order within the sample is not known it is not possible to perform tests to investigate the randomness of the data. Based on the qualitative analysis of the plot above we can assume the assumption on randomness is met. The normality assumption is met:



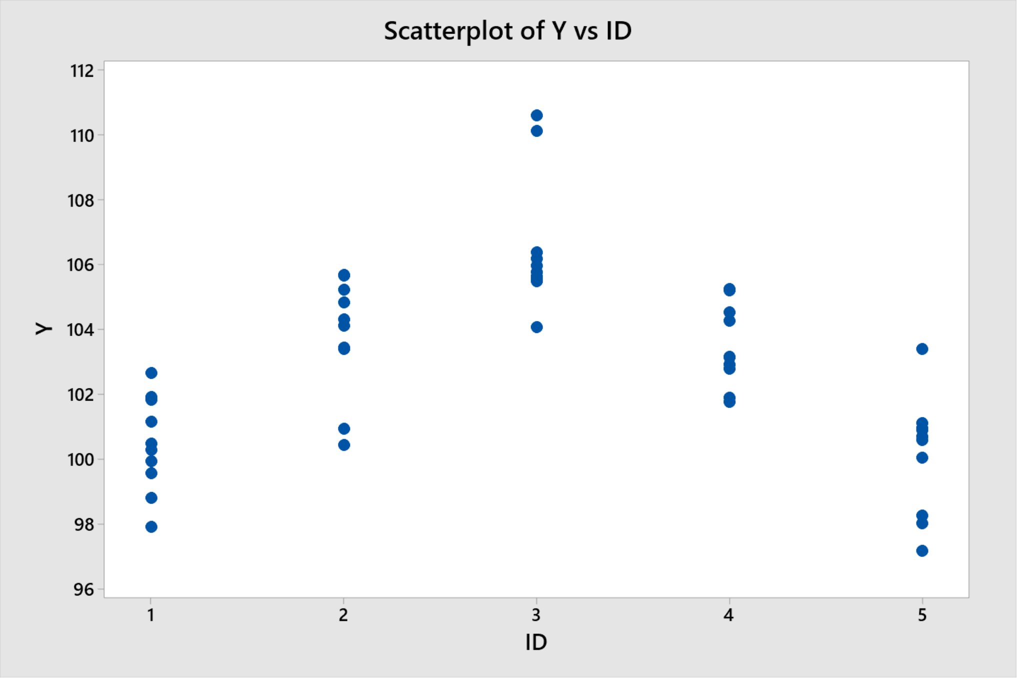
The control chart shall be designed with $ARL_0 = \frac{1}{\alpha} = 250$, and hence: $K = z_{\alpha/2} = 2{,}878$.

The resulting $\bar{X} - R$ control chart is the following:



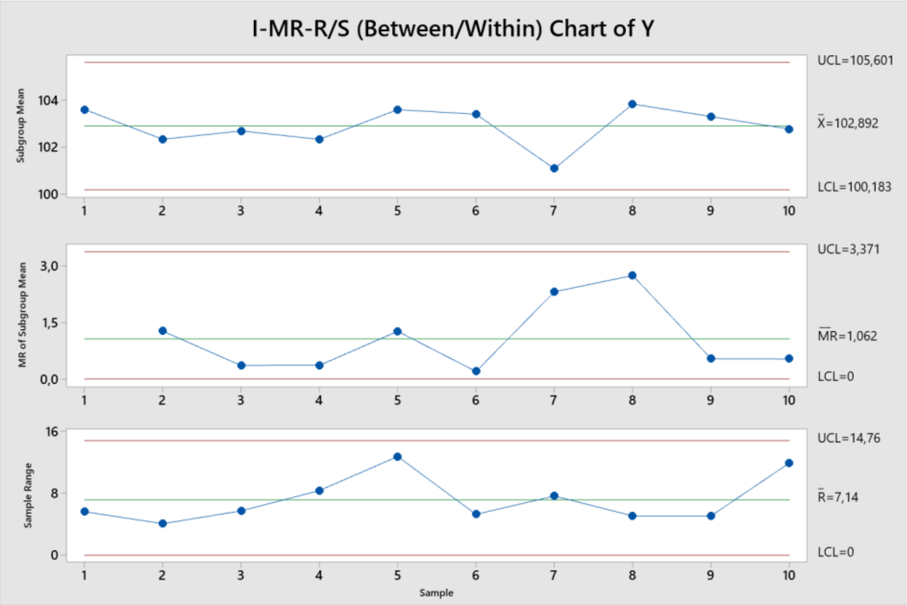
In the control chart for the mean, a hugging effect is evident. It can be caused by a violation of the randomness assumption within the sample, which in turns inflates the data variability.

The following scatter plot shows the hardness values as a function of the locations where they were measured. It is evident that there is a systematic pattern that is responsible for the hugging effects shown in the $\bar{X} - R$ control chart.



2)

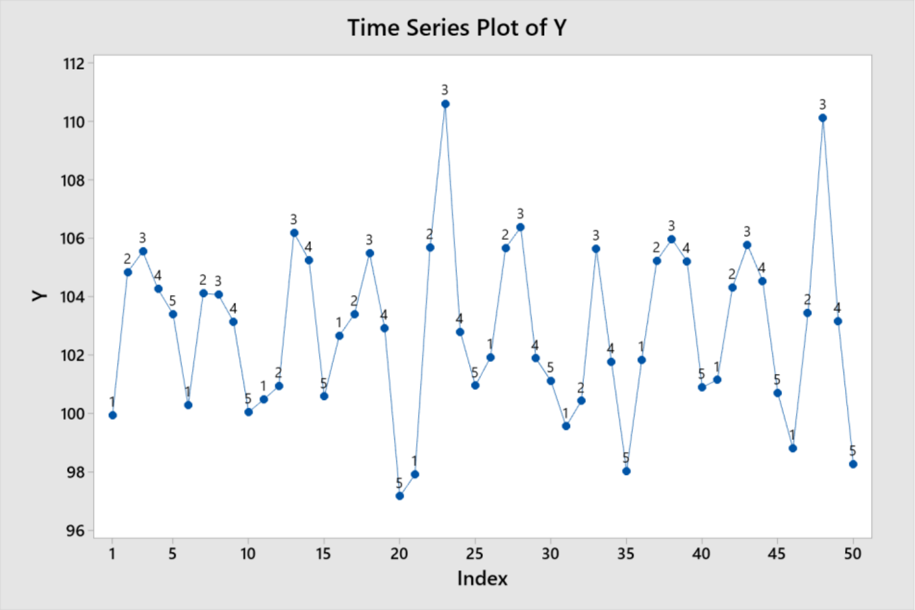
A more appropriate control chart in the presence of such hugging patterns is an I-MR-R control chart. The control chart designed with $K = z_{\alpha/2} = 2{,}878$ is the following:



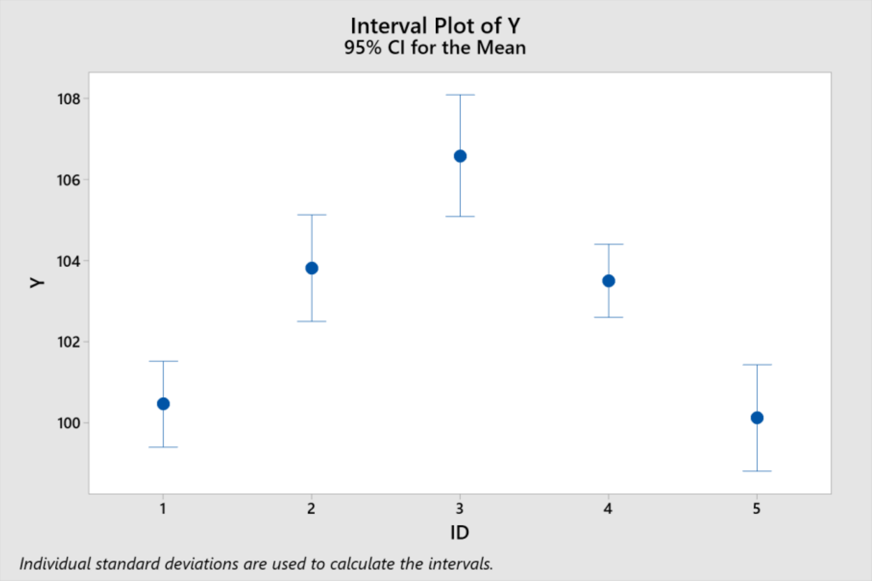
The process is in-control.

3)

Being known the time order of hardness measurements within the sample, it is possible to analyze a time series plot (where IDs of measurement locations are reported too):



It is possible to fit a time series model by using a dummy variable representing the IDs of the five different locations where hardness measurements are taken. Since some locations are not statistically different from each other, not all dummy terms are significant (as also shown by the 95% confidence intervals shown below). One possible solution is to use the stepwise regression to keep only significant dummy terms. Alternatively, the dummy can be defined assigning the same value to locations 1 and 5 (e.g., value 1) and the same value to locations 2 and 4 (e.g., value 2). In both cases the resulting model will consists of three significant dummy terms as shown below.



Interval Plot of Y
95% CI for the Mean

*Individual standard deviations are used to calculate the intervals.*

WORKSHEET 1

## Regression Analysis: Y versus ID1

### Method

Categorical predictor coding (1; 0)

### Regression Equation

Y = 100,284 + 0,0 ID1_1 + 3,370 ID1_2 + 6,299 ID1_3

### Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 100,284 | 0,379 | 264,30 | 0,000 | |
| ID1 | | | | | |
| 2 | 3,370 | 0,537 | 6,28 | 0,000 | 1,20 |
| 3 | 6,299 | 0,657 | 9,58 | 0,000 | 1,20 |

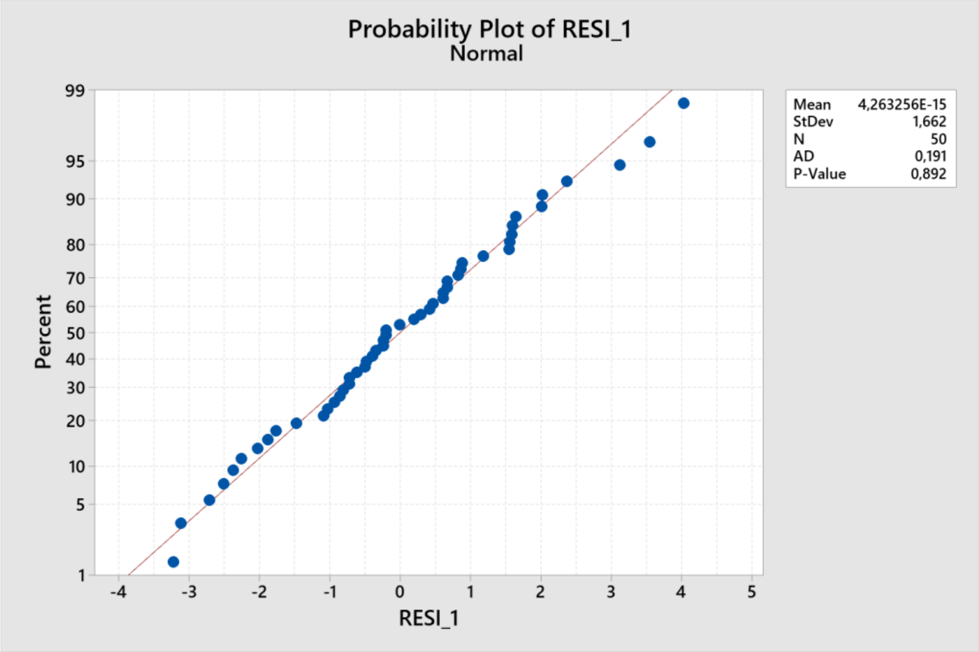
### Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 1,69691 | 67,71% | 66,34% | 63,03% |

### Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 2 | 283,8 | 141,914 | 49,28 | 0,000 |
| ID1 | 2 | 283,8 | 141,914 | 49,28 | 0,000 |
| Error | 47 | 135,3 | 2,879 | | |
| Total | 49 | 419,2 | | | |

The residuals are normal and independent:

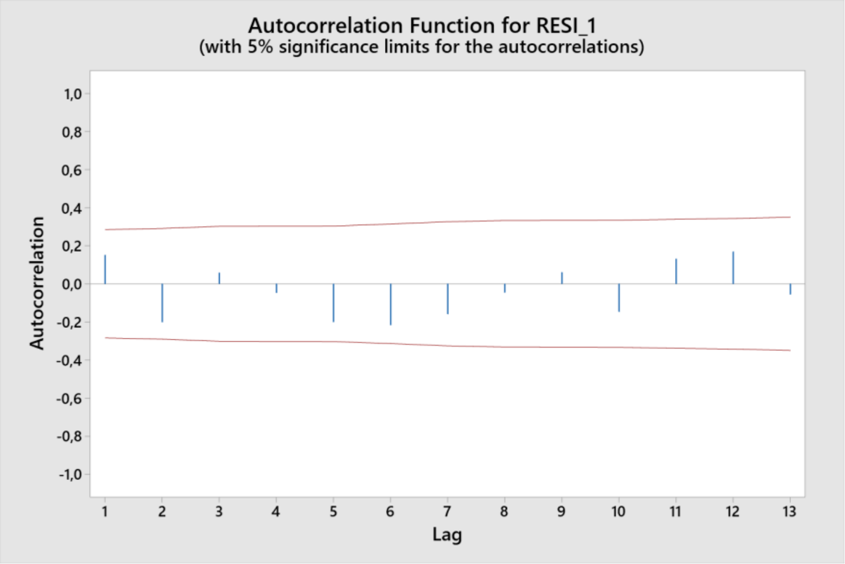


## Test

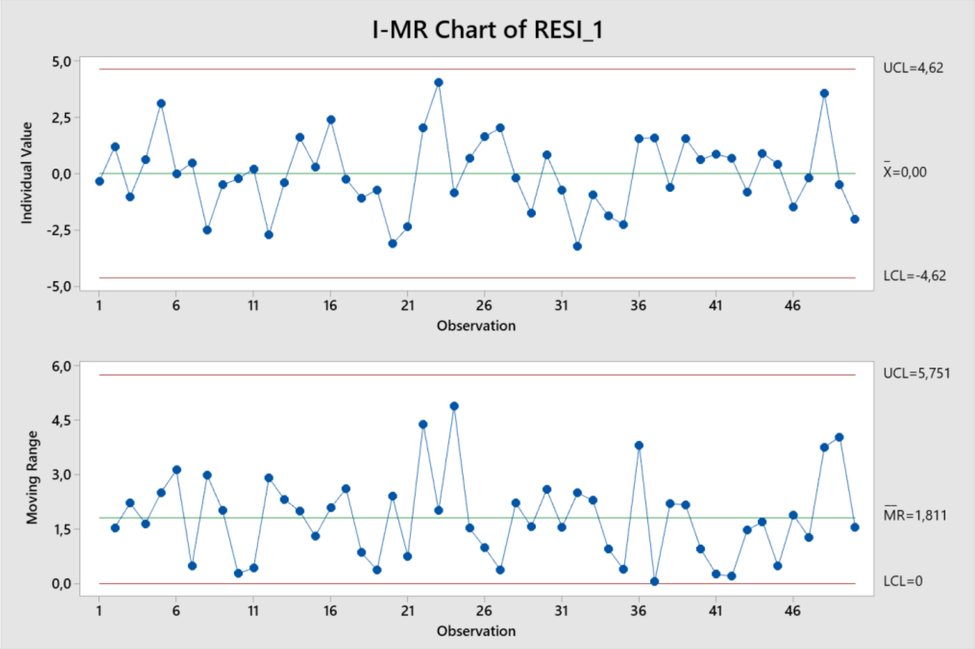
Null hypothesis H$_0$: The order of the data is random

Alternative hypothesis H$_1$: The order of the data is not random

| Number of Runs | | |
|---|---|---|
| Observed | Expected | P-Value |
| 25 | 25,84 | 0,809 |



A special cause control chart with $K = z_{\alpha/2} = 2{,}878$ is the following:

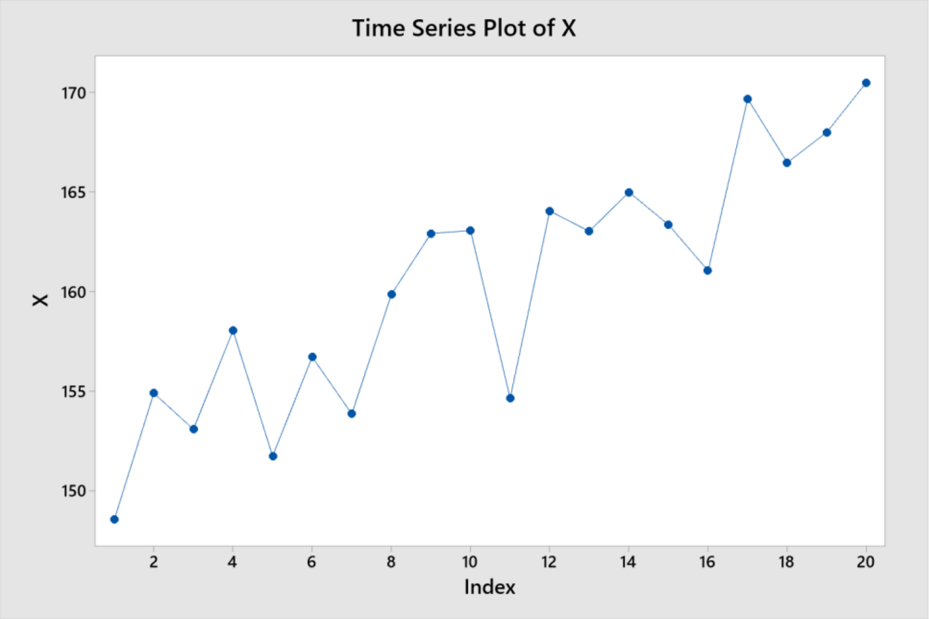


The process is in-control. This control chart allows monitoring each individual hardness measurement taking into account the location effect (assuming that such location effect is part of the natural signature of the process under control).

## Exercise 2

1)

Data snooping:



The time series exhibits an evident trend.

The trend model is the following:

EXERCISE2

# Regression Analysis: X versus t

## Regression Equation

X = 150,61 + 0,935 t

## Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 150,61 | 1,39 | 108,53 | 0,000 | |
| t | 0,935 | 0,116 | 8,07 | 0,000 | 1,00 |

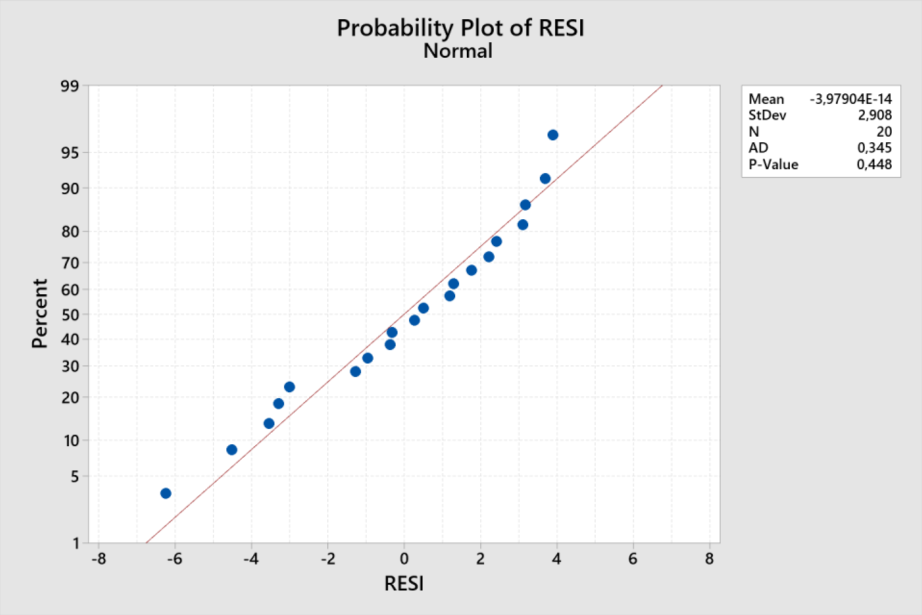
## Model Summary

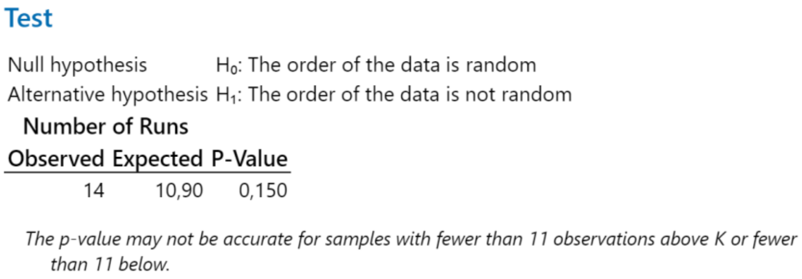
| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 2,98749 | 78,34% | 77,14% | 73,99% |

## Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 1 | 581,1 | 581,135 | 65,11 | 0,000 |
| t | 1 | 581,1 | 581,135 | 65,11 | 0,000 |
| Error | 18 | 160,7 | 8,925 | | |
| Total | 19 | 741,8 | | | |

The residuals are normal and independent:





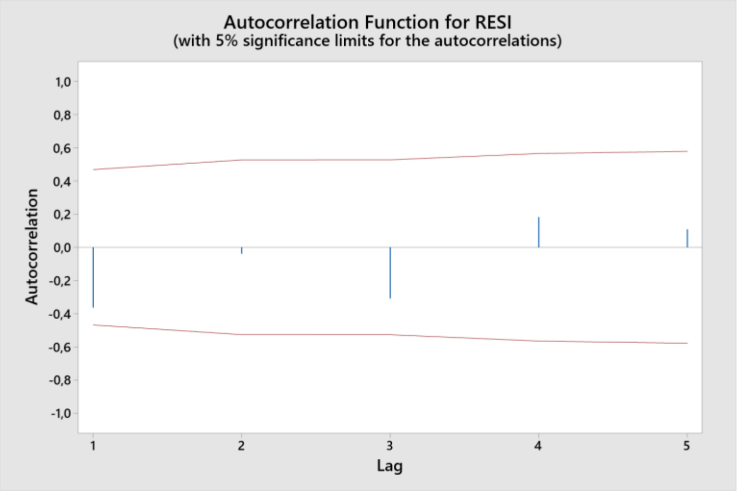
### Test

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 |
| 14 | 10,90 | 0,150 |

*The p-value may not be accurate for samples with fewer than 11 observations above K or fewer than 11 below.*



The trend control chart is the following:

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

$$CL = \ b_0 + b_1 t$$

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

where;

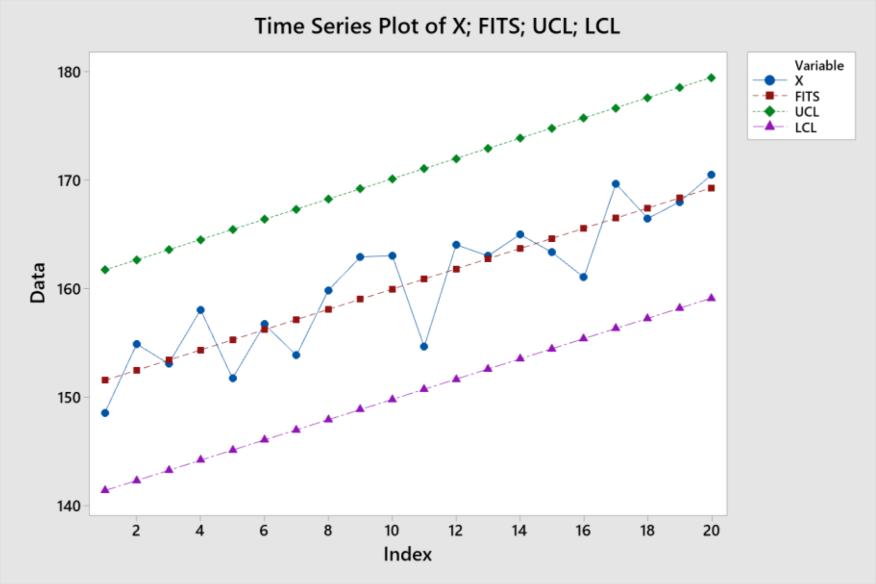
$$b_0 = 150{,}61,$$

$$b_1 = 0{,}935,$$

$$\widehat{\sigma_\varepsilon} = \frac{\overline{MR}}{d_2(2)} = \frac{3{,}99}{1{,}128} = 3{,}537$$

$$ARL_0 = \frac{1}{\alpha} = 250, K = z_{\alpha/2} = 2{,}878$$

The control chart is the following:



Assuming that the increasing trend of zinc release is the natural signature of the process, the process itself is in-control.

2)

Under the following conditions:

$$y_t = \mu_t + \sigma_\varepsilon$$

$H_0 : \mu_t = \beta_0 + \beta_1 t$

$H_1 : \mu_t = (\beta_0 + \delta_0 \sigma_\varepsilon) + \beta_1 t$

It is possible to estimate the Type II error as a function of an absolute shift of the intercept of the model as follows:

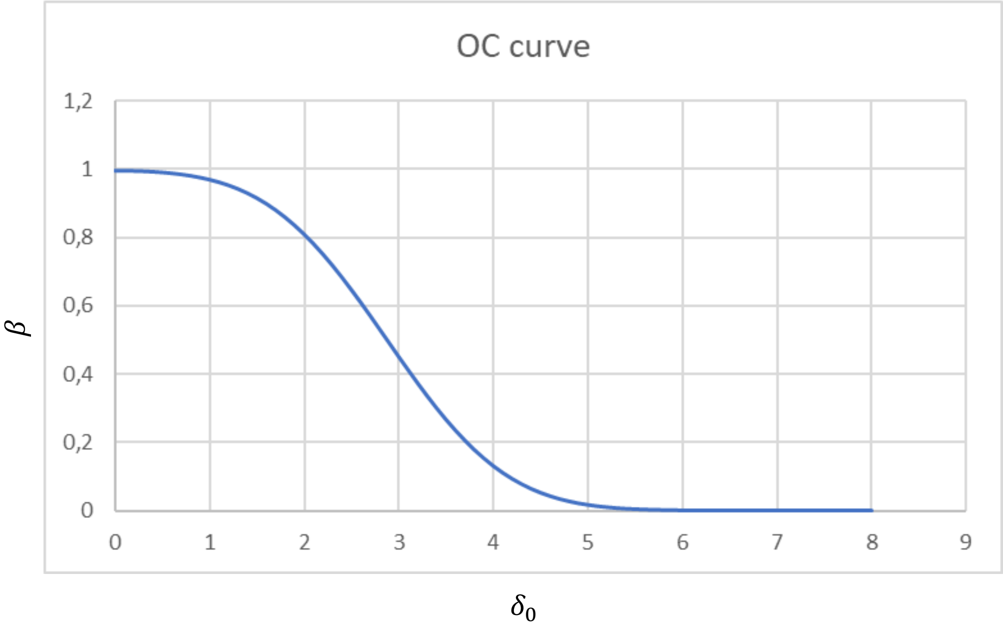
$$\beta(\delta_0) = \Phi\left(\frac{UCL - \mu_t}{\sigma_\varepsilon} | H_1\right) - \Phi\left(\frac{LCL - \mu_t}{\sigma_\varepsilon} | H_1\right) =$$

$$= \Phi\left(\frac{UCL - (\beta_0 + \delta_0 \sigma_\varepsilon) - \beta_1 t}{\sigma_\varepsilon}\right) - \Phi\left(\frac{LCL - (\beta_0 + \delta_0 \sigma_\varepsilon) - \beta_1 t}{\sigma_\varepsilon}\right) =$$

$$= \Phi\left(\frac{\beta_0 + \beta_1 t + K\sigma_\varepsilon - (\beta_0 + \delta_0 \sigma_\varepsilon) - \beta_1 t}{\sigma_\varepsilon}\right) - \Phi\left(\frac{\beta_0 + \beta_1 t - K\sigma_\varepsilon - (\beta_0 + \delta_0 \sigma_\varepsilon) - \beta_1 t}{\sigma_\varepsilon}\right) =$$

$$= \Phi(K - \delta_0) - \Phi(-K - \delta_0)$$

The corresponding OC curve is:



Type II error values:

| $\delta_0$ | $\beta$ |
|---|---|
| 2 | 0,810 |
| 4 | 0,131 |

3)

Under the following conditions:

$H_0: \mu_t = \beta_0 + \beta_1 t$

$H_1: \mu = \beta_0 + (\beta_1 + \delta_1 \sigma_\varepsilon)t$

The Type II error can be estimated as follows:

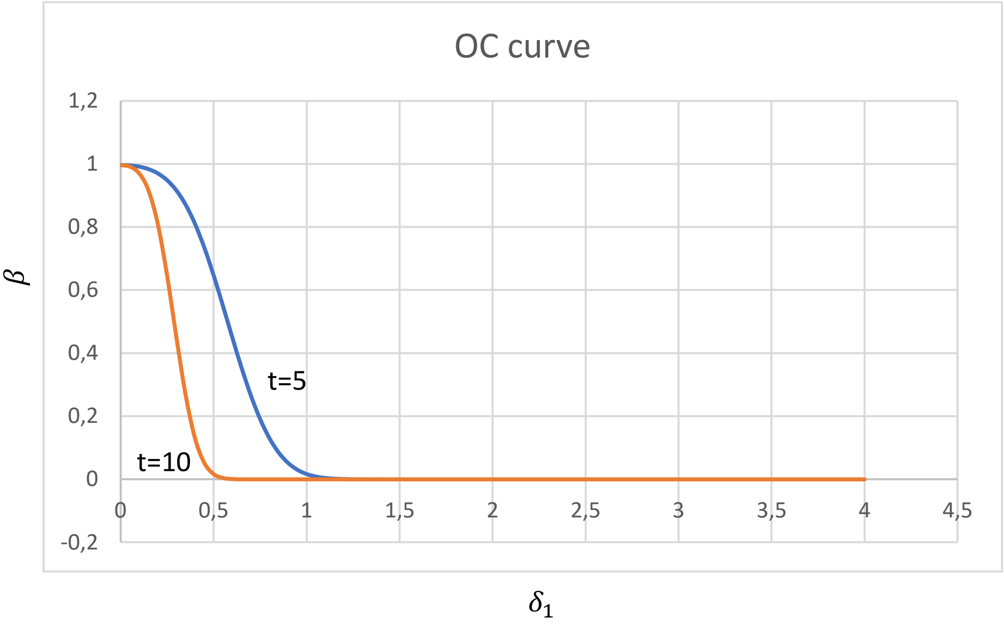
$$\beta(\delta_1) = \Phi\left(\frac{UCL - \beta_0 - (\beta_1 + \delta_1\sigma_\varepsilon)t}{\sigma_\varepsilon}\right) - \Phi\left(\frac{LCL - \beta_0 - (\beta_1 + \delta_1\sigma_\varepsilon)t}{\sigma_\varepsilon}\right) =$$

$$= \Phi\left(\frac{\beta_0 + \beta_1 t + K\sigma_\varepsilon - \beta_0 - (\beta_1 + \delta_1\sigma_\varepsilon)t}{\sigma_\varepsilon}\right) - \Phi\left(\frac{\beta_0 + \beta_1 t - K\sigma_\varepsilon - \beta_0 - (\beta_1 + \delta_1\sigma_\varepsilon)t}{\sigma_\varepsilon}\right) =$$

$$= \Phi(K - \delta_1 t) - \Phi(-K - \delta_1 t)$$

In this case, the Type II error depends not only on the shift on the $\beta_1$ parameter, but also on time t.

The two OC curves at time t = 5 min and time t = 10 min are shown below:



Values of the Type II error as a function of the absolute shift and at different times:

| $\delta_1$ | t | $\beta$ |
|---|---|---|
| 0,25 | 5 | 0,948 |
| | 10 | 0,647 |
| 0,5 | 5 | 0,647 |
| | 10 | 0,017 |

If a change of slope occurs, it is more difficult to detect it at the beginning of the process, but it becomes more and more likely to detect such shift as the out-of-control model deviates from the in-control one as the process goes on (i.e., at larger t).

### Exercise 3

1)

Being known that, under the given assumptions:

$S^2 \sim [\sigma^2/(n-1)]\chi^2(n-1)$

Then:

$P(F_{1-\frac{\alpha}{2},n-1,n-1} \leq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \leq F_{\frac{\alpha}{2},n-1,n-1}) = 1-\alpha,$

and the probabilistic control limits for the ratio between the two variances can be computed as:

$$UCL = \frac{\sigma_1^2}{\sigma_2^2} F_{\frac{\alpha}{2},n-1,n-1} = 12{,}44$$

$$CL = \frac{\sigma_1^2}{\sigma_2^2} = 1{,}4$$

$$LCL = \frac{\sigma_1^2}{\sigma_2^2} F_{1-\frac{\alpha}{2},n-1,n-1} = 0{,}158$$

Where: $\sigma_1^2 = 3{,}5$, $\sigma_2^2 = 2{,}5$, $\alpha = 0.01$ and $n = 8$.

2)

The probability to detect an increase of the ratio between the variances when $H_1\!: \frac{\sigma_{1,1}^2}{\sigma_{1,2}^2} = 4\frac{\sigma_{0,1}^2}{\sigma_{0,2}^2}$ can be estimated as follows:

$$1-\beta = 1 - P\left(\frac{S_1^2}{S_2^2} \in [LCL, UCL] \middle| \frac{\sigma_{1,1}^2}{\sigma_{1,2}^2}\right) = 1 - \left[P\left(\frac{S_1^2}{S_2^2} < UCL \middle| \frac{\sigma_{1,1}^2}{\sigma_{1,2}^2}\right) - P\left(\frac{S_1^2}{S_2^2} < LCL \middle| \frac{\sigma_{1,1}^2}{\sigma_{1,2}^2}\right)\right]$$

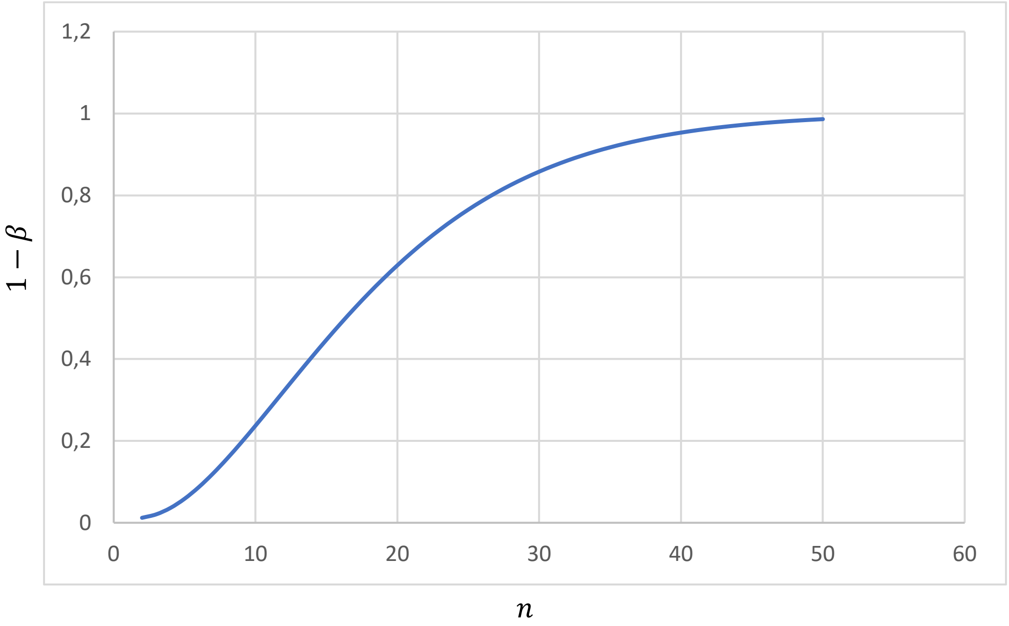
$$P\left(\frac{S_1^2}{S_2^2} < UCL \middle| \frac{\sigma_{1,1}^2}{\sigma_{1,2}^2}\right) = P\left(\frac{S_1^2}{S_2^2}\frac{\sigma_{0,2}^2}{\sigma_{0,1}^2}\frac{1}{4} < \frac{\sigma_{0,1}^2}{\sigma_{0,2}^2}\frac{\sigma_{0,2}^2}{\sigma_{0,1}^2}\frac{1}{4}F_{\frac{\alpha}{2},n-1,n-1}\right) = P\left(F_{n-1,n-1} < \frac{1}{4}F_{\frac{\alpha}{2},n-1,n-1}\right)$$

$$P\left(\frac{S_1^2}{S_2^2} < LCL \middle| \frac{\sigma_{1,1}^2}{\sigma_{1,2}^2}\right) = P\left(\frac{S_1^2}{S_2^2}\frac{\sigma_{0,2}^2}{\sigma_{0,1}^2}\frac{1}{4} < \frac{\sigma_{0,1}^2}{\sigma_{0,2}^2}\frac{\sigma_{0,2}^2}{\sigma_{0,1}^2}\frac{1}{4}F_{1-\frac{\alpha}{2},n-1,n-1}\right) = P\left(F_{n-1,n-1} < \frac{1}{4}F_{1-\frac{\alpha}{2},n-1,n-1}\right)$$

Thus:

$$1-\beta = 1 - \left[P\left(F_{n-1,n-1} < \frac{1}{4}F_{\frac{\alpha}{2},n-1,n-1}\right) - P\left(F_{n-1,n-1} < \frac{1}{4}F_{1-\frac{\alpha}{2},n-1,n-1}\right)\right]$$

This probability as a function of the sample size n can be represented as follows:



The smallest sample size to detect the out-of-control state with a probability ≥70% is $n = 23$.