SHEWHART'S CONTROL CHARTS OF SAMPLE MEANS FOR NONNORMAL DISTRIBUTION OF QUALITY VARIABLES

Ing. Jan Král [1], *RNDr. Jiří Michálek, CSc.* [2], *Ing. Josef Křepela* [3]

Key words: Shewhart's control chart; control limits; Box-Cox and Johnson transformations; CLT

1 INTRODUCTION

Using Shewhart´s control charts for process control in practice, we should assume the following two conditions are fulfilled:

1) a manufacturing process is statistically managed and stable (mean value and variability do not change at time),
2) normal distribution of the monitored quality variables.

The first condition is not often met in practice, the Six Sigma methodology assumes a fluctuating mean value within a certain small range (±1,5 σ) about a target value. It causes a significant increase in the risk of false alarms, which is set in Shewhart´s chart at α=0,0027, taking into account both control limits. It implies that a point outside the control limits can only appear in average once in 740 cases. Hence, changes in mean value increase this risk, sometimes up to α=0,10, which means that a point can appear randomly outside the control limits once in 20 cases in average. These are false alarms that operators have to solve although any intervention is not necessary. Such a situation leads to a worse motivation in work and the application of control charts becomes ineffective.

The second condition – condition of normality – is usually in practice ignored because it is very often supposed that normality of sample means within subgroups is automatically satisfied. This approach is motivated by the practical use of Central Limit Theorem (CLT) but its application is strongly influenced by the range of a subgroup and the shape of the probability distribution of an observed quality variable. If we assumed incorrectly the validity of normal distribution this fact could cause the enormous number of false alarms similarly as in the first condition.

While the first problem with nonstability in means can be relatively easily fixed by extended control limits that are constructed on total variability which contains both variability inside subgroups and variability among subgroups caused by the behaviour of means, for more details see e.g. [1], [2].
The second condition is more complex as the correct calculation of control limits is strongly dependent on the concrete probability distribution and to find a suitable model describing the underlying variable can be difficult. The main goal of this contribution is to show in a case study what danger is hidden in a formal application of original Shewhart´s limits and to give an advice how to overcome this problem by a suitable transformation of data.

2 CONVERGENCE OF DISTRIBUTION FUNCTIONS AND CENTRAL LIMIT THEOREM

In this part we need some theoretical notions and to say something about Central Limit Theorem (CLT) Let us have a sequence $\{X_n\}$ of random variables and the sequence of the corresponding distribution funtions $\{F_n(x)\}$. Further, we have a random variable X having distribution F(x). When

$$\lim_{n\to\infty} F_n(X) = F(X) \qquad (1)$$

at all points of continuity of F(x) we say that the sequence $\{X_n\}$ tends to X in distribution. Distribution function F(x) is then called the asymptotic distribution for the given sequence of random variables. In other words, for large n distribution function $F_n(x)$ can be approximated by F(x). A very important case is when the limiting distribution is normal $N(\mu, \sigma^2)$.Then we say that random variable has asymptotic normal distribution and

$$\lim_{n\to\infty} F_n(X) = F(X) = \Phi\left(\frac{X - \mu}{\sigma}\right) \qquad (2)$$

is valid at all real numbers. For the simplest case of CLT we will assume that a sequence of random variables is formed by mutually stochastically independent variables and identically distributed with a quite arbitrary distribution function having finite dispersion.

The Central Limit Theorem is a very important tool in probability theory and mathematical statistics. Formally, the theorem can be expressed as follows:

$$\sqrt{n}\left(\frac{1}{n}\sum_{1}^{n} X_i - \mu\right) \to N(0,\sigma^2) \qquad (3)$$

in distribution. This convergence means that distribution function of $\sqrt{n}(\overline{X} - \mu)$ converges to distribution function of $N(0,\sigma^2)$ at every real number. A very interesting question is the rate of this convergence. The classical result due to Berry-Esseen states that the rate of convergence is $n^{-0,5}$, more precisely, under the existence of the absolute third moment ρ, there exists a positive constant *C* such that for all real *x* and all *n*

$$\left|F_n(x) - \Phi(x)\right| \le \frac{C\rho}{\sigma^3\sqrt{n}}\,. \qquad (4)$$

The value of constant *C* originally was estimated as 7,59, the last results from the year 2010 show the value 0,4785. On the other side, it is possible to give for *C* a lower bound, namely *C* must be greater than 0,40973. Usually, based on experience, a sufficient number *n* of observations is 30-35 for a good approximation by normal distribution. But, when we apply control charts for subgroups the range of a subgroup is much smaller, even two or three pieces only. Then the behaviour of calculated averages within subgroups may be very difficult to be approximated by normal distribution. The convergence to normal distribution is dependent not only on number *n* but also on the shape of the underlying distribution of a quality variable. The skewness is also very important factor. The following case study shows that even by a very frequently used in technical practice the log-normal distribution the situation with a formal application of classical Shewhart´s control limit can lead to a sufficiently greater number of false alarms.

The convergence of sample means is very well illustrated on the following Figure 1 where three histograms are depicted. The underlying data are originated from the uniform distribution. Histogram 1 depicts the original data, Histogram 2 shows the sample means of two values and Histogram 3 shows the sample means of four values. The sample size was 250 in every case.

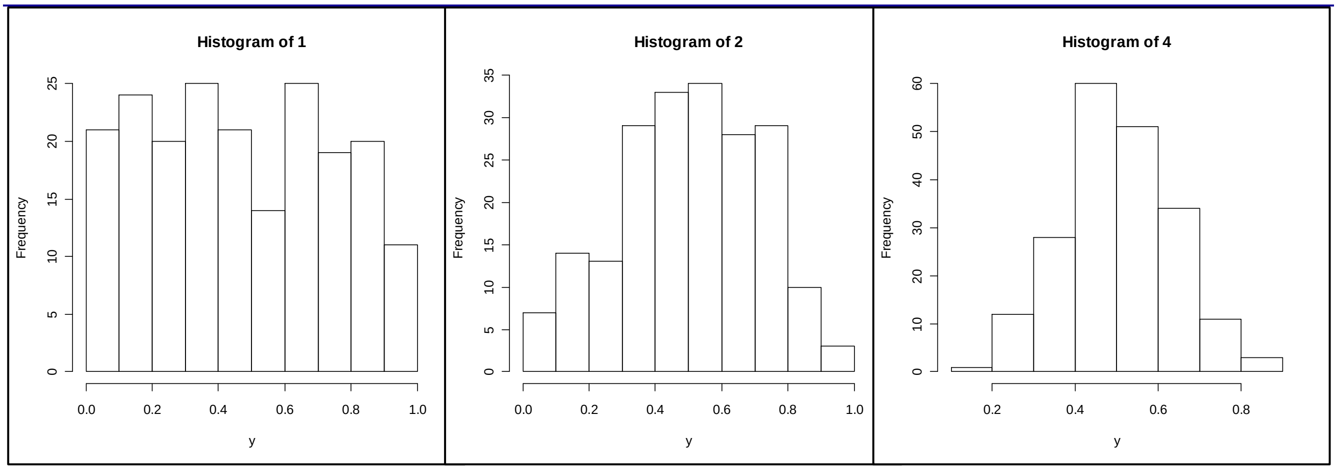


Figure 1: Convergence of sample means to normal distribution

3 CASE STUDY

3.1 INITIAL CONDITIONS

In the study we will analyze data coming from a manufacturing process and it is assumed to have a relatively large sample for the possibility of a good knowledge of probability distribution. The sample size of data in Fig.2 is 10000. The empirical distribution of data is seen from the corresponding histogram and this histogram is compared to the curve of normal distribution with parameter μ equal to arithmetic mean and σ equal to sample standard deviation of data. This chart illustrates an evident difference between normal distribution and distribution of given data.

One of goodness fit tests for testing normality is based on sample skewness and sample kurtosis.

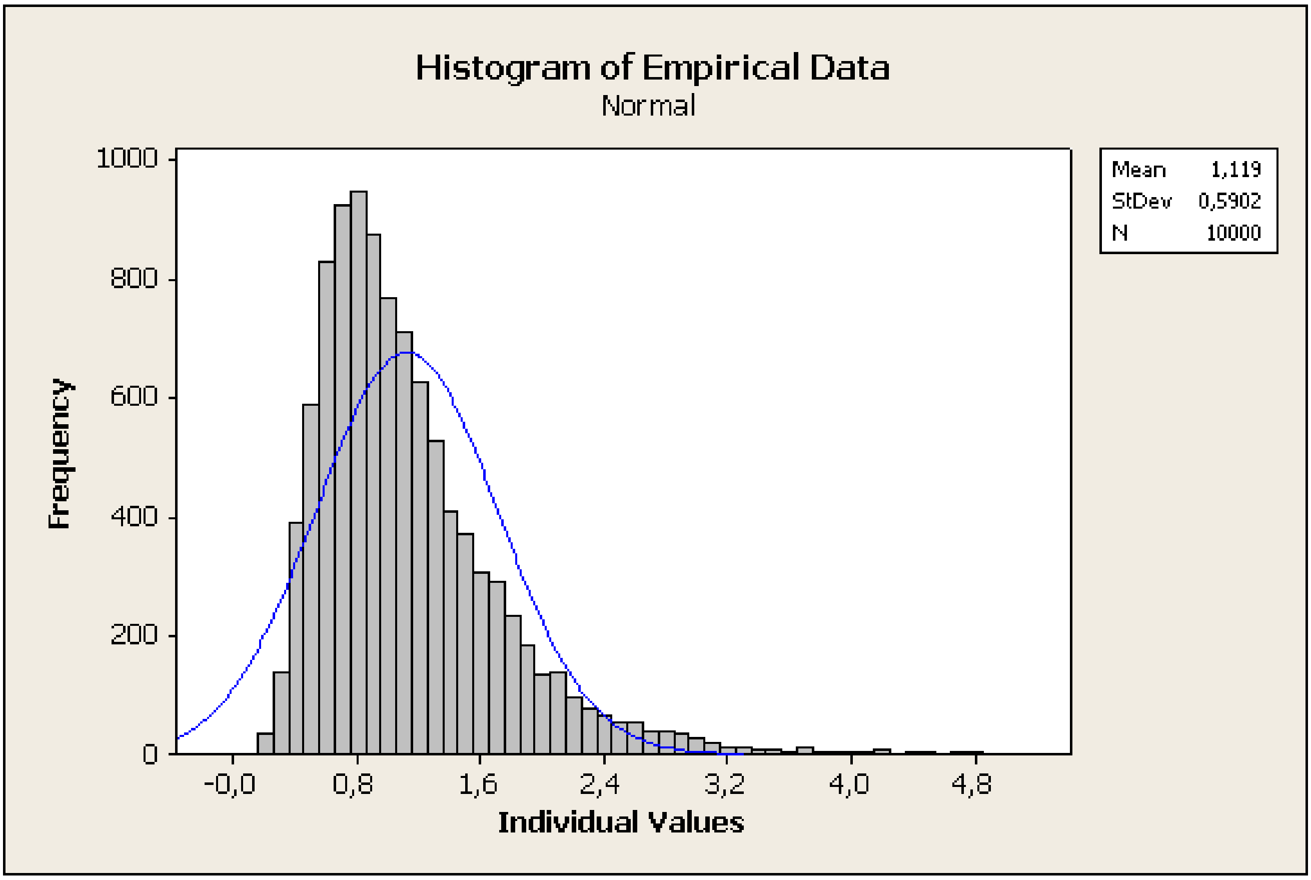


Figure 2: Histogram of data with a curve of normal distribution

Coefficient of skewness describes how a distribution function is symmetric about its mean value. For normal distribution skewness is 0. This value is compared to the sample coefficient of skewness that is 1,60687 in our case. The test says the difference is statistically significant and data cannot come from normal population. A similar situation we can see by kurtosis. Coefficient of kurtosis is a measure of concentration of probability about mean value and in the case of normal distribution is equal to 3. Its sample version obtained from data is 4,21652 and is also significant. This fact also confirms nonnormality of data.

Another possibility for comparing data to a suitable model given by a distribution function is probability chart. Here the considered distribution is transformed together with data in a chart where the distribution is depicted as a straight line. The closer data are to this line, the better fit with the model. The obtained result is quite clear, data cannot be described by the model with normal distribution.

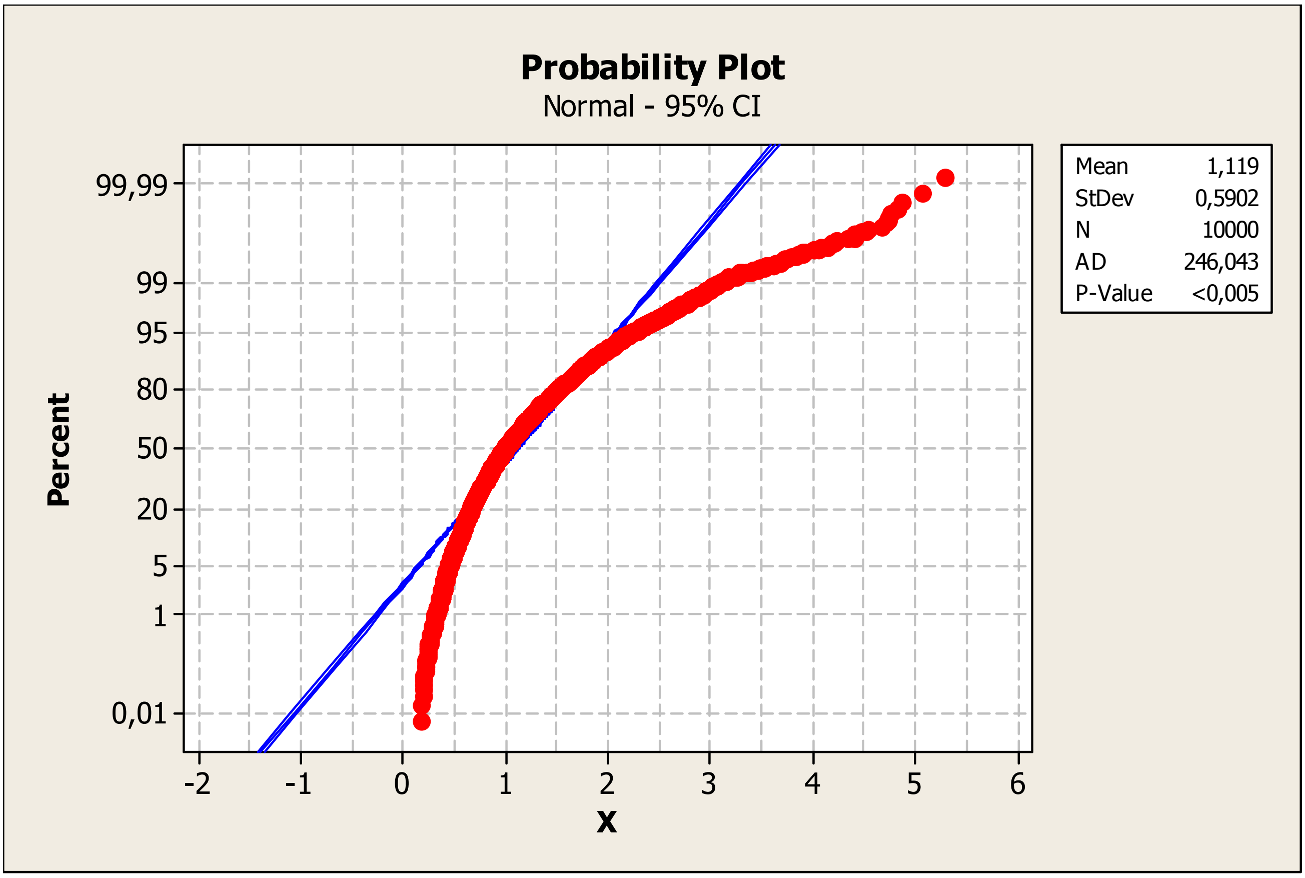


Figure 3: Probability chart

The result is given by the application of Anderson-Darling test, the corresponding p-value is almost negligable, smaller than 0,005 and with the level of significance 5% we must reject normality.

3.2 CONVENTIONAL APPROACH IN PRACTICE

Our task is to statistically control this process by use of Shewhart´s charts. The recommended chart in many cases is of the type (xbar, R). Each subgroup contains three checked pieces.

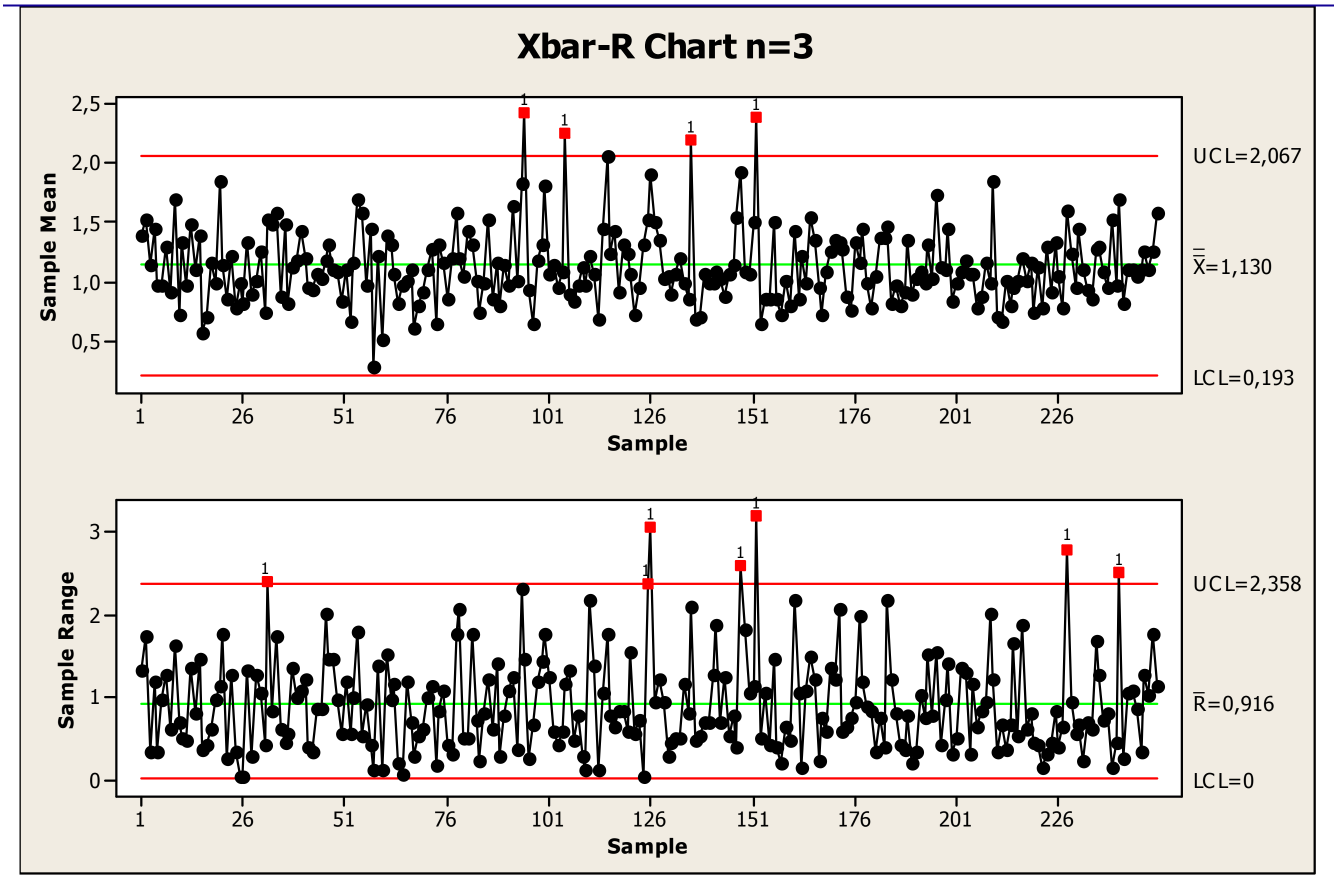


Figure 4: Control chart (xbar,R)

As seen from Fig.4 there are false alarms caused by the too low upper control limit that is calculated under the assumption of normality. This fact is provoked by nonsymmetry in distribution of data and sample means of the size 3 cannot remove this influence.

3.3 SUGGESTED APPROACH FOR SAMPLE MEANS

Observing a quality variable, which is not normally distributed the first step is to attempt to identify the probability distribution for sample means. It is quite natural to use some statistical software by identifying a suitable model. E.g., Minitab 16 offers 14 types of models that are very frequent in technical practice. At this moment there are three possible scenarios.

3.3.1 Identification of a distribution function for sample means

In the next Fig. 5 we can see that from all the possibilities offered by Minitab 16 only three are acceptable, the case with normal distribution is only for comparison. A potentional model expressing our data is log-normal distribution, other possibilities are based on transformations of original data into new data having normal distribution. The first transformation is called Box-Cox, the other is Johnson transformation, which transforms data even into new data with N(0,1). For the next decision, which model should be chosen, we will progress according to the greatest p-value of Anderson-Darling test.
As for the case of log-normal distribution the corresponding p-value is 0,211, we cannot reject this model. A similar result is given by Box-Cox transformation. But, we see that Johnson transformation is the winner with p-value equal to 0,888 (see Table 1).

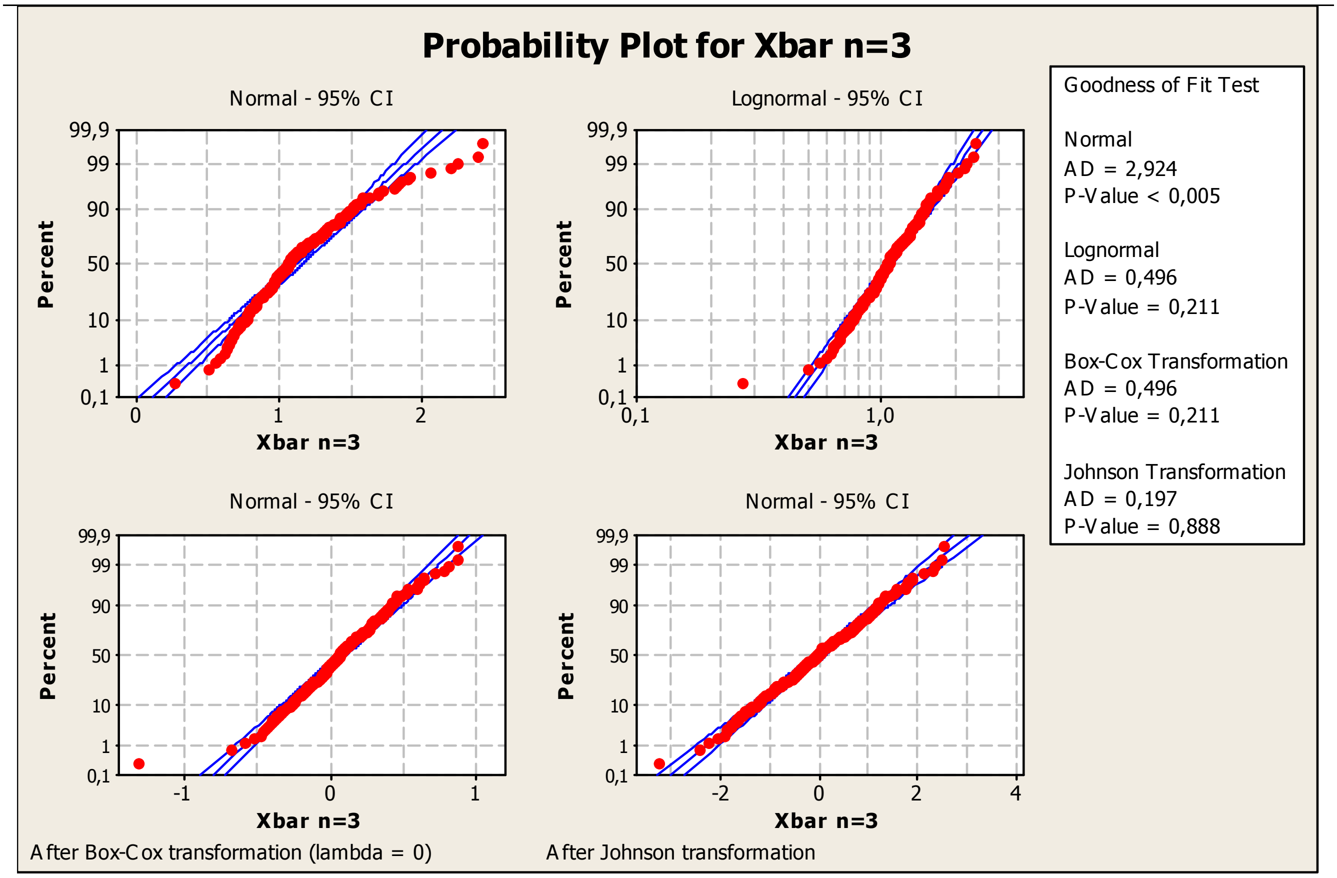
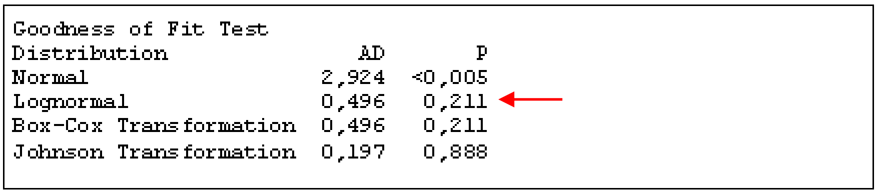


Figure 5: Identification of Distribution function

When we have chosen a priori the significance level 5%, we cannot reject any of those three models presented above. As Johnson transformation returns the maximal p- value, we can proceed in the following manner. The transformed data are distributed N(0,1) and hence the control limits with the false alarm risk 0,27% must be theoretically -3, +3. Using the inverse Johnson transformation, we will obtain new control limits respecting the probability distribution of sample means. If the winner were log-normal distribution the progress would be the following. On the basis of the corresponding density function we would calculate 0,135% and 99,865% quantiles and their values would be control limits for subgroup means. Parameters of log-normal density function are usually estimated by the maximal likelihood principle (see Table 2).

Table 1: Goodness of Fit



```
Goodness of Fit Test
Distribution                  AD       P
Normal                     2,924  <0,005
Lognormal                  0,496   0,211 ←
Box-Cox Transformation     0,496   0,211
Johnson Transformation     0,197   0,888
```

In case no model and also no transformation were found the situation is somewhat worse because in such a case we are obliged to estimate 0,135% and 99,865% quantiles directly from the data. We will need a statistical software to do it.

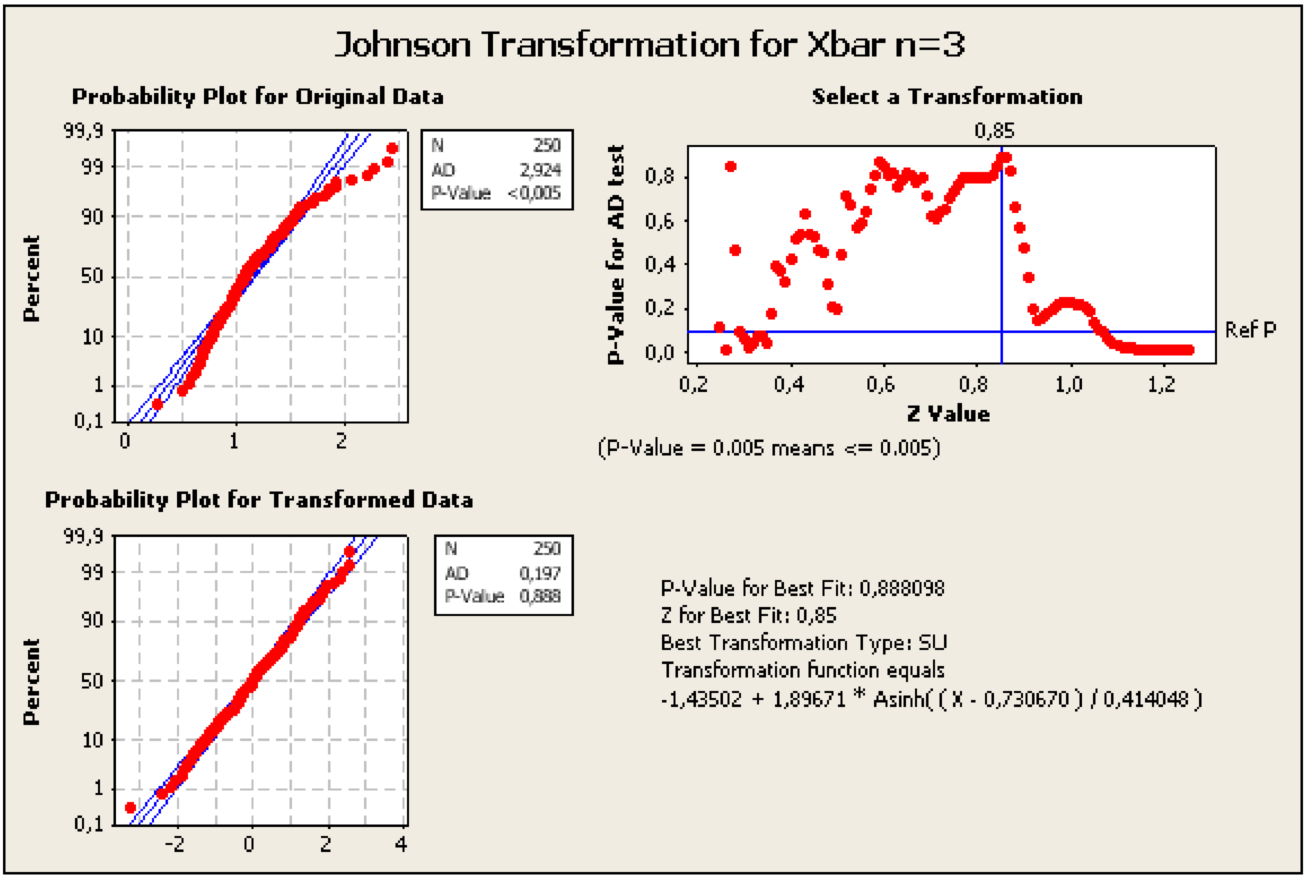
In the next table we see the corresponding quantiles for log-normal distribution and these values can be used as modified control limits for sample means within subgroups. The central line of the chart is defined by sample median.

Table 2: Table of Percentiles

Lognornal:	
Percent	Percentiles
0,135	0,46178
50	1,08602
99,865	2,55414

3.3.2 Johnson transformation

Here the approach based on Johnson transformation is described more in detail. All the calculations are carried out in Minitab 16. First, we need to find a transformation equation that will transform original data into data with distribution N(0,1). In Fig.6 we have everything prepared by software.



N	250
AD	2,924
P-Value	<0,005

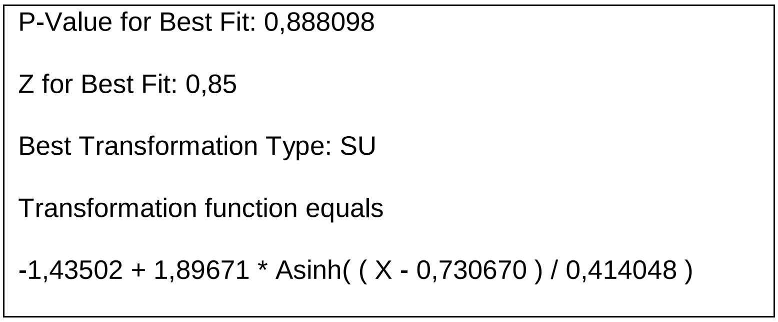
(P-Value = 0.005 means <= 0.005)

N	250
AD	0,197
P-Value	0,888

P-Value for Best Fit: 0,888098
Z for Best Fit: 0,85
Best Transformation Type: SU
Transformation function equals
-1,43502 + 1,89671 * Asinh((X - 0,730670) / 0,414048)

Figure 6: Johnson transformation

The equation defining a suitable transformation is presented in Fig.7.



```
P-Value for Best Fit: 0,888098

Z for Best Fit: 0,85

Best Transformation Type: SU

Transformation function equals

-1,43502 + 1,89671 * Asinh( ( X - 0,730670 ) / 0,414048 )
```

Figure 7: Equation for transformation

Transformed values of sample means are depicted in a classical Shewhart´s chart for individual values and control limits were calculated from transformed data. As seen in Fig.8 these limits are almost identical to theoretical ones -3 and +3.

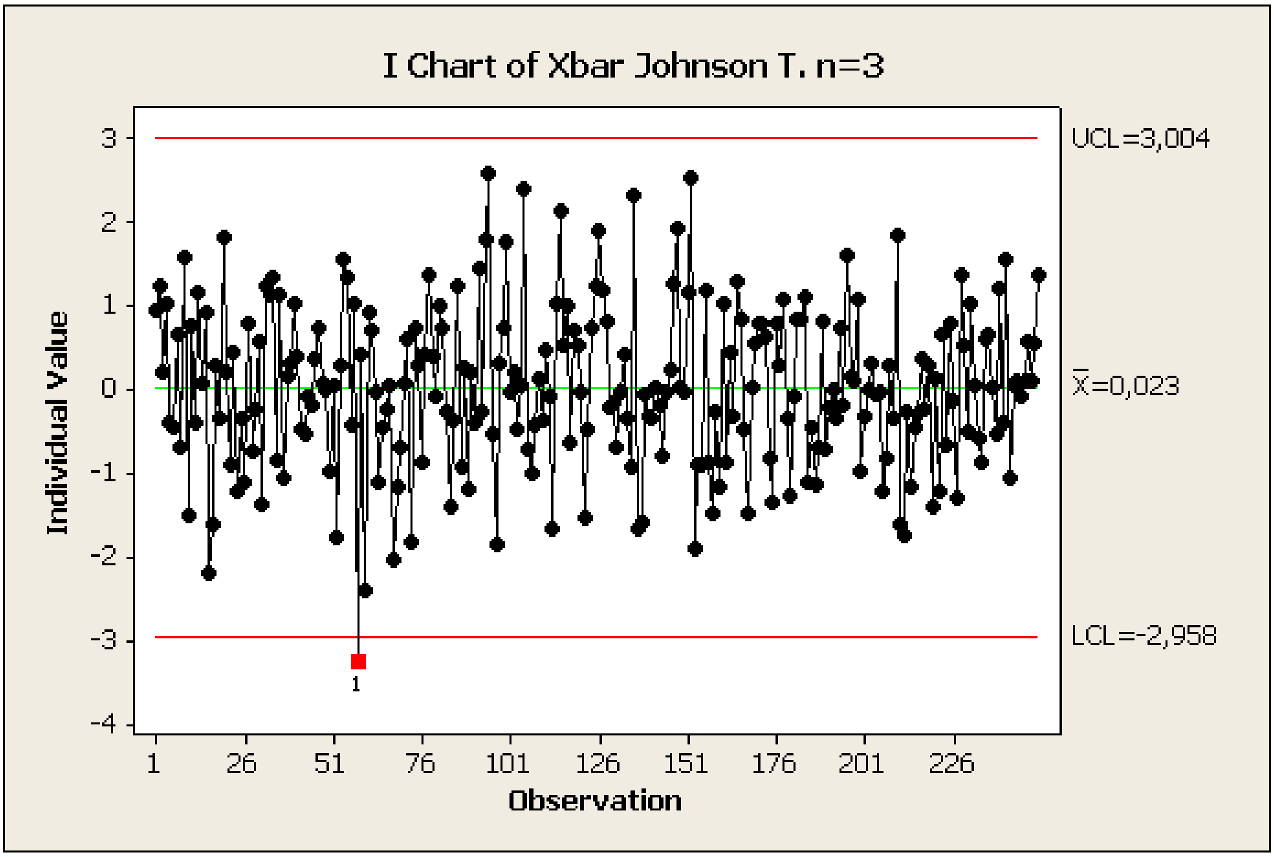


Figure 8: Control chart for individual transformed values

The control limits found in this way will be transformed by use of the inverse Johnson transformation into original sample means and a modified control chart with new control limits can be constructed. The result is seen in Fig. 9 where new limits are compared to original Shewhart´s limits for sample means of the size 3.

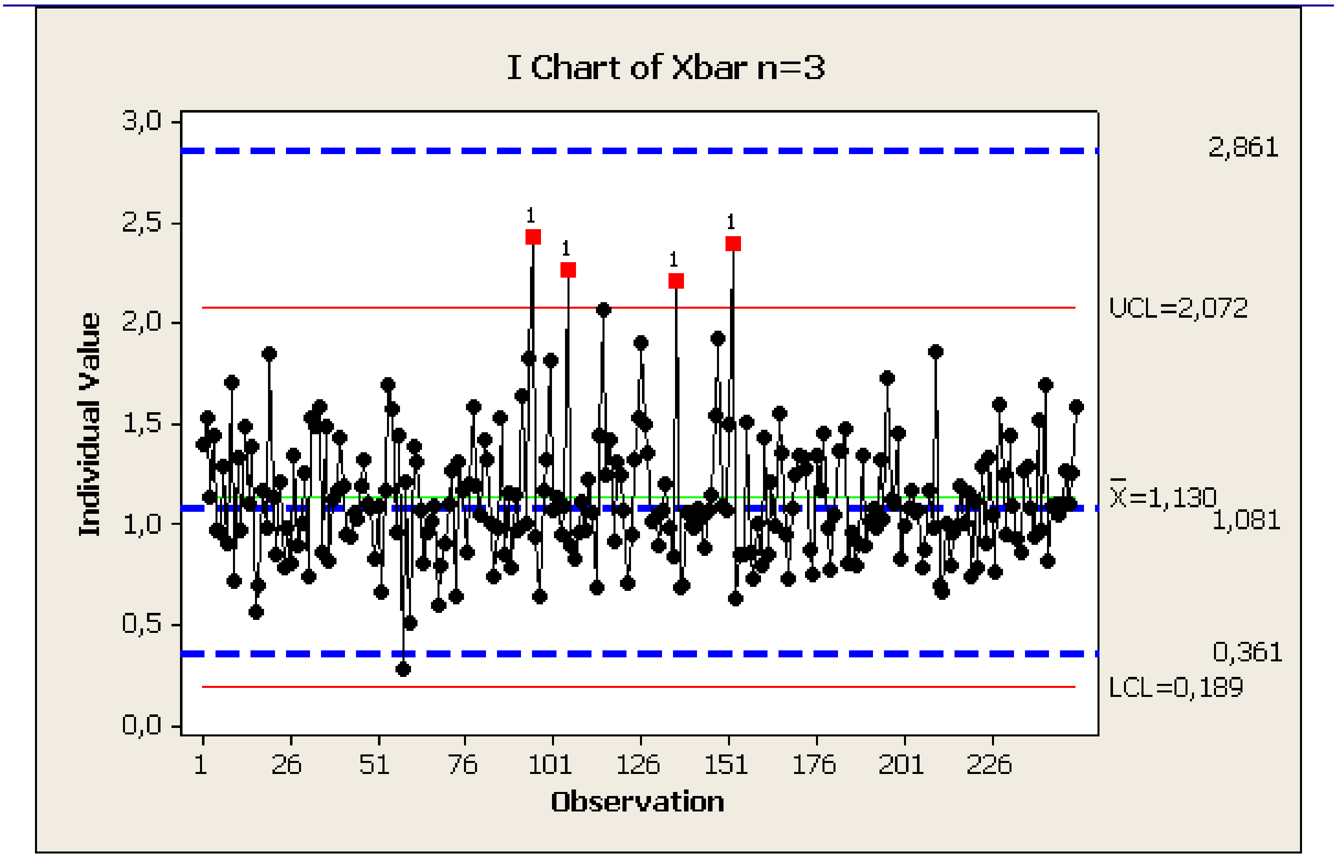


Figure 9: Control chart with modified limits

At the first sight we immediately see no false alarms in this modified case, limits obtained by the above described approach can serve for statistical control of our manufacturing process. The similar progress based on estimated quantiles is applicable when a concrete model with a suitable distribution function is found.

3.3.3 Calculation based on empirical percentiles

In the worst case when no model nor any transformation were found we in fact have the only possibility to estimate suitable limits directly from original sample means via empirical percentiles corresponding to 0,135% and 99,865%. But, this situation is very strongly dependent on number of data and this case can very often occur in practice because reliable estimates of quantiles need a lot of data. There exists a possibility to use bootstrap technique for improving estimates. In the next Fig.10 we see the final result and we can compare all the approaches suggested in this contribution. The original Shewhart´s limits are in red colour (solid line), the limits based on Box-Cox transformation with log-normal model are in green color (dash-dot line) and the limits given by Johnson transformation are blue (dashed line).

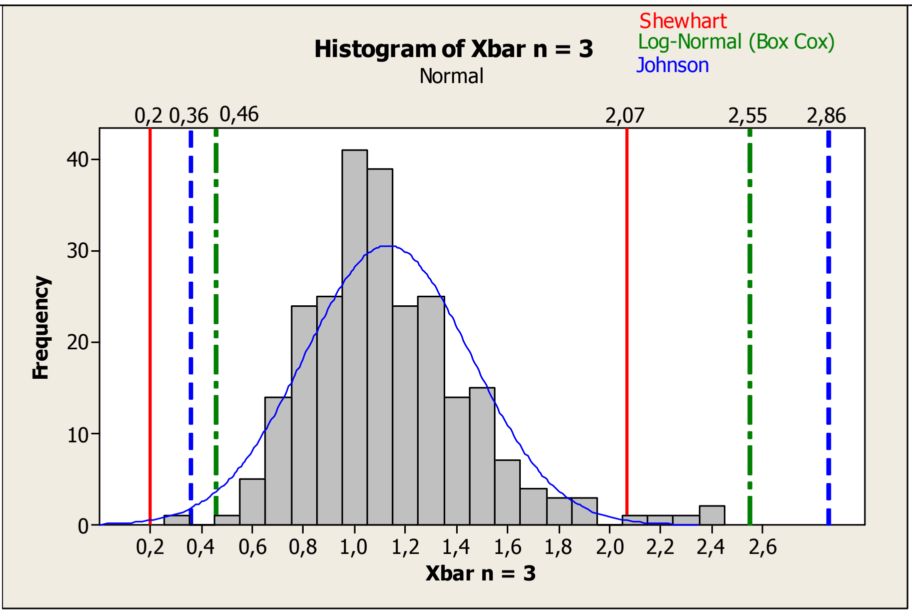


Figure 10: Histogram with all control limits

4 CONCLUSION

One can use a classical control chart for statistical control of a quality variable that is not normally distributed without any apprehension when the size of a subgroup is relatively large (n >> 10). Unfortunately, such a case is rare in practice. The size of subgroups is usually small, a few of pieces. In this situation it is necessary to analyze the probability distribution of sample means from subgroups. When a suitable model is found then the corresponding quantiles can serve as modified control limits. If no model was identified the further approach is to use some transformations that would transform original sample means into new normally distributed data. Statistical softwares very often offer Box-Cox or Johnson transformations. Then the modified control limits are calculated via inverse transformations. On the basis of these facts it is necessary to realize that a formal and automatic application of classical Shewhart´s charts can provoke unpleasant reaction by the statistical control of a manufacturing process caused by a series of false alarms. CLT is a very mighty probabilistic tool but needs a careful manipulation.

Acknowledgement

This contribution is supported by the project No.1M06047 Center of Quality and Reliability in Production of Ministry of Education, Youth and Sport of the Czech Republic.

References

[1] KRÁL, Jan. Modified Shewhart's control charts implementation. In: *Current trends in statistics in V6 region: Proceedings of Student`s Conference: Prague, 5.9.-6.9.2008*. Praha: Czech Statistical Society, 2009. ISBN 978-80-904330-0-7. p. 72-85.

[2] FABIAN, František., HORÁLEK Vratislav., KŘEPELA Josef., MICHÁLEK Jiří., CHMELÍK Václav., CHODOUNSKÝ Jiří., KRÁL Jan. *Statistical Methods in Quality Management*, Prague, Czech Society for Quality, 2007, ISBN 978-80-02-01897-1, 390 pp.(in Czech)

Autors

1 ISQ PRAHA s.r.o., Pechlátova 19, CZ-150 00, Prague 5, Czech Republic

2 Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, CZ-182 08, Prague 8, Czech Republic

3 Czech Technical University, Faculty of Mechanical Engineering, Department of Technology, Technická 4, CZ-166 07, Prague 6 Dejvice, Czech Republic