



¹ Miriam ANDREJIOVÁ, ² Zuzana KIMÁKOVÁ

THE OPEN SOURCE SOFTWARE “R” IN THE STATISTICAL QUALITY CONTROL

^{1,2} TECHNICAL UNIVERSITY IN KOŠICE, FACULTY OF MECHANICAL ENGINEERING, KOŠICE, DEPARTMENT OF APPLIED MATHEMATICS AND INFORMATICS, SLOVAKIA

ABSTRACT: In today's modern society, the importance of the word “quality” is markedly increasing, whether it is product quality from customers' perspective or from the perspective of producers, distributors, service providers, etc. Quality is becoming a very important tool for keeping the customers satisfied and maintaining the competitiveness of producers and organizations.

KEYWORDS: statistical process quality control, R project

INTRODUCTION

Quality has recently become a matter of fact in all spheres of life. According to international standard ISO 9000, “quality is a measure of the degree to which the output achieves the task goals”. Quality control is a sum of all means used to meet the quality criteria required by a customer. These criteria, along with specific requirements on properties of products or services, are gaining an ever increasing importance in the competition fight.

Any production process yields a large amount of recorded data that need to be processed and analyzed. Statistical methods and tools, which are currently considered a powerful means of quality control, enter the QC process at this point. It is important to be familiar with the basic characteristics and methods used in applying statistics to quality control. Correct application of these methods is a key element of all phases of process control.

Statistical methods used in statistical process control stem from basic knowledge and principles of the probability theory and mathematical statistics. These methods represent a set of mathematical and statistical tools that facilitate achieving and maintaining a production process at such a level as to ensure product compliance with specified requirements. Statistical methods are usually divided into three categories:

1. Simple (basic, elementary) statistical methods, i.e. the seven basic quality tools (check sheet, histogram, flowchart, cause-and-effect diagram, Pareto chart, scatter diagram, control chart);
2. Semi-sophisticated statistical methods, such as distribution, estimation theory, hypothesis testing, statistical thinning, the theory of errors, ANOVA, regression and correlation analysis, reliability evaluation methods, or the methods of experiment design and evaluation;
3. Sophisticated statistical methods, such as combined experiment design methods, multidimensional regression and correlation analysis, multifactor analysis, multidimensional statistical methods, time series analysis, or the methods of operational research.

THE “R” PROJECT AND STATISTICAL METHODS OF QUALITY CONTROL IN “R”

Computers loaded with appropriate statistical software are used nowadays to analyze and evaluate data. In addition to professional commercial statistical software (e.g. STATISTICA, SAS, or SPSS), there are many other programs that contain basic statistical functions. To this category belongs, for instance, mathematical software (MAPEL, MATHEMATICA), EXCEL spreadsheet software, or OriginLab software for graphical data processing. To some extent, it is possible to substitute open source statistical software like “R” project for these programs.

“R” is a language and an integrated environment for data analysis, statistical and mathematical computing, and graphical data processing and representation. This free software is very similar to the S language and runs on most UNIX/Linux platforms, Macintosh and Windows. “R” belongs to open source software and its latest version can be downloaded directly from its home page at <http://www.r-project.org/>.

“R” is a tool for accomplishing many conventional as well as modern statistical computing and analytical tasks. The basic environment contains several standard, recommended packages, and many more are available through the CRAN archive on the home page, including updates.

“R” includes a wide variety of statistical quality control techniques. The basic “stats” package, implemented directly in the main menu, can be extended with other statistical quality control packages from the CRAN archive on the home page, such as “qcc” or “qualityTools”.

PACKAGE “STATS”

The basic statistical “stats” package is one of the default “R” packages. It enables computation of numerical data characteristics. Table 1 lists basic functions (commands) for numerical characteristics of a statistical ensemble.

Table1. Computation of basic numerical characteristics within the “stats” package

Function	Characteristics	Function	Characteristics
<code>mean()</code>	Arithmetic mean	<code>length()</code>	Ensemble size
<code>median()</code>	Median	<code>var()</code>	Sample variance
<code>mode()</code>	Mode	<code>sd()</code>	Sample standard deviation
<code>max()</code>	Maximum value	<code>quantile()</code>	Quantiles
<code>min()</code>	Minimum value	<code>IQR()</code>	Interquartile range
<code>summary()</code>	Summary	<code>range()</code>	Value range (min., max.)

Beside descriptive statistics, the package also offers hypothesis testing, graphical representation of results, and many other statistical methods.

	A	B
1	data	sample
2	4,734	1
3	4,729	1
4	4,724	1
5	4,72	1
6	4,729	1
7	4,718	2
8	4,728	2
9	4,723	2
10	4,714	2
11	4,723	2
12	4,734	3

Figure 1. Data

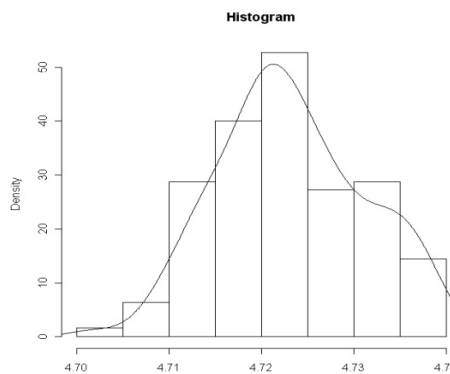


Figure 2. Histogram

To demonstrate the package features, we are going to analyze measurements of the diameter [mm] of a component of a 3D BluRay player drive. We open a new EXCEL file and divide the measured data into 25 subgroups, 5 readings per each subgroup. We label the data to be analyzed “diameter” and save the file as `data.csv`.

The process analysis starts with loading the prepared data file:

```
> table=read.table("data.csv",header=T,sep=";",dec=".",")
```

Arithmetic mean and sample variance of the “data” file will be computed as follows:

```
>mean(table$data)
```

```
[1] 4.723248
```

```
> var(table$data)
```

```
[1] 6.20751e-05
```

In terms of graphical representation, the “stats” package offers a boxplot, a Q-Q plot, a histogram, a pie chart, and many other types of charts.

```
>hist(table$data,nclass=7,prob=T,xlab="Data",main="Histogram");
lines(density(table$data))
```

The `prob=T` argument represents a probability scale; `prob=F` means that absolute frequencies are allowed for; `nclass=5` determines the number of classes.

The `lines(density(x))` function is a probability density function.

```
>boxplot(table$data,main="Boxplot Data")
```

```
>qqnorm(table$data,main="Normal Q-Q Plot");qqline(table$data)
```

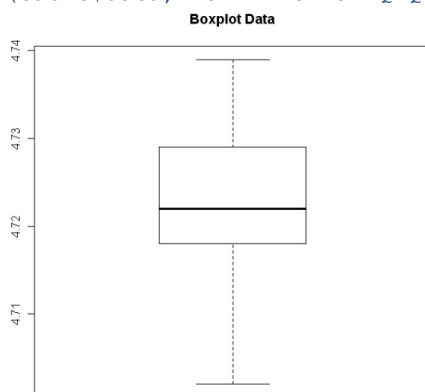


Figure 3. Boxplot

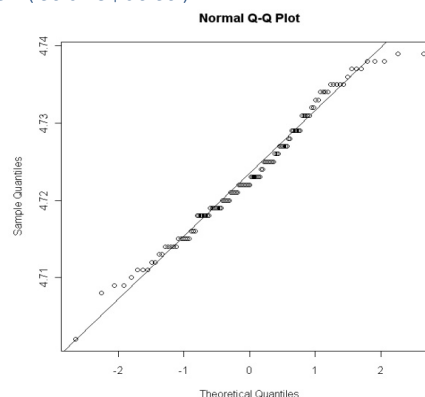


Figure 4. Normal Q-Q Plot

Normal distribution is the most frequent in industrial practice. The Shapiro-Wilk normality test (`shapiro.test()`), included among the basic functions of the “stats” package, can be used for normality testing.

```
>shapiro.test(table$data)
Shapiro-Wilk normality test
data:  table$data
W = 0.9826, p-value = 0.1089
```

PACKAGE “qcc”

The latest version 2.2 offers a multitude of functions for computing and plotting of basic quality control charts, CUSUM and EWMA charts, process capability analysis, Pareto analysis, etc.

The `qcc()` command can be used to create basic types of variable control charts. In the simplest case, the command can read as follows: `qcc(data,type=“ ”)`. Control limits are usually set as at a distance of 3σ from the centerline. This default setting can be changed by the `nsigmas` command, e.g. `qcc(data, type=“ ”,nsigmas=2)`.

```
>Data=qcc.groups(table$data,
table$sample)
>Data
      [,1] [,2] [,3] [,4] [,5]
1  4.734 4.729 4.724 4.720 4.729
2  4.718 4.728 4.723 4.714 4.723
3  4.734 4.737 4.727 4.723 4.725
4  4.738 4.713 4.726 4.718 4.725
5  4.725 4.720 4.711 4.715 4.727
```

```
>objR=qcc(Data,type="R")//R bar chart
```

Centerline $CL=0.0132$, lower control limit $LCL=0$, and upper control limit $UCL=0.0356$. The R chart (Figure 5) shows that no value is outside the control limits. We can assume that the variability of the monitored process is under statistical control. To get a control chart for the diameter (X bar chart), we need to change the type of chart, e.g. `type=“xbar”` (Table 2).

```
>objX=qcc(Data,type="xbar")//X bar chart
```

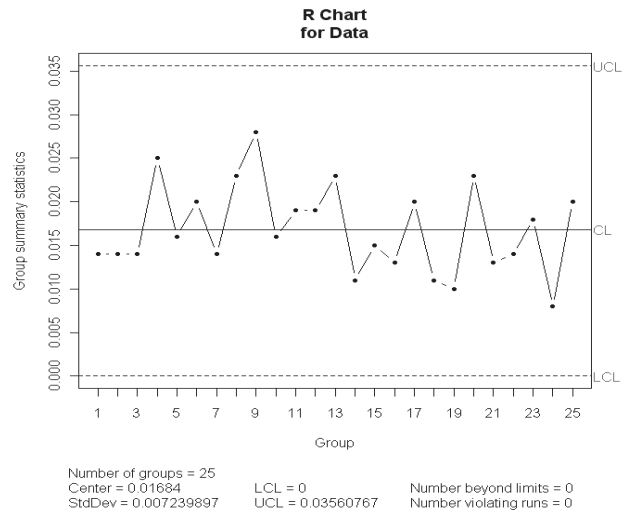


Table 2. Types of variable control charts

Type	Control Chart	Type	Control Chart
“xbar”	X chart	“S”	S chart
“R”	R chart	“xbar.one”	X chart for one-at-time data

Attribute control charts can be created by changing the chart type (`type=“np”`, or “c”, “u”, “p”).

A lower specification limit (LSL), an upper specification limit (USL), and a target value (T), given by customer requirements, must be known in order to perform a process capability analysis. Process capability indices are computed by the `process.capability()` function. In our example, the lower specification limit of $LSL=4.70$ mm and the upper specification limit of $USL=4.74$ mm were defined by the `spec.limits()` command.

```
> process.capability(objX,spec.limits=c(4.7,4.74))
Process Capability Analysis
Call:
process.capability(object = objX, spec.limits = c(4.7, 4.74))
Number of obs = 125          Target = 4.72
      Center = 4.723248      LSL = 4.7
      StdDev = 0.007239897   USL = 4.74

Capability indices:
Value  2.5%  97.5%
Cp      0.9208  0.8063  1.0352
Cp_l    1.0704  0.9483  1.1924
Cp_u    0.7713  0.6770  0.8656
Cp_k    0.7713  0.6589  0.8837
Cpm     0.8402  0.7277  0.9524
Exp<LSL 0.066%  Obs<LSL 0%
Exp>USL 1%    Obs>USL 0%
```

The program output not only contains the values of individual capability indices, but also their 97.5 percent reliability intervals. A target value of $T=4.72$ mm, representing the midpoint of the interval defined by USL and LSL , was used to determine the capability index C_{pm} . This value can be specified according to customer requirements using the “target=” function.

The graphical output is a histogram with the computed capability indices (Figure 6).

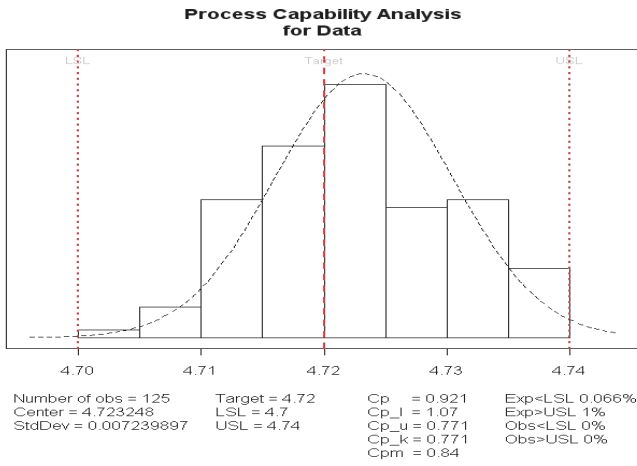


Figure 6. Process Capability Analysis

The older version 2.0 also includes the `process.capability.sixpack()` function for performing a final process capability analysis (version 2.2 lacks this function).

```
>process.capability.sixpack(objX, spec.limits=c(4.70,4.74))
```

Other functions of the “qcc” package are listed in Table 3.

In the next step, we are going to perform a Pareto analysis and identify the ultimate causes of poor quality. A certain type of product exhibited six types of quality defects during inspection. The variety and frequency of these defects is shown in the table 4.

Table 3. Other control chart functions

Type	Function	Type	Function
<code>cusum()</code>	CUSUM chart	<code>oc.curves()</code>	Operating characteristic curves
<code>ewma()</code>	EWMA chart	<code>pareto.chart()</code>	Pareto chart
<code>shewhart.rules</code>	Functions specifying rules for Shewhart charts		

Table 4. Defects and their frequency during quality inspection of a certain product

Defect	Frequency	Defect	Frequency
A Scratches	41	D Incomplete product	98
B Damaged packaging	115	E No operating manual	20
C Malfunction	72	F Other defect	8

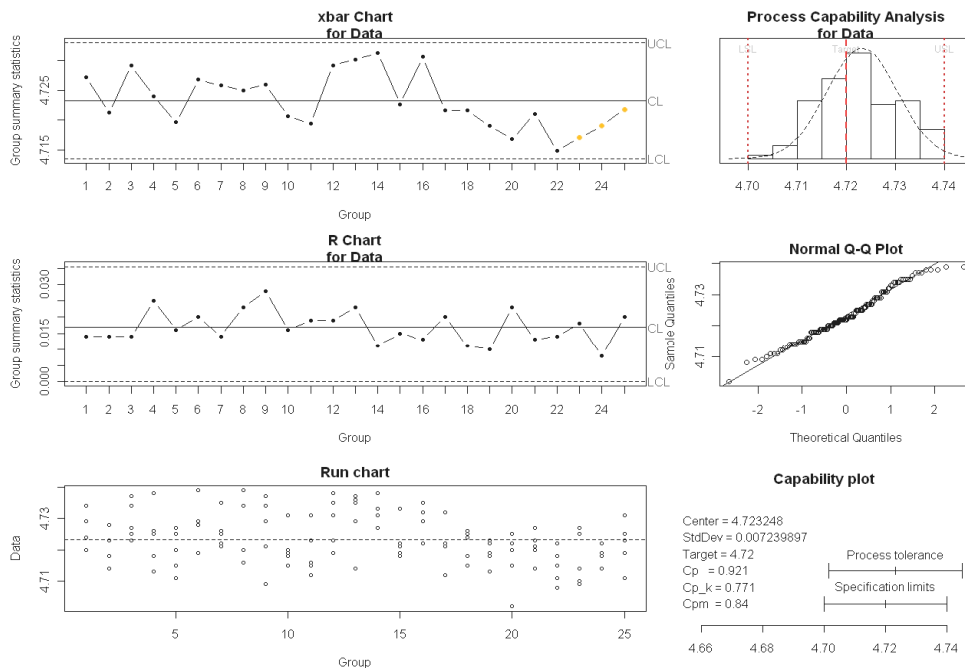


Figure 7. Process Capability Analysis

Pareto analysis for defect in R:

```
> defect=c(41,115,72,98,20,8)
> names(defect)=c("A", "B", "C", "D", "E", "F")
> pareto.chart(defect)
```

Pareto chart analysis for defect

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
B	115.000000	115.000000	32.485876	32.485876
D	98.000000	213.000000	27.683616	60.169492
C	72.000000	285.000000	20.338983	80.508475
A	41.000000	326.000000	11.581921	92.090395
E	20.000000	346.000000	5.649718	97.740113
F	8.000000	354.000000	2.259887	100.000000

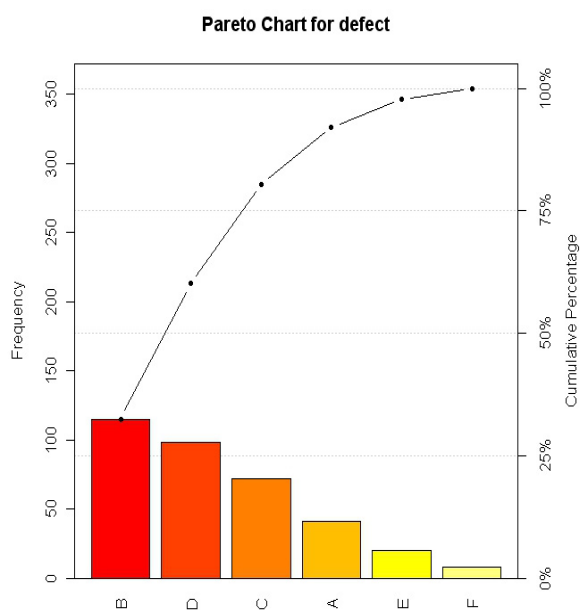


Figure 8. Pareto Analysis

The result of the analysis is a frequency chart, the graphical output being a Pareto chart with a Lorenz curve. According to the analysis, the root causes of poor quality are defects B, D and C.

PACKAGE “QUALITYTOOLS“

The “qualityTools“ package contains methods connected with the DMAIC model (Define, Measure, Analyze, Improve, Control) of the Six Sigma method. Among the tools included within this package are: process capability analysis in case of normal or other than normal distribution, measurement system analysis (repeatability, reproducibility), Pareto analysis, distribution functions, experiment design, etc.

Measurement capability analysis based on variation analysis is a part of the package. Let’s take three randomly selected operators and ten randomly selected samples, each of the operators measuring each sample twice. The measurement system capability can be analyzed using the `gageRRDesign()`, `gageRR()` and `plot()` functions.

```
>library(qualityTools)\\ installation package qualityTools
>design = gageRRDesign(Operators=3, Parts=10, Measurements=2, randomize=FALSE)
>response (design)=c(10.96,10.99,10.94,10.90,10.89,10.85,10.67,10.68,10.71,10.35,10.42,
10.36,10.71,10.73,10.73,10.82,10.80,10.75,10.55,10.54,10.48,10.65,10.64,10.65,10.46,10
.45,10.42,10.55,10.57,10.55,10.97,10.96,10.91,10.89,10.93,10.84,10.68,10.74,10.64,10.4
0,10.39,10.33,10.68,10.72,10.65,10.81,10.84,10.76,10.56,10.60,10.49,10.63,10.69,10.66,
10.47,10.43,10.45,10.57,10.54,10.53)
>analysis = gageRR(design)
AnOVA Table - crossedDesign
      Df SumSq MeanSq      F valuePr(>F)
Operator  2 0.01886 0.009432  15.8515 2.005e-05 ***
Part      9 1.87130 0.207922 349.4494 < 2.2e-16 ***
Operator:Part 18 0.01080 0.000600  1.0087  0.4778
Residuals 30 0.01785 0.000595
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
-----
AnOVA Table WithoutInteraction - crossedDesign
      Df SumSq MeanSq      F valuePr(>F)
Operator  2 0.01886 0.009432  15.80 5.344e-06 ***
Part      9 1.87130 0.207922 348.31 < 2.2e-16 ***
Residuals 48 0.02865 0.000597
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
-----

Gage R&R
      VarComp VarCompContrib Stdev StudyVar StudyVarContrib
totalRR      0.001039      0.0292 0.0322      0.193      0.171
repeatability 0.000597      0.0168 0.0244      0.147      0.130
reproducibility 0.000442      0.0124 0.0210      0.126      0.111
Operator      0.000442      0.0124 0.0210      0.126      0.111
Operator:Part 0.000000      0.0000 0.0000      0.000      0.000
Part to Part  0.034554      0.9708 0.1859      1.115      0.985
totalVar      0.035593      1.0000 0.1887      1.132      1.000
```

Figure 9: Output of Gage R&R

It follows from the first ANOVA chart results that the interaction between the operators and the samples is insignificant (p -value=0.4778). Since the interaction is insignificant, modeling will be repeated without interactions. Based on a study, the standard deviation of the process was estimated at 0.193 (Figure 9). Next results indicate that repeatability and reproducibility (R&R) represent 2.92 percent of the total variance caused by the measurement process, where 1.68 percent is repeatability (equipment variability) and 1.24 percent reproducibility (operator variability). The remaining 97.08 percent represents part to part variability. The graphical output is shown in Figure 10.

```
> plot (analysis)
```

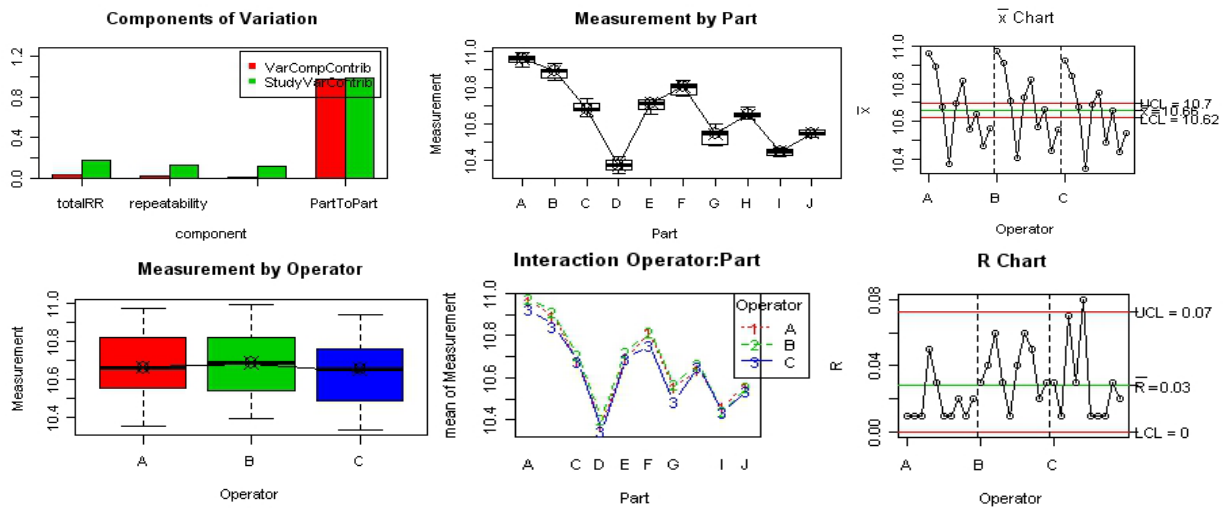



Figure 10: Graphical output of Gage R&R

OTHER PACKAGES

There are also other packages for process quality assessment through statistical methods that can be installed additionally from the CRAN archive, which currently holds about 500 libraries, containing also extension (special) functions:

- Hmisc, pastecs, psych – descriptive statistics;
- Outliers – testing of extreme values and outliers;
- fBasics, nortest – normality testing;
- lmtest – linear regression;
- tseries, forecast – time series analysis;
- IQCC, QCnalyt – statistical quality control, Shewhart control charts, reliability indices, etc.

CONCLUSIONS

In today's modern society, the importance of the word "quality" is markedly increasing, whether it is product quality from customers' perspective or from the perspective of producers, distributors, service providers, etc. Quality is becoming a very important tool for keeping the customers satisfied and maintaining the competitiveness of producers and organizations. Statistical methods and tools are a powerful quality control means, and their correct application represents a very important aspect of all process control phases. The result of their application is an analysis that forms a basis for continuous improvement as well as adopting process control measures.

Computers loaded with appropriate statistical software are used nowadays in statistical quality control. Users may find the open-source R software complicated at first, but the detailed manuals and various discussion forums contain a lot of useful information that even beginners can easily look up to grasp the basic principles of R. A big advantage of R is its neither financially nor time limited accessibility to general public. In many cases, R can be considered a satisfactory replacement for professional commercial statistical software.

REFERENCES

- [1.] Hrubec, J., Virčíková, E. a kol.: Integrated Management System (in Slovak). Nitra SPU, Nitra 2009, 543 p.
- [2.] Nenádl, J., Noskievičová, D., Petříková, R., Plura, J., Tošenovský, J.: Modern quality management (in Czech). Praha, Management Press Praha, 2008, 377 p.
- [3.] Roth, T.: Package „qualityTools“. In: <http://cran.r-project.org/web/packages/qualityTools/qualityTools.pdf>
- [4.] Scrucca, L.: Package “qcc“. In: <http://cran.r-project.org/web/packages/qcc/qcc.pdf>.
- [5.] Verzani, J.: Using R for Introductory statistics. In: <http://www.math.csi.cuny.edu>



ANNALS OF FACULTY ENGINEERING HUNEDOARA



– INTERNATIONAL JOURNAL OF ENGINEERING



copyright © UNIVERSITY POLITEHNICA TIMISOARA,
FACULTY OF ENGINEERING HUNEDOARA,
5, REVOLUTIEI, 331128, HUNEDOARA, ROMANIA
<http://annals.fih.upt.ro>