Testing of Normality of Data Files for Application of SPC Tools

PAVEL MACH, HANA HOCHLOVÁ
Department of Electrotechnology, Faculty of Electrical Engineering, Czech Technical University in Prague
Technická 2, 166 27 Prague 6, Czech Republic
Phone (++420) 224 352 122, Fax (++420) 224 353 949, E-mail: mach@fel.cvut.cz

Abstract

Different types of statistical tests have been used for evaluation of normality of selected data files. The files have been simulated to be on the border of normality. The results of the tests have been compared. It has been found that some files have been, using some type of a test, marked as normal, but using another type of a test, marked as non-normal. It is possible to meat similar results by normality testing of data files obtained from production. Therefore complementary tools for assessment of normality of the files under test have to be used.

The paper compares results of normality testing of some simulated files by three basic tests of normality: Test of elective skewness and kurtosis, Anderson – Darling test and Kolmogorov – Smirnov test. It gives also directions which other tools of exploratory analysis have to be used for verification of results obtained by different tests of normality.

Keywords: normality test, Test of elective skewness and kurtosis, Anderson – Darling test, Kolmogorov – Smirnov test, exploratory analysis, SPC.

Introduction

SPC (Statistical Production Control) is represented by different types of statistical tools, which are used for improvement of quality of production [1]. Chart diagrams, histograms, frequency tables, Pareto diagrams and others can be involved among these tools. The basic research in the field of SPC has been done by Ishikawa, who has developed "Six Ishikawa tools". These basic tools have been completed by many others and different other ways have been also successfully used (e.g. Six sigma). The high interest for this area of research follows of the fact that it is not possible to realize successful production based on "high technology processes" without sophisticated quality control. This rule is valid especially in electronic production, but in many others, too.

SPC tools are usable not for control of quality only, but they make evaluation of capability of a fabrication process to fulfil requests of customers possible and also give a possibility of economical checking of production. This information is very significant for manufacturers.

The basic assumption for the successful use of different types of SPC tools is that the data files have normal distribution. Many tests of normality have been developed for normality testing. However, there are two questions joined with this testing:

- What is necessary to do with the files, which have not normal distribution to transform them to normality (with the goal the use of some type of SPC tool for their processing)
- Conclusions given by different types of the tests of normality can differ – which way can be done the right conclusion?

The files, which have not passed through the test of normality, are transformed to normality usually [2]. There are used different types of transformations, which are mostly very efficient. After the transformation appropriate SPC tool is applied and conclusions are made over the transformed file. At the end have to be these conclusions (e.g. about the limits of the chart diagram) transformed again into beginning variables.

A problem is, when conclusions given by different types of normality tests differ. It is recommended to make a final decision about normality of a file under test after the use of more tests of normality or some other tools of exploratory analysis.

The paper is focused on comparison of results of three very often used tests of normality. Data files, which have been "on the border" of normality, have been tested. The files have been simulated. There are also shown some selected tools of exploratory analysis, which can be used for evaluation of normality of data files.

1. Normal distribution

Normality of data files is a basic assumption for the use of SPC tools. It is defined by the density of probability

$$f = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1)$$

0-7803-8422-9/04/$20.00 c2004 IEEE

Where μ ... mean value, σ ... standard deviation.

Parameters that describe shape of the distribution are also skewness (marked usually g_1 ... third central moment) and kurtosis (marked usually g_2 ... fourth central moment).

$$g_1 = \frac{\sqrt{n\sum_{i=1}^{n}(x_i-\mu)^3}}{\left[\sum_{i=1}^{n}(x_i-\mu)^2\right]^{\frac{3}{2}}} \tag{2}$$

$$g_2 = \frac{n\sum_{i=1}^{n}(x_i-\mu)^4}{\left[\sum_{i=1}^{n}(x_i-\mu)^2\right]^2} \tag{3}$$

Skewness of the normal distribution is 0 and the kurtosis is 3.

2. Basic statistical tests for testing of normality

2.1 Test based on evaluation of sampling skewness and kurtosis

This test is based on testing criterion given by the equation

$$C_1 = \frac{g_1^2}{D(g_1)} + \frac{[g_2 - E(g_2)]^2}{D(g_2)} \tag{4}$$

Here g_1, $D(g_1)$ ... sampling skewness and its dispersion, g_2, $D(g_2)$, $E(g_2)$ ... sampling kurtosis and its dispersion and mean value. Calculations of dispersions and mean value are carried out according to the formulas

$$D(g_1) = \frac{6(n-2)}{(n+1)(n+3)} \tag{5}$$

$$D(g_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} \tag{6}$$

$$E(g_2) = 3 - \frac{6}{n+1} \tag{7}$$

If the data have normal distribution, parameter C_1 has the X^2 (2) distribution. Therefore the calculated value C_1 has to be compared with a quantil $X_{1-\alpha}^2(2)$. If

$$C_1 \rangle X_{1-\alpha}^2, \tag{8}$$

a data file is not normally distributed (α is level of significance, usually is chosen 1 %, 5 % or 10 %).

2.2 Anderson-Darling test

Testing criterion for this type of the test is given by equation

$$AD = \frac{\sum_{i=1}^{n}(2i-1)(\ln Z_i + \ln(1-Z_{n-i+1}))}{n} - n \tag{9}$$

Parameter

$$Z_i = \phi\frac{x_i-\mu}{s} \tag{10}$$

is a value of a distribution function of normal distribution, s ... standard deviation, which is calculated of the equation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{11}$$

Calculated value of the parameter D is compared with the quantile $D_{1-\alpha}$. If

$$AD \rangle D_{1-\alpha} \tag{12}$$

then a hypothesis about normality of data is neglected. Values $D_{1-\alpha}$ can be calculated, for level of significance 0.05, according to the formula

$$D_{0.95} = 1.0348\left(1 - \frac{1.013}{n} - \frac{0.93}{n^2}\right) \tag{13}$$

2.3 Kolmogorov-Smirnov test

This test is based on coincidence of two distributions. The distribution function $S_n(x)$ of a sampling distribution is compared with the distribution function F(x) of the normal distribution. Maximum difference between these two distribution functions is investigated.

$$D = \max|S_n(x) - F(x)| \qquad \text{for} \quad -\infty < x < \infty \tag{14}$$

The distribution function $S_n(x)$ is calculated of the formula

$$S_n(x) = \frac{M}{n} \tag{15}$$

Where M ... number of values X_1, X_2, ... , X_n, which are $\leq x$.

Calculated values of D are compared with a critical value of D (critical values of D for Kolmogorov-Smirnov test are involved in many handbooks

of statistical tables). If the calculated value is higher than the value taken of the table, then the hypothesis about normality of the data file under test has to be rejected.

3. Exploratory analysis

The goal of exploratory analysis is a diagnostics of basic assumptions, which data have to fulfill for the use in majority of statistical methods [3]. Instead normality following properties are investigated: constant dispersion and mean value, homogeneity, outliers, independence.

The most often used graphic tools are: a histogram, a box plot, a Q-Q plot, and graphical expression of density of probability.

Histogram is a very good known tool for investigation of the type of data distribution. Its shape gives a good possibility for estimation of the type of distribution. The shape of the histogram depends significantly on the number of classes. Standardized number of classes is from 7 to 20. The values of a random variable are collected is subgroups according their values. Every subgroup has lower and upper limit of values. The number of the classes is calculated using the formula

$$c = \frac{A_{max} - A_{min}}{r} \qquad (16)$$

Here c ... number of classes, A_{max}, A_{min} ... maximal and minimal value of the data file, r ... numbers 2 or 5. The difference A_{max}-A_{min} is divided by 2 or 5 or by 0.2, 0.5, 20, 50 etc. to obtain the number of classes between 7 and 20. If two numbers of classes inside this interval are possible for the use, the higher number is used if the data file has more than 100 values and vice versa.

A very significant tool is a Q-Q plot. This graph gives information about normality, symmetry, skewness, kurtosis, homogeneity of sampling and outliers. Horizontal axis of this graph is the axis of theoretical values of quantiles of the normal distribution; vertical axis is the axis of sampling quantiles of an examined distribution. In the case that investigated distribution would be normal one, the Q-Q plot will be a straight line. An example of a Q-Q plot is shown in Fig. 1.

A box plot and a graph of density of probability are very good known tools, and it is possible to find them in many statistical handbooks [4], [5].

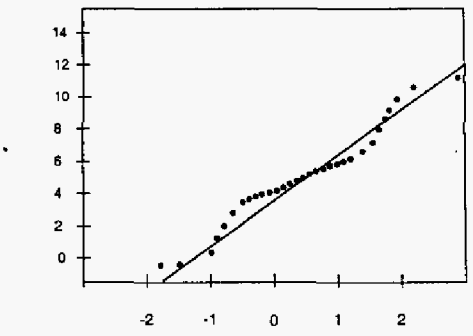

Fig. 1 An example of a Q-Q plot

4. Area of normality of data

Typical parameters of the normal distribution are mean value, standard deviation, skewness and kurtosis. The value of skewness of the normal distribution is 0, the value of kurtosis is 3. However, it has been found that the test based on evaluation of sampling skewness and kurtosis marks as normal distributed also files with different skewness and kurtosis.

Therefore area of normality of data with respect to the skewness and kurtosis has been investigated. The goal of the work has been to find such values of sampling skewness g_1 and kurtosis g_2 to obtain resulting value of criterion C_1 (see equation 4) equal or lower than the quantile $X^2_{(0.95)}$ (see equation 8). Its value is 5.991465. Pairs of values of g_1 and g_2, which fulfil the condition $C_1=X^2_{(0.95)}$ create border of normality of data.

The calculations have been carried out in Excel by the use of a tool "Solver". The result is shown in Fig. 2.

5. Comparison of basic statistical tests of normality

Three types of tests of normality have been compared – The test based on evaluation of sampling skewness and kurtosis (J-B test), Anderson-Darling test (A-D test) and Kolmogorov-Smirnov test (K-S test). Different types of data files have been simulated for testing: data with the normal distribution, data with other than normal distribution and data on the border of normality. The data files under test have number of values 50, 100, 200, 500 or 1000.

The results of checking of normality of data files with 500 values are presented in Tab. 1. It has been found that the results are very different. The results of the tests are the same for normally distributed data only. For data files with other than normal distribu-

tion the results are significantly different and also differences have been found in the results of the tests for data on the border of normality.

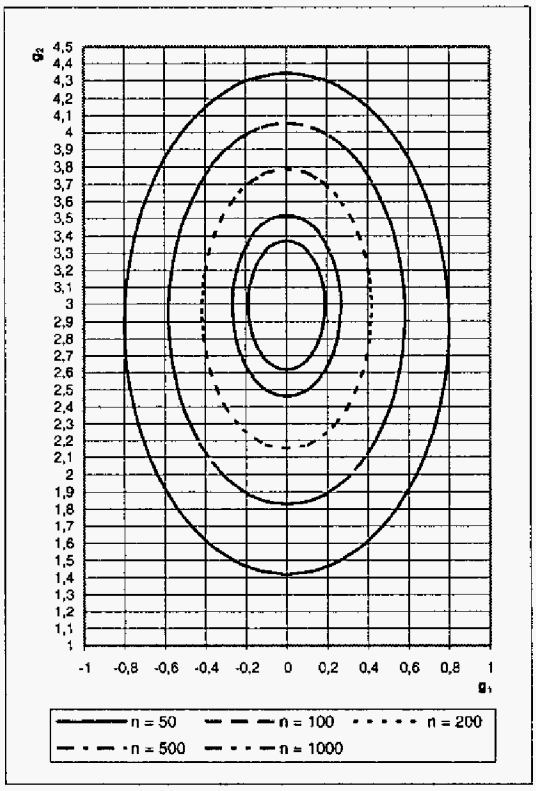

Fig. 2. Borders of normality for data files with different sizes of random samplings

TABLE 1. Comparison of results of different tests of normality. The tested data files have had 500 values

J-B test		A-D test		K-S test	
C1	Norm	AD	Norm	C1	Norm
Data with normal distribution					
3.607	Y	0.994	Y	0.141	Y
5.208	Y	1.007	Y	0.114	Y
4.107	Y	1.054	Y	0.129	Y
Data with other than normal distribution					
30.689	N	2.096	N	0.132	Y
834.378	N	3.209	N	0.19	N
7.167	N	0.913	Y	0.112	Y
12.887	N	0.757	Y	0.11	Y
49.231	N	0.969	Y	0.128	Y
Data on the border of normality					
5.99	Y	1.979	N	0.161	Y
5.976	Y	1.045	N	0.112	Y
5.994	N	0.898	Y	0.141	Y

Conclusions

Testing of normality of the same data files using different tests of normality can give different results (see Tab. 1). Therefore it is necessary to use more tests of normality for normality checking and to complete these tests by some other tools of exploratory analysis. Such the tool can be e.g. a Q-Q diagram, but histogram or plot of density of probability can also give very useful information.

According to our experience, the most useful test of normality seems to be the J-B test. Using this test you will achieve the best correlation with the other tools of the exploratory analysis. On the other hand, the other types of normality tests can be successfully used for verification of conclusions obtained by the use of J-B test.

References

1. MESSINA S. WILLIAM, *Statistical Quality Control for Manufacturing Managers*, New York: J. Wiley&Sons, Inc., 1987
2. MONTGOMERY, D.C, *Introduction to Statistical Quality Control*, Willey and Sons, 2000
3. MONTGOMERY, D.C, *Design and analysis of experiments*, Willey and Sons, 2000
4. Server http://deming.eng.clemson.edu/pub/tutorials/qctools/ccmain1.htm
5. WISE, S.A., FAIR, D.C., Innovative Control Charting: Practical SPC Solutions for Today's Manufacturing Environment. ASQ, 1997