Transformation of Data for Statistical Processing

Pavel Mach, Josef Thuring, David Šámal

Department of Electrotechnology, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
mach@fel.cvut.cz

Abstract: *The use of many statistical tools depends on normality of processed data. There are different methods for transformation of non-normally distributed data sets toward to normally distributed ones. The goal of the work has been to investigate usability of four types of transformations (Box-Cox, exponential, power and logarithmic) for transformation of data sets with four non-normal distributions (logarithmic-normal, exponential, gamma, and Weibull) toward to normally distributed data. The usability and efficiency of individual transformation functions for transformation of data sets with different types of distributions have been found.*

1. Introduction

The use of SPC (Statistical Production Control) tools as well as many other statistical tools depends strongly on normality of processed data [1], [6]. If distribution of data is not normal, the use of many statistical tools is not possible because false results are obtained.

Our research is focused on investigation of properties of electrically conductive adhesives. It has been necessary, for analysis of correlation among changes of electrical and mechanical properties of joints fabricated of different types of electrically conductive adhesives, to check normality of measured data, to delete outliers (if they have been found), and, when the data have not been normally distributed, to transform them into normality. Efficiency of transformation depends strongly on selection of a proper type of a transformation function. Application of a central limit theorem instead transformation is also possible; however, if the grouping is of a higher order, the total number of processed values rapidly decreases. Therefore the use of this theorem seems to be, especially in cases, when limited volume of data is measured, disadvantageous.

Therefore the research has been focused on examination of efficiency of different types of transformation functions for transformation of data sets with different types of distributions different from the normal distribution, toward to normally distributed data sets.

2. Basic types of distributions

Data have been transformed toward to normally distributed ones. Probability density of normal distribution N (μ, σ) is described by the equation [2]:

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

Where μ … mean value, σ … standard deviation. Normal distribution is typical for the data, which are measured at the output of good stabilized fabrication processes, for the data obtained by repeated measurements, which are disturbed by random noise [3], [4].

Efficiency of transformation functions have been tested on data sets with following types of distributions:

Logarithmic-normal distribution LN(μ, σ):

$$\mathrm{f}(\ln \mathrm{x}) = \frac{1}{\sigma(\ln x)\cdot\sqrt{2\pi}} \cdot \exp\left(-\frac{(\ln x - E(\ln x))^2}{2\sigma^2(\ln x)}\right) \qquad (2)$$

Where E(ln x) … mean value of ln x, σ(ln x) … standard deviation of ln x. This distribution is used for effects, which have positive values only, e.g. for the length of a biologic life, for life time of apparatuses.

Exponencial distribution Ex (λ):

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$
$$f(x) = 0 \quad \text{for } x \lhd 0 \tag{3}$$

This distribution is used for description of lengths of intervals between adjacent drop-outs of systems and this distribution describes the life-time of equipment as well.

Gamma distribution (α, β):

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} \quad x \geq 0$$
$$f(x) = 0 \quad x < 0 \tag{4}$$

Where α, β are continuous parameters.

Weibull distribution W(α, β):

$$f(x) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(\frac{x}{\beta})^{\alpha}} \tag{5}$$

Where α, β are continuous parameters.

Weibull distribution can be modified, by the use of proper values of parameters, in other types of distributions [5]. Weibull distribution is used, instead others applications, for description of the life-time of electron tubes.

3. Transformation functions under test

Nonlinear transformation of data is a very efficient method for reduction of their asymmetry and modification of their distribution toward to the normal one. The transformation function has to be nonlinear. This function must also be monotonous, to be not changed order of data after their transformation. This property is described by the condition:

If $x_i < x_j$, then $x_i^* < x_j^*$ (6)

Where x_i, x_j … data before transformation, x_i^*,x_j^* … data after transformation. Principle of nonlinear transformation is shown in Fig. 1

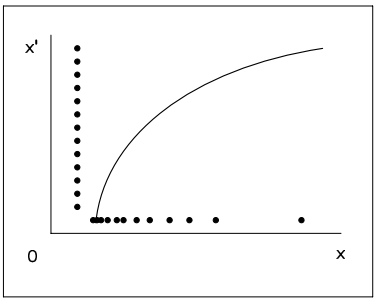

Fig. 1 Principle of nonlinear transformation

The work with a data set starts with testing of its normality usually. If it is found that the processed data are not normally distributed, they must be transformed toward to the normal ones. The transformation procedure consists of following steps:

- It is chosen a proper type of the transformation function.
- Data are transformed.
- Normality of data is tested using a proper normality test together with some tool of exploratory analysis (usually Q-Q graph and density plot are used).
- If the data are normally distributed, requested operations (usable for normally distributed data) are carried out.
- Using retransformation are the results, which have been found over normally distributed data, retransformed into original data. The retransformation is carried out using function inverse to the transformation function.

Following types of transformations have been tested:

Box-Cox transformation

This type of transformation is the most commonly used type of a power transformation. It is described by the equation:

$$F(x) = \begin{cases} \dfrac{x^r - 1}{r} & \text{for r = 0} \\ \ln r & \text{for r} \neq 0 \end{cases} \qquad (7)$$

The course of the Box-Cox function in dependence on the value of the parameter *r* is shown in Fig. 2.

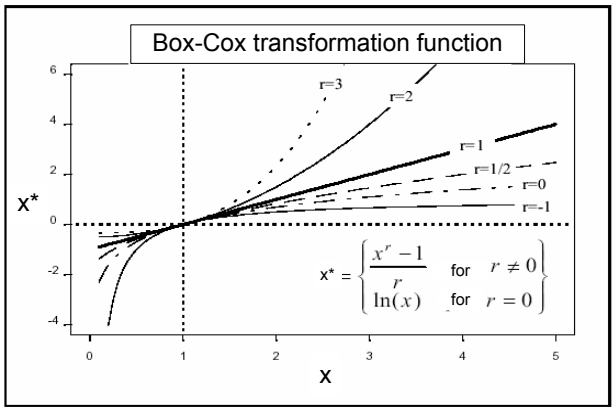

Fig. 2 Box-Cox transformation function

Exponential transformation

Exponential transformation is also used very often, especially when the data have distribution, which is near to lognormal distribution. The transformation function is as follows:

$$F(x) = e^{x \cdot r} \qquad (8)$$

The course of this function strongly depends on the value of the parameter r.

Simple power transformation

The simple power transformation is described by the formula:

$$F(x) = x^r \qquad (9)$$

This transformation is actually simply, but it gives results of sufficient quality in many cases.

Logarithmic transformation

Logatithmic transformation is based on the function:

$$F(x) = \ln(x \cdot r) \qquad (10)$$

Transformation functions presented in equations (8) to (10) are written in the simplest form. More complex functions, with more parameters, e.g. m-degree polynomial function, are also used. On the one hand, the more complicated is the transformation function, the higher quality of transformation can be achieved. On the other hand, the more complicated is the transformation function, the more difficult is finding of optimum parameters for the best transformation.

4. Testing of normality

Quality of transformation has been verified by testing of normality of transformed data. There are many types of normality tests, which are used. We have used the "*Test of combination of sample skewness and kurtosis*". This test seems to be a very strong and it is usable also in the cases, when the other tests fail. The testing criterion C_1 for this test is as follows:

$$C_1 = \frac{\hat{g}_1^{\,2}}{D(\hat{g}_1)} + \frac{[\hat{g}_2 - E(\hat{g}_2)]^2}{D(\hat{g}_2)} \qquad (11)$$

Where the sample skewness g_1 and its variance $D(g_1)$, and sample kurtosis g_2, its mean value $E(g_2)$, and its variance $D(g_2)$ are calculated using formulas, which can be found in [7]. If the data are normally distributed, the variable C_1 has distribution of the type $\chi^2(2)$.

If

$$C_1 > \chi_{1-\alpha}^{\,2}(2) \quad , \qquad (12)$$

hypothesis about normality of the data has to be rejected.

The normality tests have to be completed, when normality is tested, by some tools of exploratory analysis. The test informs about the fact if a data set is or is not normally distributed only. Test does not give information about the reason of non-normality, if it is caused by limited number of outliers only, or, if the data set is not normal in principle. Therefore the normality test has to be completed with some tool of exploratory analysis.

Exploratory analysis is a group of graphical methods for investigation of normality of data. The

most frequent used methods are a Probability density plot and a Q-Q plot.

The Probability density plot is based on drawing of probability density curve of an investigated data set and probability density curve of normal distribution. It shows, on the first view, which is difference of investigated data curve from the normal curve. The probability density plot is shown in Fig. 3.

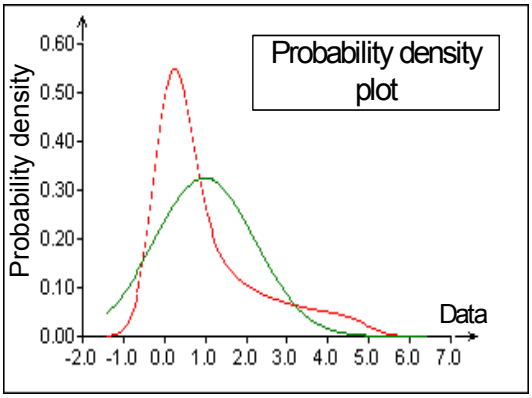

Fig. 3 Probability density plot. The probability density of investigated data set is dashed, the probability density of normal distribution is drawn by solid line

The Q-Q (Quantile-Quantile) Plot is a graphical tool for investigation if two data sets come of populations with a common distribution. There are drawn quantiles of normal distribution on the horizontal axis and quantiles of an investigated data set on the vertical axis.

A 45-degree reference line is also plotted. If the investigated data set comes from a population with the normal distribution, the points should fall approximately on this reference line. The higher is the distance of the points from this reference line, the higher is deviation of investigated data set from normality.

The use of the Q-Q plot is advantageous, because its drawing is very simple and this diagram affords information about outliers, about the level of difference between distribution of investigated data and normal distribution, and about the type of skewness if the distribution of the investigated data fail is not symmetrical. An example of a Q-Q plot is shown in Fig. 4.

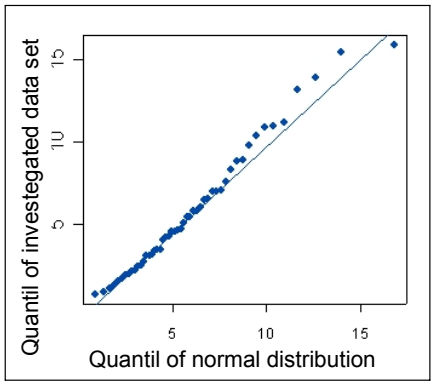

Fig. 4 Q_Q plot

5. Program for transformation

A program for application of different types of transformation functions on data sets with different types of distributions has been developed in Excel. Optimization of parameters of transformation functions has been carried out by calculation of the value of skewness in every step. This value should be 0 for normal distribution. Therefore parameters of the transformation functions have been changed in order to zero skewness would be achieved. A tool "solver" has been used for this calculation.

6. Calculations

Data sets with four distributions mentioned above, have been simulated. Every data set has been simulated of 500 values; five data sets have been created for every type of distribution.

Every of above mentioned transformation functions has been used for transformation of every data set toward to normality. Optimum parameters of the transformation functions have been found using a program described above.

Simulations of data sets with requested distributions have been carried out using the program QC Expert.

The results are presented in Tab. 1.

		Type of distribution			
		Lognormal	Exponencial	Gamma	Weibull
Type of transformation	Box-Cox	Y	Y	**Y**	**Y**
	Exponencial	Y	Y	Y	Y
	Simple power	Y	**Y**	**Y**	**Y**
	Logarithmic	**Y**	Y	**N**	**N**

Tab. 1 Results of application of different types of transformation functions on data sets with different types of non-normal distributions. The data have been transformed toward to normally distributed data.
Y in a dark cell … highly recommended
Y in a light cell … usable
N in a dark cell … unusable

7. Conclusions

Four types of transformation functions have been tested for transformation of data sets with four different non-linear distributions of data toward to normality. Proper combinations of transformation function and type of distribution have been found.

It has been also found high efficiency of the simple power transformation and of the Box-Cox transformation.

The data with requested types of distributions have been simulated and usability of transformation functions has been evaluated.

In principle, it is also possible to meet with data sets, which are not normally distributed, and which can not be transformed to normally distributed data. Statistical tools, which have been developed for non-normally distributed data, must be used for such the data sets (e.g. Hotelling diagram). The number of these statistical tools is limited.

References

[1] MESSINA, S. W. *Statistical Quality Control for Manufacturing Managers*. WILEY-INTERSCIENCE PUBLICATIONS, 1999.

[2] MONTGOMERY, Douglas C. *Introduction to Statistical Quality Control*. 4th edition, 2001, New York, N.Y, John Wiley and Sons. ISBN 0-471-31648-2.

[3] HILL, Terry. *Productions Operations Management.* 2nd edition, Prentice Hall, New Jersey 1991.

[4] FORREST, W. Breyfogle III. *Implementing SIX SIGMA. Smarter Solutions Using Statistical Methods*. 1st edition, Wiley Interscience Publications, 1999. ASIN 0471296597.

[5] LUCEY, Terry. *Quantitative Techniques*. 5th edition, 1996. DP Publications, London. ISBN 1-85805-183-5.

[6] DEMING, Edward. *Theories*.
http://www.ealtd.com/aboutus.asp?strPage=Theories

[7] MELOUN, M., MILITKÝ, J.: Statistical processing of experimental data (in Czech), Plus, 1994, Prague. ISBN 80-85297-56-6.

Acknowledgment

The work has been supported by a research project Nr. MSM6840770021 “Diagnostics of materials”.