THE BOX-COX TRANSFORMATION FOR RIGOROUS STATISTICAL ANALYSIS OF METALLURGICAL DATA

Meloun M.[1], Kupka K.[2]
[1] Department of Analytical Chemistry, Faculty of Chemical Technology, University Pardubice, 532 10 Pardubice, Czech Republic, email: milan.meloun@upce.cz
[2] Trilobyte Statistical Software Ltd., 530 02 Pardubice, Czech Republic, email: kupka@trilobyte.cz

BOX-COXOVA TRANSFORMACE PRO RIGOROZNÍ STATISTICKOU ANALÝZU METALURGICKÝCH DAT

Meloun M.[1], Kupka K.[2]
[1] Katedra analytické chemie, Fakulta chemickotechnologická, Univerzita Pardubice, 532 10 Pardubice, Česká republika, email: milan.meloun@upce.cz
[2] Trilobyte s. r. o., 530 02 Pardubice, Česká republika, email: kupka@trilobyte.cz

Abstrakt

Běžný software ke statistickému zpracování dat sice nabízí paletu odhadů rozličných parametrů polohy a rozptýlení avšak zřídka umožní vybrat danému rozdělení odpovídající parametr střední hodnoty. Jedině průzkumovou analýzou a vyšetřením předpokladů o výběru se dojde ke správnému parametru. Rigorózní chemometrická metoda, prezentovaná na analýze metalurgických dat při určení střední hodnoty a vybraných dolních kvantilů meze pevnosti konstrukční oceli ilustruje navržený postup. Je-li výběrové rozdělení sešikmené a nehomogenní s odlehlými hodnotami, musí být původní data transformována. Box-Coxova transformace zlepší symetrii rozdělení a stabilizuje rozptyl. Graf logaritmu věrohodnostní funkce umožní nalézt vhodný transformační exponent. Navržená metoda poskytuje spolehlivé odhady parametru polohy (střední hodnoty) především všude tam, kde aritmetický průměr nelze použít.

Abstract

Available software for the statistical data treatment offers a list of various point estimates of the location and spread parameters but rarely is able to help the user to choose the parameter statistically adequate to an actual sample batch. The exploratory data analysis and an examination of sample assumptions will find an answer to this question. Rigorous chemometrical methodology presented on the metallurgical data concerning an estimation of the mean value and requested lower quantiles of the tensile strength of a construction steel illustrates the proposed procedure of a statistical data treatment with the exploratory data analysis. The sample distribution is skewed. Under such circumstances the original data should be transformed. The Box-Cox transformation improves a sample symmetry and also make a stabilization of a spread. The plot of logarithm of the likelihood function enables us to find an optimal transformation parameter. Proposed procedure gives reliable estimates of location parameter for asymmetric distribution when an arithmetic mean can not be used.

Key words: Univariate data; Exploratory data analysis; Data transformation; Power transformation, Box-Cox transformation, Hines-Hines selection graph; Chemometrics; Tensile strength of construction steel,

1. Introduction

Many widely used statistical methods work best when sample distribution is symmetrical with no severe outliers. However, when an exploratory data analysis [1 - 3] indicates that the sample distribution strongly differs from the symmetrical and normal one, we are faced with the problem how to analyze a data. Raw data may require re-expression to produce an informative display, effective summary, or a straightforward analysis [4 - 11]. Difficulties may arise because the raw data have (i) a strong asymmetry, (ii) batches at different levels with a widely differing spread. By altering the shape of the batch or batches we may alleviate these problems. We transform the data by applying a single mathematical function to all raw data values [11]. We may need to change not only the units in which the data are stated, but also the basic scale of the measurement. Changes of origin and scale mean the linear transformations, and they leave a shape alone. Nonlinear transformations such as the logarithm and square root are necessary to change shape. The reasons for transforming a batch of original data include the following:

Transforming for symmetry: symmetry of a data batch is often desirable property as many estimates of location work best and are best understood when the data come from a symmetric distribution. In a perfectly symmetric batch, all midsummaries would be equal to the median. If the data were skewed to the right, the midsummaries would increase as they came from letter values further into the tails. For data skewed to the left, the midsummaries would decrease.

Transforming for stable spread: the data sometimes come to us in several batches at different levels and we often find a systematic relationship between spread and level: increasing level usually brings increasing spread. Individual batches become more nearly symmetric and have fewer outliers.

This paper brings a description of the Box-Cox transformation and a re-expression of statistics for transformed data. The procedure of the Box-Cox transformation is illustrated on a typical metallurgical study case concerning an chemometrical estimation of the mean value and lower quantiles of the tensile strength of a construction steel.

2. Methodology of data transformation

Examining data we must often find the proper transformation which leads to symmetric data distribution, stabilizes the variance or makes the distribution closer to normal. Such transformation of original data x to the new variable value $y = g(x)$ is based on an assumption that the original metallurgical data represent a nonlinear transformation of normally distributed variable $x = g^{-1}(y)$.

i) *Transformation for variance stabilization* implies ascertaining the transformation $y = g(x)$ in which the variance $\sigma^2(y)$ is constant. If the variance of the original variable x is a function of the type $\sigma^2(x) = f_1(x)$, the variance $\sigma^2(y)$ may be expressed by $\sigma^2(y) = \left(\frac{\mathrm{d}\,g(x)}{\mathrm{d}\,x}\right)^2 f_1(x) = C$, where C is a constant. When the dependence $\sigma^2(x) = f_1(x)$ is of power (exponent) nature, the optimal transformation will also be a power transformation.

Since for a normal distribution the mean is not dependent on a variance, a transformation that stabilizes the variance makes the distribution closer to normal.

ii) *Transformation for symmetry* is carried out by a simple power transformation

$$y = g(x) = \begin{cases} x^{\lambda} & \text{for parameter } \lambda > 0 \\ \ln x & \text{for parameter } \lambda = 0 \\ -x^{-\lambda} & \text{for parameter } \lambda < 0 \end{cases} \qquad (1)$$

which does not retain the scale, is not always continuous and is suitable only for positive x. Optimal estimates of parameter λ are sought by minimizing the absolute values of particular characteristics of an asymmetry. In addition to the classical estimate of a skewness $g_1(y)$, the robust estimate $g_{1R}(y)$ is used

$$g_{1R}(y) = \frac{(y_{0.75} - y_{0.50}) - (y_{0.50} - y_{0.25})}{(y_{0.75} - y_{0.25})} \qquad (2)$$

The robust estimate of asymmetry $g_p(y)$ may be also expressed with the use of a relative distance between the arithmetic mean $\bar{y}$ and the median $y_{0.50}$ by

$$g_p(y) = \frac{\bar{y} - y_{0.50}}{\sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}} \qquad (3)$$

as for symmetric distributions it is equal to zero, $g_p(y) = 0$.

iii) *Transformation leading to the approximate normality* may be carried out by a family of Box-Cox transformation defined as

$$y = g(x) = \begin{cases} (x^{\lambda} - 1)/\lambda & \text{for parameter } \lambda \neq 0 \\ \ln x & \text{for parameter } \lambda = 0 \end{cases} \qquad (4)$$

where x is a positive variable and λ is real number. The curves of Box-Cox transformation $g(x)$ are monotonic and continuous with respect to parameter λ because $\lim_{\lambda \to 0} \frac{(x^{\lambda} - 1)}{\lambda} = \ln x$ and all transformation curves share one point [$y = 0$, $x = 1$] for all values of λ. The Box-Cox transformation can be applied only on the positive data. To extend this transformation means to make a substitution of x values by $(x - x_0)$ values which are always positive. Here x_0 is the threshold value $x_0 < x_{(1)}$. To estimate the parameter λ in Box-Cox transformation, the method of maximum likelihood may be used because for $\lambda = \hat{\lambda}$ a distribution of transformed variable y is considered to be normal, $N(\mu_y, \sigma^2(y))$. The logarithm of the maximum likelihood function may be written as

$$\ln L(\lambda) = -\frac{n}{2} \ln s^2(y) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i \qquad (5)$$

where $s^2(y)$ is the sample variance of transformed data y. The function $\ln L = f(\lambda)$ is expressed graphically for a suitable interval, for example, $-3 \leq \lambda \leq 3$. The maximum on this curve represents the maximum likelihood estimate $\hat{\lambda}$ Fig.1. The asymptotic $100(1 - \alpha)$ % confidence interval of parameter λ is expressed by $2[\ln L(\hat{\lambda}) - \ln L(\lambda)] \leq \chi^2_{1-\alpha}(1)$, where $\chi^2_{1-\alpha}(1)$ is the quantile of the χ^2 distribution with 1 degree of freedom. This Box-Cox transformation is less suitable if the confidence interval for λ is too wide.

Critical criterion: when the value $\lambda = 1$ is also covered by this confidence interval, the data transformation is not efficient.

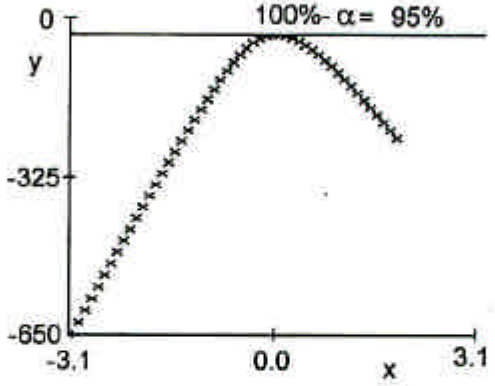

Fig.1 The plot of the logarithm of maximum likelihood estimate of λ for the statistical probability 95%

Re-expression of the statistical measures: after an appropriate transformation of the original data $\{x\}$ has been found, so that the transformed data give approximately normal symmetrical distribution with constant variance, the statistical measures of location and spread for the transformed data $\{y\}$ are calculated. These include the sample mean $\bar{y}$, the sample variance $s^2(y)$, and the confidence interval of the mean $\bar{y} \pm t_{1-\alpha/2}(n-1)\, s(y)/\sqrt{n}$. These estimates must then be recalculated for original data $\{x\}$. Two different approaches to re-expression of the statistics for transformed data can be simply used: (a) Rough re-expressions represent a single reverse transformation $\bar{x}_R = g^{-1}(\bar{y})$. This re-expression for a simple power transformation leads to the general re-expressed mean

$$x_R = x_\lambda = \left[\frac{\sum_{i=1}^{n} x_i^{\lambda}}{n} \right]^{\frac{1}{\lambda}} \qquad (6)$$

where for $\lambda = 0$, $\ln x$ is used instead of x^{λ} and e^x instead of $x^{1/\lambda}$. The re-expressed mean $\bar{x}_R = \bar{x}_{-1}$ stands for the harmonic mean, $\bar{x}_R = \bar{x}_0$ for the geometric mean, $\bar{x}_R = \bar{x}_1$ for the arithmetic mean and $\bar{x}_R = \bar{x}_2$ for the quadratic mean.

(b) The more correct re-expressions are based on the Taylor series expansion of the function $y = g(x)$ in a neighbourhood of the value $\bar{y}$. The re-expressed mean $\bar{x}_R$ is then given

$$x_R = g^{-1}\left\{\bar{y} - \frac{1}{2}\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)\right\} \qquad (7)$$

For variance it then holds $s^2(x_R) = \left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$, where individual derivatives are calculated at the point $x = \bar{x}_R$. The 100(1 - α)% confidence interval of the re-expressed mean for the original data is

$$I_L \le \mu \le I_U \qquad (8)$$

where $I_L = g^{-1}\left[\bar{y} + G - t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right]$, $I_U = g^{-1}\left[\bar{y} + G + t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right]$

and $G = \frac{1}{2}\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$.

On the basis of the (known) actual transformation $y = g(x)$ and the estimates $\bar{y}$, $s^2(y)$ it is easy to calculate re-expressed estimates $\bar{x}_R$ and $s^2(\bar{x}_R)$:

1. For a logarithmic transformation (when $\lambda = 0$) and $g(x) = \ln x$ the re-expressed mean and variance are calculated $\bar{x}_R = \exp\left[\bar{y} + 0.5\, s^2(y)\right]$ and $s^2(\bar{x}_R) = \bar{x}_R^2(y)\, s^2(y)$.

2. For $\lambda \neq 0$ and the Box-Cox transformation the re-expressed mean $\bar{x}_R$ will be represented by one of the two roots of the quadratic equation

$$x_{R,1,2} = \left[0.5(1+\lambda\bar{y}) \pm 0.5\sqrt{1 + 2\lambda\left(\bar{y} + s^2(y)\right) + \lambda^2\left(\bar{y}^2 - 2s^2(y)\right)}\right]^{1/\lambda} \qquad (9)$$

which is closest to the median $x_{0.5} = g^{-1}(\bar{y}_{0.5})$. If $\bar{x}_R$ is known the corresponding variance may be calculated from $s^2(x) = \bar{x}_R^{(-2\lambda+2)}\, s^2(y)$.

3. Experimental procedure

Procedures Power transformation and Box-cox transformation of the statistical systems ADSTAT [11] search parameters of a simple power transformation and parameters of the normalizing Box-Cox transformation of data. It also enables the exploratory data analysis of transformed data. For a transformation the different measures of symmetry are calculated and the sample skewness in range $-3 \le \lambda \le 3$ with a step 0.1 and the optimal values of these measures are printed. For the transformation the estimate λ maximizing $\ln L(\lambda)$ is calculated. Selected λ is used in calculation of estimates $\bar{y}$, $s^2(y)$, $g_1(y)$, and $g_2(y)$. Then from these estimates, the re-expressed estimates of original variables $\bar{x}_R$, $s^2(\bar{x}_R)$, and the 95% confidence interval of the re-expressed variable of location μ are calculated, [11-12].

4. Results and discussion

Many statistical programs, for example NCSS2000 [13], offer a list of various point parameters of a location and spread but rarely help the user to choose the statistically adequate parameter to an actual sample batch. The exploratory data analysis and an examination of sample assumptions will find an answer to this question. The following study case with methodology runs on typical metallurgical sample data will illustrate a rigorous procedure of the statistical treatment of univariate data with the exploratory data analysis.

Study Case: *Estimation of the mean value and lower quantiles of the tensile strength of a construction steel*

The goal is to estimate the parameters of location and spread of the tensile strength R_m of a construction steel when two data sets were analysed. The original large data set (Set A) was of size $n = 279$ while the second smaller one (Set B) was created by a random selection of sample elements from the first one to the resulting size $n = 41$ (Table 1). The smaller subsample was analyzed to illustrate the importance of the sample size for reliability of the estimates and their confidence interval. On the other hand, it is to show the stability of the results even for relatively smaller samples. Generally, the sample should be as large as possible, namely for determination of very low (or high) quantiles. The steel company requires such a steel quality that a number of samples elements having the tensile strength R_m equal or less than 500 MPa should be less than 1% what expressed statistically means that 1% quantile of the tensile strength R_m should be greater than 500 MPa. Find a rigorous mean value of both data sets and estimate 1% quantile.

Table 1 Set B: measured values of the tensile strength [MPa] of a construction steel

668	621	620	551	592	588	591	613	628	576	569	608
556	548	557	605	542	559	567	539	524	552	532	565
577	557	557	539	534	555	518	523	514	577	713	520
552	545	533	554	532							

(1) *Survey of descriptive statistics* for Set A and for Set B (in brackets): the statistical software NCSS2000 [13] for an actual sample batch calculates a survey of parameters of location and spread (an elucidation of statistics *cf. ref.* [11-12]). Then, on base of the exploratory data analysis the user should select the most convenient parameter of location from following available estimates (in {.} brackets {for Set B}): the arithmetic mean $\bar{x}$ = 567.5 {567.6} MPa with the confidence limits L_L = 563.1 {554.4} MPa and L_U = 571.8 {580.8} MPa, the median $x_{0.5}$ = 561.0 {556.5} MPa, the geometric mean $\bar{x}_g$ = 566.3 {566.2} MPa, the harmonic mean $\bar{x}_h$ = 565.2 {564.8} MPa and following trimmed means $\bar{x}(5\%)$ = 564.9 {563.6} MPa, $\bar{x}(10\%)$ = 563.8 {562.5} MPa, $\bar{x}(25\%)$ = 561.6 {559.1}

MPa, $\bar{x}(45\%)$ = 560.6 {556.5} MPa, the standard deviation s = 36.6 {41.9} MPa, the interquantile range R_I = 199 {204} MPa, and at last a survey of parameters of shape: the skewness g_1 = 0.99 {1.47} and kurtosis g_2 = 3.99 {5.64} which prove skewed distribution.

(2) *Exploratory data analysis:* some selected diagnostic plots prove asymmetric distribution, (*cf.* pp. 35 - 67 in *ref.* [11]). Both quantile plots for Set A (Fig.2a) and also for Set B (Fig.2b) show an asymmetric distribution with at least 2 outliers, the kernel estimate of the probability density function (Fig.3a, Fig.3b) and the quantile-quantile plot (Fig.4a, Fig.4b) prove also an asymmetric distribution. The quantile-quantile plot also called the rankit plot checking a normal distribution does not exhibit close agreement of sample points with a straight line of the normal distribution. Not constant halfsums Z_Q and positive skewness S_Q clearly indicate a skew distribution. Tails length T_Q for this distribution cannot be used for deeper analysis.

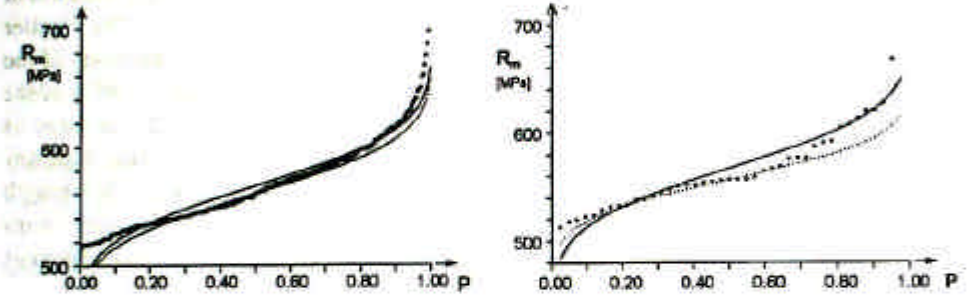

Fig.2 The quantile plot of the steel data: (a) the data set A, (b) the data set B

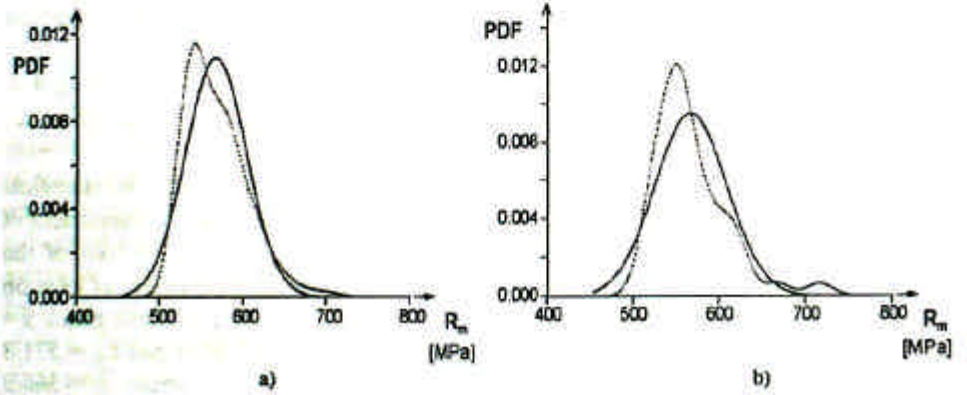

Fig.3 The kernel estimator of the probability density plot of the steel data. The empirical curve (dashed curve) and the normal distribution approximation (dot curve): (a) the data set A, (b) the data set B

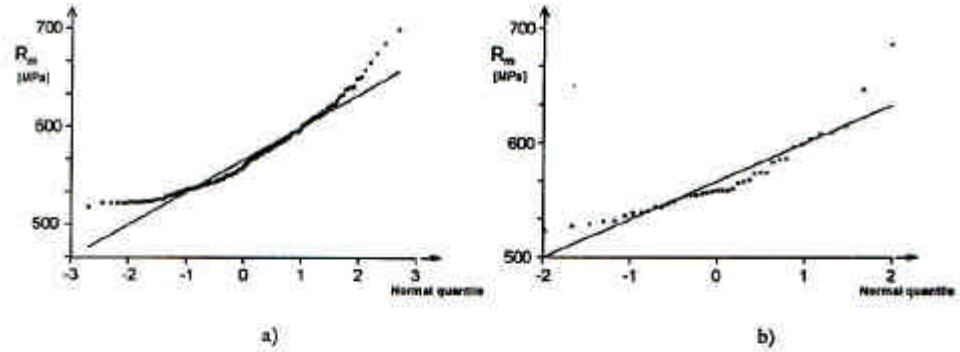

Fig.4 The quantile-quantile plot (for normal distribution called the rankit plot) of the steel data: (a) the data set A, (b) the data set B

Table 2 The quantile measures of location, spread and shape for data set B from Table 1

Quantile	P	Lower quantile Q_L	Upper quantile Q_U	Range R_Q	Halfsum Z_Q	Skewness S_Q	Tails Length T_Q	Pseudo-Sigma G_Q
Median	0.5	557	557	-	-	-	-	-
Quartile	0.25	539	588	49	563.5	10.63	0	36.35
Octile	0.125	529	613	84	571	6.2	0.539	36.52
Sedecile	0.0625	521.5	624.5	103	573	5.08	0.743	33.66

(3) *Basic assumptions about the sample* (*cf.* pp. 78 - 82 in *ref.* [11]): applying an analysis of basic assumptions about data the following conclusions were met: a test of sample elements independence leads to the test statistic t_{17} = 1.132 < $t_{0.975}(42)$ = 2.018 and therefore an independence is not rejected. A combined sample skewness and curtosis test of normality leads to the test statistic C_1 = 16.98 {9.16} > $\chi^2(0.95, 2)$ = 5.992 {5.992} and therefore a normality of data distribution was rejected. Because data are skewed, an analysis of the sample homogeneity based on a normality assumption cannot be used.

(4) *Data transformation:* most diagnostic plots of EDA exhibit an asymmetric distribution of an original sample data and therefore necessity of data transformation is proved. In case of Box-Cox transformation the true mean value of a sample distribution with both confidence limits L_L and L_U are calculated. From the plot of the logarithm of the likelihood function for the Box-Cox transformation the maximum of the curve is at λ = -1.20 {-1.82}, Fig.5. For both samples the corresponding 95% confidence interval of a found exponent λ -⟨1.8; -0.6⟩ and {⟨-3.5; -0.3⟩} does not contain the exponent value λ = 1, so the transformation is statistically significant.

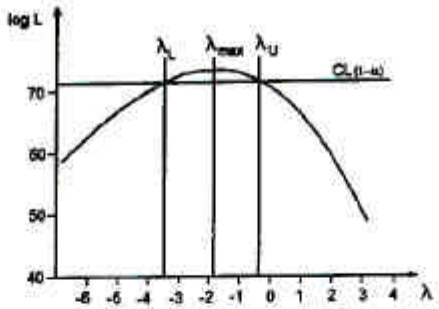

Fig.5 The plot of the logarithm of likelihood (L) in dependence on the power λ for the steel data B and estimation of the optimal power λ_{max} with its lower λ_L and upper λ_U limits of the confidence interval for the confidence level $(1-\alpha)$, $CL_{1-\alpha}$

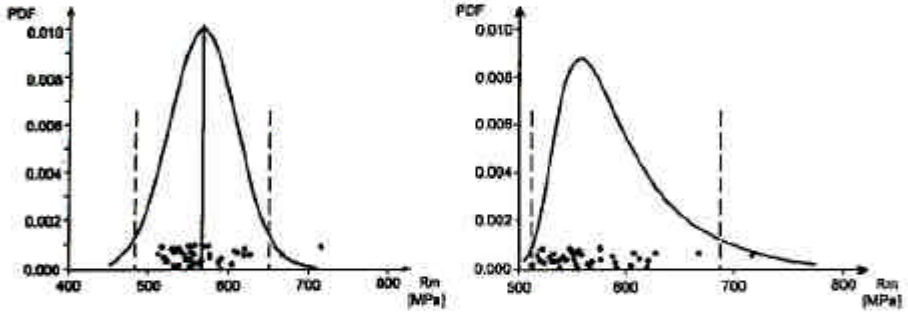

Fig. 6 (a) The normal curve and (b) the transformed normal density for the steel data B in Table 1. The vertical dashed lines indicate the interval $\bar{x} \pm 2s$

(5) *Conclusion:* original data should be transformed to improve a symmetry of data distribution. While the classical measures of location, spread and shape for the original data can not be used as they lead to false values, i. e. mean $\bar{x}$= 567.6 {567.6} MPa, the corrected mean value $\bar{x}_R$= 560.4 {558.5} MPa found with the Box-Cox transformation. The classical approach based on the assumption of normality gave the 1% and 0.25% quantiles 470.2 MPa and 450.1 MPa what means that a required quality was not reached. However, using a data transformation, much more realistic values of 1% and 0.25% quantiles are 509.4 MPa and 502.9 MPa which state that a steel quality is fully acceptable. The difference between the transformed and untransformed estimates is also obvious on Fig.6a and 6b. It may be concluded that the arithmetic mean can be used only for a symmetrical distribution. For an asymmetrical distribution the data transformation should be applied.

Acknowledgement

Financial support of the Grant Agency of the Czech Republic (Grant No 303/00/1559) is thankfully acknowledged.

Literature

[1] Tukey, J. W.: Exploratory Data Analysis. Reading, Addison Wesley, Massachusetts 1977
[2] Chambers, J., Cleveland, W., Kleiner, W. and Tukey, J. W.: Graphical Methods For Data Analysis. Duxbury Press, Boston, 1983
[3] Hoaglin, D. C., Mosteler, F. and Tukey, J. W.: Exploring Data Tables, Trends and Shapes. J. Wiley and Sons, New York, 1985
[4] Silverman, B. W.: Density Estimation. Chapman and Hall, London, 1986
[5] Lejenne, M., Dodge, Y. and Koelin, E.: Proceedings of the Conference COMSTAT'82, Vol III, p. 173, Toulouse 1982
[6] Hoaglin, D. C., Mosteler, F. and Tukey, J. W. (Editors): Understanding Robust And Exploratory Data Analysis. J. Wiley and Sons, New York, 1983
[7] Kafander, K. and Spiegelman, C. H.: An Alternative to Ordinary Q-Q Plot, Comput. Stat. Data Anal., vol. 4, 1986, p. 167
[8] Hines, W. G. S. and Hines, R. J. H.: Quick Graphical Power Law Transformation Selection, Am. Statist., vol. 41, 1987, p. 21
[9] Hoaglin, D. C.: Performance of Some Resistant Rules For Outlier Labeling, J. Am. Statist. Assoc., vol. 81, 1986, p. 991
[10] Stoodley, K.: Applied And Computational Statistics, Ellis Horwood, Chichester, 1984
[11] Meloun, M., Militký, J. and Forina, M.: Chemometrics For Analytical Chemistry, Vol. 1, PC-Aided Statistical Data Analysis. Ellis Horwood, Chichester, 1992
[12] Statistical package ADSTAT, TriloByte Statistical Software, Pardubice, 1999
[13] NCSS Statistical Software, 329 North 1000 East, Kaysville, Utah 84037, Email: sales@ncss.com