Clin Chem Lab Med 2000; 38(6):553–559 © 2000 by Walter de Gruyter · Berlin · New York

Transformation in the PC-Aided Biochemical Data Analysis

Milan Meloun[1], Martin Hill[2], Jiří Militký[3] and Karel Kupka[4]

[1] Department of Analytical Chemistry, Faculty of Chemical Technology, University Pardubice, Pardubice, Czech Republic
[2] Institute of Endocrinology, Prague, Czech Republic
[3] Department of Textile Materials, Technical University, Liberec, Czech Republic
[4] Trilobyte Statistical Software Ltd., Pardubice, Czech Republic

Data transformations enable expression of original data in a new scale, more suitable for data analysis. In computer-aided interactive analysis of biochemical and clinical data an exploratory data analysis often finds that the sample distribution is systematically skewed or does not accept a sample homogeneity. Under such circumstances the original data should be transformed. The power transformation and the Box-Cox transformation improve sample symmetry and also stabilize variance. Both the Hines-Hines selection graph and the plot of logarithm of a maximum likelihood function allow selection of an optimum transformation parameter. The proposed procedure of data transformation in univariate data analysis is illustrated on a determination of 17-hydroxypregnenolone in umbilical blood of a population of newborns. Lower levels of free 5-ene steroids in umbilical blood and elevated levels of 5-ene steroid sulfates indicate a congenital sex-specific placental sulfatase insufficiency. After examination of statistical assumptions by diagnostic plots of an exploratory data analysis the best estimate of a mean value of 17-hydroxypregnenolone is derived.

Key words: Data transformation; Exploratory analysis; Hines-Hines selection graph; 5-ene steroids; 17-hydroxypregnenolone; Umbilical blood.

Abbreviations: EDA, exploratory data analysis.

Introduction

When an exploratory data analysis (1–3) indicates that the sample distribution strongly differs from the normal one, we are faced with the problem of how to analyze biochemical or medical 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) marked 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 involve linear transformations, and they do not affect the shape. Non-linear 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 to enhance interpretability: changing the scale of measurement is natural because it gives an alternative way to report information. The implied transformation is to the more convenient scale, *e.g.* a merit of the temperature scale is that the zero of the Celsius scale corresponds to a natural and widely understood phenomenon, the freezing point of water. Therefore, we transform Fahrenheit degrees (*F*) to Celsius degree (*C*) using a linear transformation, C = (5/9)(F-32).

Transforming for symmetry: symmetry of a batch is often a desirable property as many estimates of location work best and are best understood when the data come from a symmetric distribution. A simple way to check on symmetry is to define a set of midsummaries (each midsummary is the average of the two corresponding quantiles (also known as "letter values") Q_L and Q_U: for example, lower and upper quartiles F_L and F_U, lower and upper octiles E_L and E_U, lower and upper sedeciles D_L and D_U, etc.). 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: biochemical 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. When this relationship is strong, we have several reasons for transforming the data in a way that reduces or eliminates the dependence of spread on level. The transformed data will be better suited for comparison and visual exploration. The transformed data may be better suited for common confirmatory techniques. Individual batches become more nearly symmetric and have fewer outliers.

This paper gives a description of the power transformation and the Box-Cox transformation and re-expression of statistics for transformed data. The procedure of the power transformation and the Box-Cox transformation is illustrated on a typical biochemical study case concerning a determination of 17-hydroxypregnenolone in umbilical blood of newborns.

Methods

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 biochemical 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

$$\sigma^2(y) = \left(\frac{\mathrm{d}\, g(x)}{\mathrm{d}\, x}\right)^2 f_1(x) = C \qquad [1]$$

where C is a constant. The chosen transformation $g(x)$ is then the solution of the differential equation

$$g(x) \approx \int C \frac{\mathrm{d}\, x}{\sqrt{f_1(x)}} \qquad [2]$$

In some instrumental methods of analytical chemistry, biochemistry and clinical chemistry the relative standard deviation $\delta(x) = \sigma(x)/\bar{x}$ of the measured variable is constant. This means that the variance $\sigma^2(x)$ is described by a function $\sigma^2(x) = f_1(x) = \delta^2(x)\, x^2 = \text{const}\, x^2$. After substitution into equation [2] and solution of differential equation, the transformation $g(x) = \ln x$ results. Optimal transformation of original data is the logarithmic transformation. This transformation leads to the use of a geometric mean.

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 the 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

$$y = g(x) = \left\{ \begin{array}{ll} x^{\lambda} & \text{for parameter } \lambda > 0 \\ \ln x & \text{for parameter } \lambda = 0 \\ -x^{-\lambda} & \text{for parameter } \lambda < 0 \end{array} \right\} \qquad [3]$$

which does not retain the scale, is not always continuous and is suitable only for positive x. Optimal estimates of parameter $\hat{\lambda}$ are sought by minimizing the absolute values of particular characteristics of an asymmetric distribution. In addition to the classical estimate of skewness $\hat{g}_1(y)$, the robust estimate $\hat{g}_{1,R}(y)$ is used

$$\hat{g}_{1,R}(y) = \frac{(\tilde{y}_{0.75} - \tilde{y}_{0.50}) - (\tilde{y}_{0.50} - \tilde{y}_{0.25})}{(\tilde{y}_{0.75} + \tilde{y}_{0.25})} \qquad [4]$$

The robust estimate of an asymmetry $\hat{g}_P(y)$ may be also expressed with the use of a relative distance between the arithmetic mean $\bar{y}$ and the median $\tilde{y}_{0.50}$ by

$$\hat{g}_P(y) = \frac{\bar{y} - \tilde{y}_{0.50}}{\sqrt{\dfrac{\sum_{i=1}^{n} (y_1 - \bar{y})^2}{n-1}}} \qquad [5]$$

as for symmetric distributions it is equal to zero, $\hat{g}_P(y) \approx 0$.

iii) Transformation leading to approximate normality may be carried out by the use of the family of Box-Cox transformations (11), defined as

$$y = g(x) = \left\{ \begin{array}{ll} (x^{\lambda} - 1)/\lambda & \text{for parameter } \lambda \neq 0 \\ \ln x & \text{for parameter } \lambda = 0 \end{array} \right\} \qquad [6]$$

where x is a positive variable and λ is a real number. The Box-Cox transformation has the following properties:

a) The curves of transformation $g(x)$ are monotonic and continuous with respect to parameter λ because

$$\lim_{\lambda \to 0} \frac{(x^{\lambda} - 1)}{\lambda} = \ln x \qquad [7]$$

b) All transformation curves share one point [$y = 0$, $x = 1$] for all values of λ. The curves nearly coincide at points close to [0, 1]; *i.e.*, they share a common tangent line at that point.

c) The power transformations of exponent –2; –3/2; –1; –1/2; 0; 1/2; 1; 3/2; 2 have equal spacing between curves in the family of Box-Cox transformation graphs.

The Box-Cox transformation can be applied only to positive data. To extend this transformation it is necessary to make a substitution of x values by $(x - x_0)$ values which are always positive. Here x_0 is the threshold value $x_0 < x_{(1)}$.

An excellent diagnostic tool enabling estimation of parameter λ is represented by the Hines-Hines selection graph (8). It is based on the equation

$$\left(\frac{\tilde{x}_{Pi}}{\tilde{x}_{0.5}}\right)^{\lambda} + \left(\frac{\tilde{x}_{0.5}}{\tilde{x}_{1-Pi}}\right)^{-\lambda} = 2 \qquad [8]$$

valid for distribution symmetrical around a median. For the cumulative probability $P_i = 2^{-i}$, the letter values F, E, $i = 2, 3$ are usually chosen.

To compare the empirical course of experimental points with the ideal one, ideal curves for various values of parameter λ are drawn in a selection graph. These curves λ represent

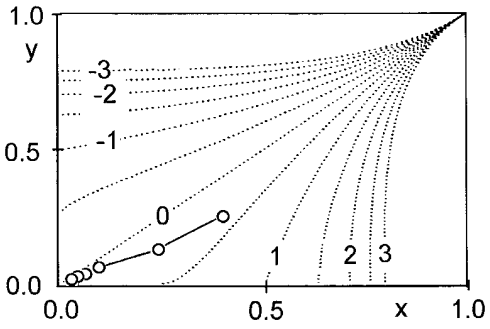

Fig. 1 Graphical estimation of λ from a Hines-Hines selection plot in the range [–3; +3]. Circles denote sample points.

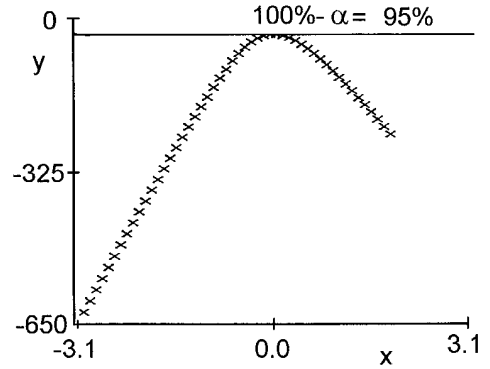

Fig. 2 The plot of the logarithm of maximum likelihood estimates λ for the statistical probability 95%.

a solution of the equation $y^{\lambda} + x^{-\lambda} = 2$ in the range $0 \le x \le 1$ and $0 \le y \le 1$:

1. For $\lambda = 0$ the solution is a straight line $y = x$.
2. For $\lambda \le 0$ the solution is in a form $y = (2 - x^{-\lambda})^{1/\lambda}$.
3. For $\lambda \ge 0$ the solution is in a form $x = (2 - y^{\lambda})^{-1/\lambda}$.

The estimate $\hat{\lambda}$ is guessed from a selection graph, according to the location of experimental points near to the various ideal curves. 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

$$\ln L(\lambda) = -\frac{n}{2} \ln s^2(y) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i \qquad [9]$$

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, $-3 \le \lambda \le 3$. The maximum on this curve represents the maximum likelihood estimate $\hat{\lambda}$. The asymptotic $100(1 - \alpha)$ % confidence interval of parameter λ is expressed by $2\,[\ln L(\hat{\lambda}) - \ln L(\lambda)] \le \chi^2_{1-\alpha}$ where $\chi^2_{1-\alpha}(1)$ is the quantile of the χ^2 distribution with 1 degree of freedom. This interval contains all values λ for which it is true that

$$\ln L(\lambda) \ge \ln L(\hat{\lambda}) - 0.5\chi^2_{1-\alpha}(1) \qquad [10]$$

This Box-Cox transformation is less suitable if the confidence interval for λ is too wide. When the value $\lambda = 1$ is also covered by this confidence interval, the transformation is not efficient.

Re-expression of the statistical measures

After an appropriate transformation of the original data $\{x\}$ has been found, so that the transformed data give an 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 the 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

$$\bar{x}_R = \bar{x}_{\lambda} = \left[\frac{\sum_{i=1}^{n} x_i^{\lambda}}{n} \right]^{\frac{1}{\lambda}} \qquad [11]$$

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

$$\bar{x}_R \approx g^{-1}\left\{ \bar{y} - \frac{1}{2} \frac{d^2 g(x)}{dx^2} \left(\frac{dg(x)}{dx} \right)^{-2} s^2(y) \right\} \qquad [12]$$

For variance it is then valid that

$$s^2(x_R) \approx \left(\frac{dg(x)}{dx} \right)^{-2} s^2(y) \qquad [13]$$

where individual derivatives are calculated at the point $x = \bar{x}_R$. The $100(1 - \alpha)$% confidence interval of the re-expressed mean for the original data may be defined as

$$\bar{x}_R - l_L \le \mu \le \bar{x}_R + l_U \qquad [14]$$

where $l_L = g^{-1}\left[\bar{y} + G - t_{1-\alpha/2}(n-1) \frac{s(y)}{\sqrt{n}} \right]$ [14a]

and $l_U = g^{-1}\left[\bar{y} + G + t_{1-\alpha/2}(n-1) \frac{s(y)}{\sqrt{n}} \right]$ [14b]

$$G = -\frac{1}{2} \frac{d^2 g(x)}{dx^2} \left(\frac{dg(x)}{dx} \right)^{-2} s^2(y) \qquad [14c]$$

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

$$\bar{x}_R \approx \exp\,[\bar{y} + 0.5\, s^2(y)] \qquad [15]$$

and $$s^2(\bar{x}_R) \approx \bar{x}_R^2(y)\, s^2(y) \qquad [16]$$

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

$$\bar{x}_{R,1,2} = [0.5(1 + \lambda\bar{y}) \pm 0.5 \sqrt{1 + 2\lambda(y + s^2(y)) + \lambda^2(y^2 - 2s^2(y))}]^{1/\lambda} \qquad [17]$$

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

$$s^2(x) = \bar{x}_R^{(-2\lambda+2)}\, s^2(y) \qquad [18]$$

Results

Proposed procedure

Procedures Power transformation and Box-Cox transformation of the statistical systems Adstat or QC-Expert (11) search parameters of a simple power transformation and parameters of the normalized 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 the range of $-3 \le \lambda \le 3$ with a step value of 0.1, and the optimal values of these measures are printed. The selection graph is drawn as well as the points of optimal values of λ. From this graph the value of λ can be estimated. Using transformed data, the mean $\bar{y}$, the variance $s^2(y)$, the skewness $\hat{g}_1(y)$, and the curtosis $\hat{g}_2(y)$ are calculated. These computations can be repeated for various values of λ. For the transformation the estimate λ maximizing $\ln L(\lambda)$ is calculated. The selected λ is used in calculation of estimates $\bar{y}$, $s^2(y)$, $\hat{g}_1(y)$, and $\hat{g}_2(y)$. 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 μ are calculated.

Procedure Transformation in software Adstat or QC-Expert (12) searches parameters of the simple power transformation and parameters of the normalized Box-Cox transformation of data.

Discussion

Many statistical programs offer a list of various point parameters of location and spread but rarely help the user to choose the statistically adequate parameter for an actual sample batch. Exploratory data analysis and

Tab. 1 17-Hydroxypregnenolone (17-hydroxypregnenolone) (nmol/l) for sample size $n = 99$.

19.00	15.41	20.20	19.70	41.00	39.60	8.77	44.60	33.50	28.60	30.30	21.20
431.0	45.00	21.50	32.90	53.00	53.60	19.00	73.90	17.60	27.60	22.20	32.30
41.00	28.40	14.80	37.00	10.40	16.60	67.90	57.30	41.00	239.0	16.50	13.00
68.50	7.32	35.00	22.90	45.80	37.70	7.42	78.20	30.70	34.00	63.00	48.90
16.30	75.70	10.40	16.80	20.10	11.00	18.30	28.30	8.86	9.13	53.10	9.67
52.50	34.10	16.80	39.80	97.00	5.91	25.40	15.80	34.00	22.20	51.30	17.40
33.10	52.10	37.50	28.90	29.80	7.77	10.80	16.30	26.70	26.90	27.30	13.60
26.00	12.50	14.10	38.00	28.50	82.70	24.10	45.40	23.70	42.90	15.80	26.10
30.00	29.90	31.40									

an examination of sample assumptions will provide an answer to this question. First study case with methodology runs on typical biochemical sample data will illustrate a rigorous procedure of the statistical treatment of univariate data with exploratory data analysis.

Study case: Determination of 17-hydroxypregnenolone in the umbilical blood of newborns. Lower levels of free 5-ene steroids in umbilical blood and elevated levels of 5-ene steroid sulfates indicate a congenital sex-specific placental sulfatase insufficiency (13). Delayed onset of labor, frequently linked with the necessity of intervention (14), together with relatively low birth weight is the common symptom of the disease. The recessive X-linked type of defect also called the disease of "dry skin", may have phenotypic consequences in later postnatal life (15). The incidence of this disorder appears to be approximately one per 2000 male births (16). The exact assessment of the mean value and the variance of steroid levels in controls are necessary for the correct judgment of the samples from patients. Low levels of 5-ene steroid sulfates are common in pregnancies complicated by intrauterine fetal growth retardation (17). The levels of pregnenolone, 17-hydroxypregnenolone (13), (DHEA) and the levels of respective 3β-OH sulfates were evaluated. The evaluation of levels of 17-hydroxypregnenolone was chosen as an example of the correct data analysis. Statistical assumptions should be tested on the group of umbilical blood from newborns using some plots of an extended exploratory data analysis and the statistical tests of basic assumptions. The best estimate of a mean value in 17-hydroxypregnenolone is to be evaluated.

Solution

(1) Survey of descriptive statistics: the statistical software NCSS2000 (18) for an actual sample batch calculates a survey of parameters of location and spread for $n = 99$ (for an explanation of statistics *cf.* ref. (11)). On the basis of the exploratory data analysis (EDA) the user should select the most convenient parameter of location from the following available estimates: the arithmetic mean $\bar{x} = 37.2$ nmol/l, the median $\bar{x}_{0.5} = 28.4$ nmol/l, the geometric mean $\bar{x}_g = 27.4$ nmol/l, the harmonic mean $\bar{x}_h = 21.9$ nmol/l, the mode $\bar{x}_M = 41$ nmol/l, and following trimmed means $\bar{x}(5\%) = 30.6$ nmol/l with $s(5\%) = 15.7$ nmol/l and $n(5\%) = 89$, $\bar{x}(10\%) = 29.4$ nmol/l with $s(10\%) = 12.2$ nmol/l and $n(10\%) = 79$, $\bar{x}(25\%) = 28.1$ nmol/l with $s(25\%) = 10.1$ nmol/l and $n(25\%) = 50$, $\bar{x}(45\%) = 28.3$ nmol/l with $s(45\%) = 0.98$ nmol/l and $n(45\%) = 10$. A survey of parameters of spread is available: the variance $s^2 = 1562.7$, the standard deviation $s = 48.8$ nmol/l, the unbiased standard deviation $s = 48.9$ nmol/l, the interquantile range $R_F = 24.2$ nmol/l, and finally a survey of parameters of shape: the skewness $\hat{g}_1 = 6.14$, the kurtosis $\hat{g}_2 = 46.96$.

(2) Basic diagnostic plots in the EDA are used for a graphical visualization of data: the quantile plot (Figure 3) shows a strong deviation from a normal distribution as all sample points do not fit a line and two outliers at high values are indicated. Both dot diagrams (Figure 4) and the box-and-whisker plot (Figure 5) indicate five outliers at high values and an asymmetric, skewed distribution. In the halfsum plot (Figure 6) and in the symmetry plot (Figure 7) most sample points are outside the confidence limits and both diagnostic plots indicate that the sample distribution is strongly skewed. The kurtosis plot (Figure 8) indicates two outliers, being outside the confidence limits. The quantile-box plot

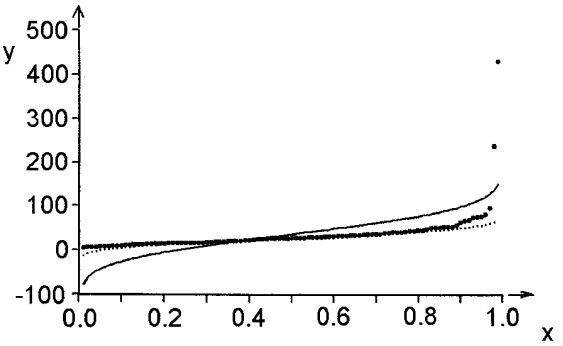

Fig. 3 The quantile plot of 17-hydroxypregnenolone data.

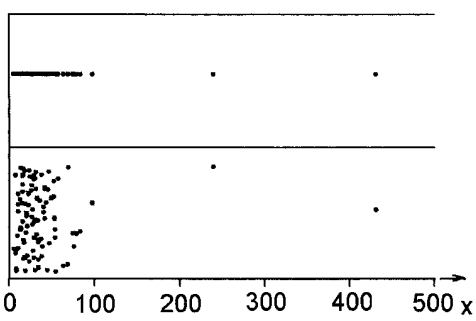

Fig. 4 The dot and jitter dot diagram of 17-hydroxypregnenolone data.

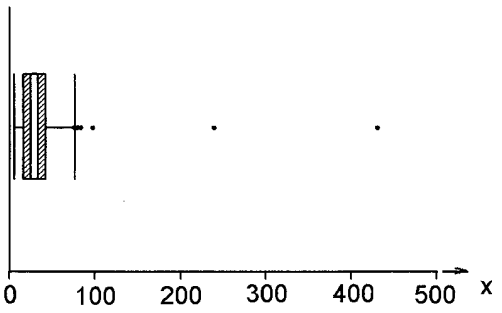

Fig. 5 The box-and-whisker plot of 17-hydroxypregnenolone data.

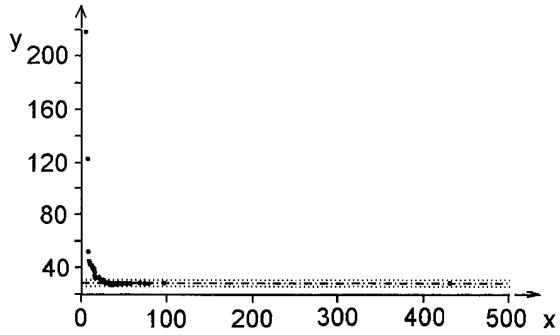

Fig. 6 The halfsum plot of 17-hydroxypregnenolone data.

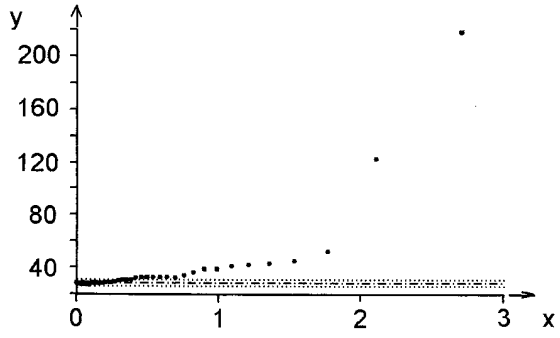

Fig. 7 The symmetry plot of 17-hydroxypregnenolone data.

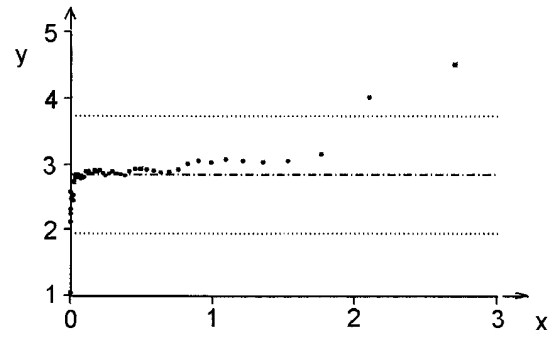

Fig. 8 The kurtosis plot of 17-hydroxypregnenolone data.

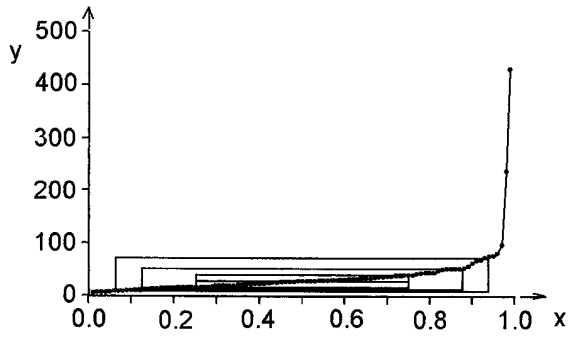

Fig. 9 The quantile-box plot of 17-hydroxypregnenolone data.

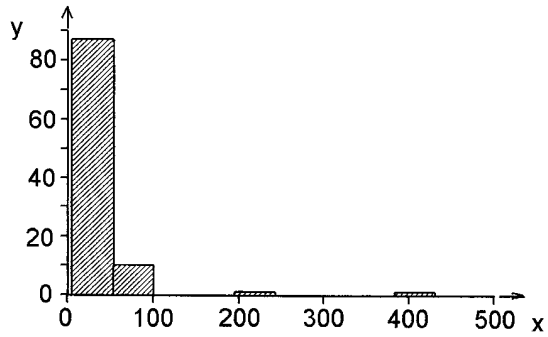

Fig. 10 The histogram of 17-hydroxypregnenolone data.

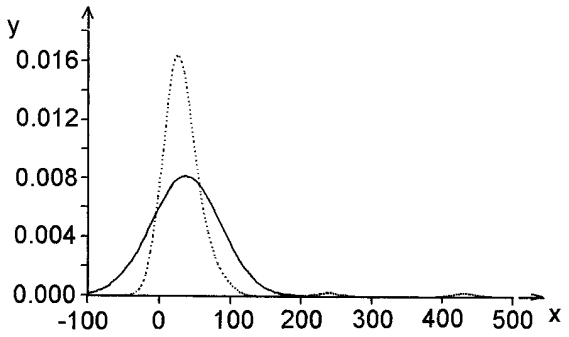

Fig. 11 The Kernel estimator of the probability density plot of 17-hydroxypregnenolone data: the empirical curve (dot curve) and the normal distribution approximation (full curve).

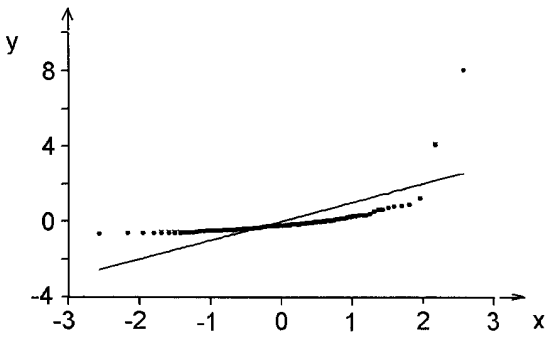

Fig. 12 The quantile-quantile plot (for normal distribution called the rankit plot) of 17-hydroxypregnenolone data.

(Figure 9) shows an asymmetric distribution with 5 outliers.

(3) Determination of sample distribution in the EDA: the sample distribution, represented by symmetry, skewness, and kurtosis is examined by four plots: the histogram (Figure 10) shows that most sample points are located in one class. The kernel density estimator of the probability density function (Figure 11) indicates a skewed sample distribution with two or three outliers. The rankit plot (Figure 12) checking a normal distribution does not exhibit close agreement of sample points

Tab. 2 The quantile measures of location, spread and shape for 17-hydroxypregnenolone data (nmol/l).

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	28.40	28.40	–				
Quartile	0.25	16.80	41.00	24.20	28.90	0.43	0.000	17.95
Octile	0.125	12.62	53.08	40.45	32.85	0.27	0.514	17.59
Sedecile	0.0625	9.20	73.23	64.03	41.21	0.12	0.973	20.92

with a straight line. The highest value of the correlation coefficient $r = 0.90795$ of the Q-Q plot is reached for the lognormal distribution. The probability-probability P-P plot (Figure 13) does not prove a normal distribution.

Not constant halfsums Z_Q and positive skewness S_Q clearly indicate a skewed distribution. Tail length T_Q for this distribution cannot be used for deeper analysis. The point estimate of skewness of 6.14 and that of kurtosis at 46.96 indicate that the sample distribution is strongly asymmetric with a slim and sharp peak and definitely not normal.

(4) 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) Examination for independence of sample elements: a test of independence of sample elements leads to the test statistic $t_{17} = 0.358 < t_{0.975}(100) = 1.984$ and therefore independence is accepted.

(b) Examination for normality of sample distribution: a combined sample skewness and kurtosis test leads to the test statistic $C_1 = 9949.0 > \chi^2(0.95, 2) = 5.992$ and therefore normality of data distribution was rejected.

(c) Examination of sample homogeneity: because data are skewed, an examination of sample homogeneity based on normality assumption cannot be used.

(5) Data transformation: most diagnostic plots of EDA exhibit an asymmetric distribution of the original sample data and therefore show the necessity for data transformation. In case of Box-Cox transformation the true mean value of a sample distribution with both confidence limits L_L and L_U was calculated. From the plot of the logarithm of the likelihood function for the power transformation, the maximum on the curve was read from a graph at $\lambda = -0.1$ (Adstat), for the Box-Cox transformation the maximum of the curve is at $\lambda = -5.6$ (Figure 14) (QC-Expert). For both transformations the corresponding 95% confidence interval does not contain the exponent value $\lambda = 1$, so all transformations are statistically significant. The rankit plot on Figure 15 shows that the Box-Cox transformation brings more accurate results.

The classical measures of location, spread and shape for the original data, *i.e.* mean $\bar{x} = 37.2$ nmol/l, standard deviation $s(x) = 48.8$ nmol/l, skewness $\hat{g}_1(x) = 6.14$ and kurtosis $\hat{g}_2(x) = 46.96$ are out of statistical significance and may be taken as false estimates. The power transformation ($\hat{\lambda} = -0.13$, Adstat) estimated the corrected mean value $\bar{x}_R = 26.5$ nmol/l, the Box-Cox transformation ($\hat{\lambda} = -5.638$, Adstat) the corrected mean value $\bar{x}_R = 26.5$ nmol/l and the exponential transforma-

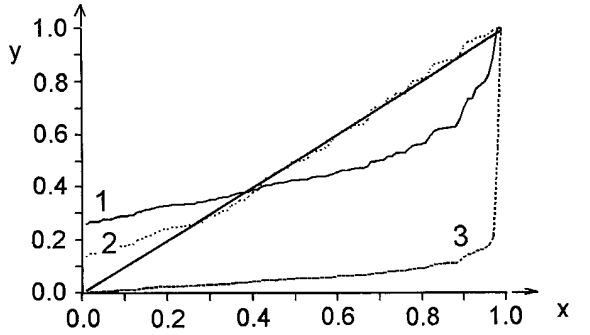

Fig. 13 Probability-probability plot of 17-hydroxypregnenolone data approximated by curve of (1) the normal distribution, (2) the Laplace distribution, and (3) the rectangular distribution.

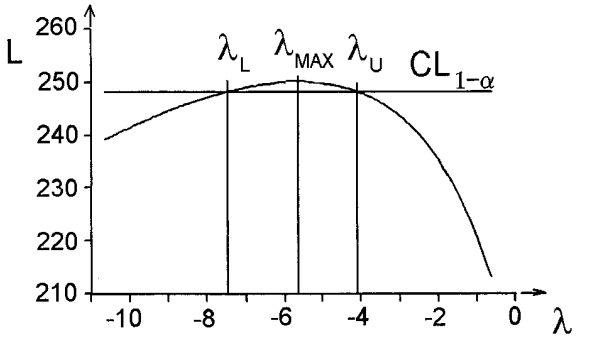

Fig. 14 The plot of the logarithm of maximum likelihood (L) in dependence on the power λ for 17-hydroxypregnenolone data and estimation of the optimal power λ_{max} with its lower λ_L and upper λ_U limits of the confidence interval for the confidence level $(1-\alpha)$, $CL_{1-\alpha}$.

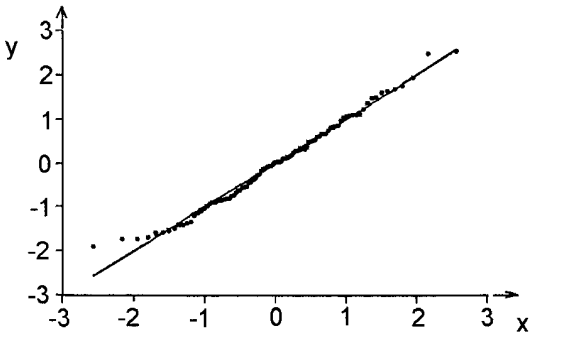

Fig. 15 The quantile-quantile plot for 17-hydroxypregnenolone data after the Box-Cox transformation. (Compare this plot before transformation on Figure 12).

tion ($\hat{\lambda}$ = 2.285, QC-Expert) found the corrected mean value $\bar{x}_R$ = 26.2 nmol/l with the confidence interval L_L = 23.3 nmol/l and L_U = 29.6 nmol/l.

(6) Conclusion: all EDA display techniques prove that the sample distribution is skewed with 3 outliers and does not come from a population with a normal distribution. For the best estimate of a location parameter the arithmetic mean gives a false value 37.2 nmol/l and can not be used. Instead of this arithmetic mean the median $\bar{x}_{0.5}$ = 28.4 nmol/l is more suitable and can be recommended. For parameter of a spread the standard deviation of the median s =2.1 nmol/l may be used, but the parameters describing a shape, *i.e.* the skewness $\hat{g}_1$ = 6.14 and the kurtosis $\hat{g}_2$ = 46.96 indicate strongly asymmetric and skewed distribution. The interval estimate for parameter of location is described by the confidence interval of the median, L_L = 24.8 nmol/l, L_U = 32.0 nmol/l, in which the unknown concentration exists with 95% confidence. On base of the quantile-quantile plot the Box-Cox transformation is considered as the most rigorous one with the corrected mean value $\bar{x}_R$ = 26.5 nmol/l and two confidence limits L_L = 23.1 nmol/l and L_U = 30.5 nmol/l. The robust estimate of median is closed to the estimated corrected mean value (Box-Cox) than the arithmetic mean of the original data batch.

Conclusions

Often, biochemical data are less ideal and do not fulfill all basic assumptions. Original data need to be transformed to improve symmetry of data distribution and achieve a variance stabilization. Statistical measures of transformed data are re-transformed to obtain unbiased and rigorous measures for original data.

Acknowledgements

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

References

1. Tukey JW. Exploratory data analysis. Reading, Massachusetts: Addison Wesley, 1977.
2. Chambers J, Cleveland W, Kleiner W, Tukey P. Graphical methods for data analysis. Boston: Duxbury Press, 1983.
3. Hoaglin DC, Mosteler F, Tukey JW. Exploring data tables, trends and shapes. New York: Wiley, 1985.
4. Silverman BW. Density estimation. London: Chapman and Hall, 1986.
5. Lejenne M, Dodge Y, Koelin E. Proceedings of the Conference COMSTAT'82. Toulouse 1982: 173p, Vol III.
6. Hoaglin DC, Mosteler F, Tukey JW, editors. Understanding robust and exploratory data analysis. New York: Wiley, 1983.
7. Kafander K, Spiegelman CH. An alternative to ordinary QQ plot, Comput Stat Data Anal 1986; 4:167.
8. Hines WGS, Hines RJH. Quick graphical power law transformation selection. Am Statist 1987; 41: 21.
9. Hoaglin DC. Performance of some resistant rules for outlier labeling. J Am Statist Assoc 1986; 81:991.
10. Stoodley K. Applied and computational statistics. Chichester: Ellis Horwood, 1984.
11. Meloun M, Militký J, Forina M. Chemometrics for analytical chemistry. Vol 1. PC-Aided statistical data analysis. Chichester: Ellis Horwood, 1992.
12. Statistical package Adstat 2.0. and QC-Expert, Pardubice: TriloByte Statistical Software, 1999.
13. Hirato K, Yanaihara T. Serum steroid hormone levels in neonates born from the mother with placental sulfatase deficiency. Endocrinol Jpn 1990; 37:731–9.
14. Rabe T, Hosch R, Runnebaum B. Sulfatase deficiency in the human placenta: clinical findings. Biol Res Pregnancy Perinatol 1983; 4:95–102.
15. Shapiro LJ, Cousins L, Fluharty AL, Stevens RL, Kihara H. Steroid sulfatase deficiency. Pediatr Res 1977; 11:894–7.
16. Lykkesfeldt G, Nielsen MD, Lykkesfeldt AE. Placental steroid sulfatase deficiency: biochemical diagnosis and clinical review. Obstet Gynecol 1984; 64:49–54.
17. Parker CR Jr, Buchina ES, Barefoot TK. Abnormal adrenal steroidogenesis in growth-retarded newborn infants. Pediatr Res 1994; 35:633–6.
18. NCSS Statistical Software, 329 North 1000 East, Kaysville, Utah 84037, Email: sales@ncss.com, Tel.: +1(801) 546-0445, Fax: +1(801) 546-3907.

Received 17 December 1999; revised 8 May 2000; accepted 10 May 2000

Corresponding author: Prof. Dr. Milan Meloun, Department of Analytical Chemistry, University Pardubice, CZ-532 10 Pardubice, Czech Republic
Tel.: +4240-603 7026, Fax: +4240-603 7068
Email: milan.meloun@upce.cz