Computer-Assisted Data Treatment in Analytical Chemometrics I. Exploratory Analysis of Univariate Data

[a]M. MELOUN and [b]J. MILITKÝ

[a]*Department of Analytical Chemistry, Faculty of Chemical Technology,*
University Pardubice, CZ-532 10 Pardubice

[b]*Department of Textile Materials, Technical University,*
CZ-461 17 Liberec

Received 19 April 1993

The first step of univariate data analysis called an exploratory data analysis (EDA) isolates certain basic statistical features and patterns of data. EDA is based on the general assumptions as a continuity and differentiability of underlying density. For visualization of data the quantile plot, the dot and jittered dot diagrams, and the box-and-whisker plot are proposed. Peculiarities of sample distribution are investigated by the midsum plot, the symmetry plot, the curtosis plot, and the quantile-box plot. Construction of sample probability density function is carried out by the kernel estimation and the histogram. The quantile-quantile plot serves for comparison of sample distribution with selected theoretical ones. EDA is illustrated on a quantitative chemical analysis of phosphorus content in blood.

No chemist can be unaware of the astonishing developments that have recently taken place in the realm of microelectronics. The rapid growth of chemometrics — the application of the mathematical methods to the solution of chemical problems of all types — is due to the ease with which large quantities of data can be handled, and complex calculations done, with calculators and computers. It is most important for the analytical chemist to remember that the availability of computer-assisted data handling facilities increases rather than decreases the need for a sound knowledge of the principles underlying statistical calculations.

A computer will rapidly perform such calculations, *whatever the data inserted*. The computer will blindly — but efficiently — perform the calculation the user requests. The analyst must thus use his knowledge of statistics to ensure that the appropriate calculation is performed. In this series we want to introduce some useful methods and new software of statistical data treatment which should bring benefits to every chemist.

Exploratory data analysis (EDA) provides the first contact with the data and serves to isolate certain basic statistical features and patterns of data. According to *Tukey* [1] EDA is a "detective work". It uses various descriptive and graphically oriented techniques which are typically free of strict statistical assumptions about data [2, 3]. These techniques are often called "distribution-free". EDA techniques are quite effective for investigation of statistical behaviour of data from new or nonstandard analytical procedures.

This paper brings a description of efficient diagnostic displays and plots of exploratory analysis of univariate data in software ADSTAT. Procedure of systematic investigation of data is illustrated on a quantitative chemical analysis of phosphorus content in blood.

THEORETICAL

Observations as results of chemical experiments are the *random quantities*. The complete collection of all possible outcomes from a chemical experiment in question is called the *population* and observations represent points in this population [4]. When only one variable is recorded, then all the observations create *univariate sample*. The sample values $x_{1,}$ ..., x_n can be sorted in order of ascending magnitude, $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ and values $x_{(1)}, x_{(2)}, ..., x_{(n)}$ are called the *order statistics* of the sample $x_1, x_2, ..., x_n$. We can define the *rank* of an observation in either of two ways [3, 5]: we may count up from the smallest value, or count down from the largest. The first of these yields is the *upward rank* of observation

$$R_{P_i} = i$$

and counting down from the largest value the *downward rank* of observation

$$K_{P_i} = n + 1 - i$$

Considering both rankings, we can see that for any data value it is valid that

$$R_{P_i} + K_{P_i} = n + 1$$

The *depth* of the *i*-th element in an ordered sample is the smaller of its upward rank and its downward rank

$$H_i = \min(R_{P_i}, K_{P_i})$$

and expresses how far it is from the low or high end of the sample.

It can be proved that order statistic $x_{(i)}$ is rough estimator of sample quantile $\tilde{x}_{P_i}$. Denote that under quantile $\tilde{x}_{P_i}$ the $100P_i$ % of the sample values lie and the parameter P_i is here the *cumulative* or *rank probability* given by

$$P_i = \frac{i}{n+1} \tag{1}$$

For a normal distribution the expression

$$P_i = \frac{i-3/8}{n+1/4} \tag{2}$$

is often used, but the EDA uses

$$P_i = \frac{i-1/3}{n+1/3} \tag{3}$$

The plot of order statistic $x_{(i)}$ against the cumulative probability P_i, for $i = 1, ..., n$ is the estimator of the *quantile function* $Q(P)$, *cf.* Ref. [6]. This is, in fact, an inverse function of the sample distribution function. For any value α from the interval $\langle 0, 1\rangle$ the 100α-th quantile $\tilde{x}_\alpha$ may be calculated by linear interpolation

$$\tilde{x}_\alpha = (n+1)\left(\alpha - \frac{i}{n+1}\right)(x_{(i+1)} - x_{(i)}) + x_{(i)} \tag{4}$$

where

$$\frac{i}{n+1} \le \alpha \le \frac{i+1}{n+1} \tag{5}$$

Quantiles and Letter Values

Some methods of EDA are based on some selected quantiles Q being calculated for selected cumulative probabilities $P_i = 2^{-i}$, $i = 1, 2, \ldots$. These quantiles are termed the *letter values* [5] (Table 1), where the symbol u_{P_i} denotes the quantile of standard normal distribution $N(0, 1)$. Except median ($i = 1$), for each $i > 1$ there is a pair of extreme quantiles, the lower (Q_L) and upper (Q_U) letter value. The lower letter value is calculated for a cumulative probability $P_i = 2^{-i}$ while for the upper one $P_i = 1 - 2^{-i}$ (Fig. 1).

Table 1. A Survey of Some Letter Values

i	*i*-th quantile	Cumulative probability for lower one	Symbol for letter value	Normal quantile u_{P_i}
1	Median	$2^{-1} = 0.500$	*M*	0
2	Quartiles	$2^{-2} = 0.250$	*F*	– 0.674
3	Octiles	$2^{-3} = 0.125$	*E*	– 1.15
4	Sedeciles	$2^{-4} = 0.0625$	*D*	– 1.53

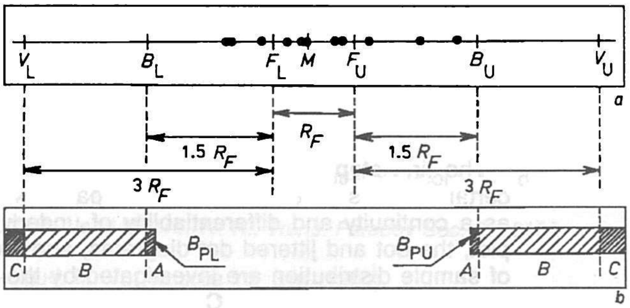

Fig. 1. Construction of the dot diagram with letter values indicating also outliers: *a*) the dot diagram with median *M*, F_L (lower) and F_U (upper) quartiles, inner B_L (lower) and B_U (upper) bounds, outer V_L (lower) and V_U (upper) bounds; *b*) the area of outliers: *A* close outliers, *B* near far outliers, *C* far outliers.

To estimate the letter value, the technique of the *rank-and-depth* is often used [5]. The depth of median is defined by

$$H_M = \frac{n+1}{2} \tag{6}$$

If H_M is an integer, the median is equal to $\tilde{x}_{0.5} = x_{(H)}$, otherwise linear interpolation between two values, $x_{(n/2)}$ and $x_{(n/2+1)}$, is applied. The depth of lower letter values is calculated by

$$H_Q = \frac{1+\mathrm{int}(H_{Q-1})}{2} \tag{7}$$

where *Q* stands for letters *F, E, D, ...* and int(*x*) means the integer part of a number *x*. If *Q* = *F*, then *Q* – 1 = *M* is used. When H_Q is an integer, then the lower quantile Q_L is $x_{(H_Q)}$, while the upper quantile Q_U is $x_{(n+1-H_Q)}$. When H_Q is not integer, the following linear interpolation is carried out

$$Q_L = \frac{x_{(\mathrm{int}(H_Q))} + x_{(\mathrm{int}(H_Q)+1)}}{2} \tag{8}$$

$$Q_U = \frac{x_{(n+1-\mathrm{int}(H_Q))} + x_{(n+2-\mathrm{int}(H_Q))}}{2} \tag{9}$$

For lower values H_Q and quantiles near values of $x_{(1)}$ and $x_{(n)}$ the procedure based on eqns (*8, 9*) is more robust than the application of eqn (*4*).

COMPUTATION

EDA Plots and Displays

Features and statistical properties of a data sample are described by a symmetry and peakedness of the sample distribution, local concentration of data and a presence of outliers. The various exploratory plots and displays offer such information.

Quantile plot [2] enables an identification of the peculiarities of shape of sample distribution which can be symmetrical, skewed to higher or lower values. To compare an actual sample distribution with the normal one, the quantile function of the normal distribution $Q(u_{P_i}) = \mu + \sigma u_{P_i}$ for $0 \langle P_i \langle 1$ is plotted:

1. Classical estimators of μ and σ^2, *i.e.* $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = s^2$ are used where $\bar{x}$ is the arithmetic mean and s^2 is the estimate of variance $\hat{\sigma}^2$.

2. Robust estimators of μ and σ^2, $\hat{\mu} = \tilde{x}_{0.5}$ and $\hat{\sigma}^2 = (R_F/1.349)^2$ are used where $\tilde{x}_{0.5}$ is the median and R_F is the interquartile range, $R_F = F_U - F_L$.

Dot diagram [1] is a one-dimensional scatter plot of data and represents a univariate projection of the quantile plot into the *x*-axis. The dot diagram simply shows a local concentration of data, outliers, and extremes in data.

Jittered dot diagram [2] is similar to the dot diagram and also represents a univariate projection of the quantile plot but sample points are randomly spread along the *y*-axis. Therefore this diagram enables more demonstrative view into local concentration of points.

Box-and-whisker plot [1] brings the overview of letter values in the form of median, two quartiles (hinges) and two extremes. This plot enables *a*) determination of a robust estimate of median *M*, *b*) illustration of a spread and skewness of the sample, *c*) examination of a symmetry and length of distribution tails, and *d*) identification of outliers. The letter values are here interpreted graphically. The box-and-whisker plot has a length R_F from the lower F_L to upper F_U quartile

$$R_F = F_U - F_L$$

and its width is proportional to the value $(n)^{1/2}$.The position of the median is marked by a vertical crossbar inside the box. This plot is useful in illustrating skewness of a sample. If the distribution has a long tail to the right (*positive skew*), the right-hand section of the box will be longer than the left one, and the upper extreme point will be further from the median than the lower extreme. The converse will be true if the distribution has *negative skew* with its longer tail to the left. We use modified box-and-whisker plot where whiskers are terminated by the adjacent values B_U and B_L. Adjacent values lie within the *inner bounds* nearest to their boundary values B_U and B_L being expressed by

$$B_U = F_U + 1.5\,R_F \qquad (10a)$$

$$B_L = F_L - 1.5\,R_F \qquad (10b)$$

Values outside inner bounds but within the *outer bounds* V_U and V_L being expressed by

$$V_U = F_U + 3\,R_F \qquad (11a)$$

$$V_L = F_L - 3\,R_F \qquad (11b)$$

are called the *near far outliers*. Observations outside inner bounds, smaller than V_L or larger than V_U, are called the *far outliers*, and are marked on this plot by the circle.

Notched box-and-whisker plot [2] also enables an examination of the variability of median which is expressed by notches given by the robust confidence interval $I_L \le M \le I_U$, where the lower and upper limits are

$$I_L = M - 1.57\,R_F/(n)^{1/2} \qquad (12a)$$

$$I_U = M + 1.57\,R_F/(n)^{1/2} \qquad (12b)$$

The notches I_L and I_U are located symmetrically around the median. The properties of the notched box-and-whisker plot are the same as in the previous plot.

EDA Diagnostics of Distribution Shape

Main statistical features of the sample distribution are represented by the asymmetry and tails length in comparison with normal (Gaussian) one. The skewness and peakedness can be characterized at different distances from the median by the following statistical diagnostics based on quantiles:

the midsum $Z_Q = (Q_L + Q_U)/2$,
the interquantile range $R_Q = Q_U - Q_L$,
the skewness $S_Q = (M - Z_Q)/R_Q$,
the pseudosigma $G_Q = R_Q\,/(-\,2u_{P_i})$,
the tails length $T_Q = \ln\,(R_Q\,/R_F)$,

where *Q* stands for the letter value and u_{P_i} is the quantile of standardized normal distribution for $P_i = 2^{-i}$. These diagnostics are summarized in Table 2.

For selected symmetric distribution the theoretical length of tails, T_E and T_D, is computed: for the normal distribution $T_E = 0.534$ and $T_D = 0.822$, for the rectangular distribution 0.405 and 0.559, and for the Laplace distribution 0.693 and 1.098.

To examine all statistical features of the sample, various plots of characteristics from Table 2 are used. For large samples, the letter values are examined, whereas for small samples the quantile $\tilde{x}_{P_i} = x_{(i)}$ usually for $P_i = (i - 1/3)/(n + 1/3)$ is used.

Table 2. Diagnostics of Distribution Shape

Characteristic	Used for exploration of	Valid for L
Midsum Z_Q	a symmetry (at $Z_Q = 0$)	$F, E, D, ...$
Interquantile range R_Q	a spread	$F, E, D, ...$
Skewness S_Q	a symmetry (at $S_Q = 0$)	$F, E, D, ...$
Pseudosigma G_Q	a peakedness (for Gaussian distribution $G_Q = 0$)	$F, E, D, ...$
Tails length T_Q	a peakedness	$E, D, ...$

Midsum plot [2] indicates a symmetry of distribution. It has on the x-axis the order statistic $x_{(i)}$ and on the y-axis the midsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2$. For symmetrical distribution the midsum plot forms a horizontal line $y = \tilde{x}_{0.5}$.

Quantile-box plot [6] for examining the statistical features of data is based on the estimate of a sample quantile function formed connecting points $\{x_{(i)}, P_i\}$ by straight lines. It has on the x-axis the order probability P_i and on the y-axis the order statistic $x_{(i)}$, where P_i is calculated by $P_i = (i - 1/3)/(n + 1/3)$. For symmetrical distributions, the sample quantile function exhibits a sigmoid shape, whereas for an asymmetrical one, the quantile function is convex or concave increasing. For easier interpretation the following quantile boxes are on the graph:

a) *The quartile box F* has on the y-axis two vertices given by quartiles F_L and F_U with corresponding values on the x-axis equal to the cumulative probability values $P_2 = 2^{-2} = 0.25$ and $1 - 2^{-2} = 0.75$.

b) *The octiles box E* has on the y-axis octiles E_L and E_U and on the x-axis the cumulative probabilities $P_3 = 2^{-3} = 0.125$ and $1 - 2^{-3} = 0.875$.

c) *The sedeciles box D* has on the y-axis sedeciles D_L and D_U and on the x-axis the cumulative probabilities $P_4 = 2^{-4} = 0.0625$ and $1 - 2^{-4} = 0.9375$.

The position of *the median M* is marked by a horizontal line inside the quartile box. Robust estimate of median confidence interval, $M \pm 1.57\ R_F/(n)^{1/2}$, is drawn as a vertical line at $P_i = 0.5$. On the basis of this plot, the following statistical features of the sample distribution may be stated [6]:

a) *The symmetric unimodal sample distribution* contains individual boxes arranged symmetrically inside themselves and the value of relative skewness is close to zero, $S_Q \approx 0$.

b) *The asymmetric sample distribution*: in the case of a distribution skewed to higher values, there are significantly shorter distances between the lower parts of the boxes when compared with those between the upper ones. The skewness S_Q then has a negative value. For a distribution skewed to lower values the skewness S_Q is positive.

c) *Outliers* are indicated by a sudden increase of the quantile function outside the F box, the slope may approach infinity limit.

d) *A multimodal sample distribution* is indicated by several parts of the quantile function inside box F reaching zero slope.

Analysis of Sample Distribution

Some graphical displays can show overall patterns or trends. They can also reveal surprising, unexpected, or amusing features of data that might otherwise go unnoticed. When a large number of observations is available, the estimation of *probability density function* or other function characterizing the data distribution can help to elucidate the structure of the sample.

Kernel estimation of the sample probability density function $\hat{f}(x)$ for small and medium samples may be calculated by the relation

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left[\frac{x - x_i}{h}\right] \qquad (13)$$

where h is bandwidth which controls a smoothness of $\hat{f}(x)$ and $K(x)$ is the kernel function. The kernel function $K(x)$ is symmetric around zero and has properties of a probability density function. We consider biquadratic kernel estimate

$$K(x) = \begin{Bmatrix} 0.9375(1 - x^2)^2 & \text{for } -1 \le x \le 1 \\ 0 & \text{for } x \text{ outside } \langle -1;\ 1 \rangle \end{Bmatrix} \qquad (14)$$

The quality of the kernel estimate $\hat{f}(x)$ is controlled by the choice of parameter h. If h is too small, the estimate is rough; if it is too large, the shape $\hat{f}(x)$ is flattened too much. For samples taken from normal distribution the optimal bandwidth h can be calculated by an expression suggested by *Scott* and *Sheater* [7]

$$h_{\text{opt}} = 2.34\ \sigma n^{-0.2} \qquad (15)$$

Selection of optimal h for EDA purposes is described by *Lejenne, Dodge,* and *Koelin* [8].

Histogram is one of the oldest classical representations of grouped frequency distributions. The vertical axis represents roughly the class frequency, and the class mid-values x_i , $i = 1, ..., k$, are plotted on the horizontal axis. With the class mid-value x_i as the centre of its base, a vertical bar of base equal to the class width and height equal to an empirical relative frequency f_i, is erected for each of the classes.

Quantile-quantile plot or the **Q-Q plot** [3, 5] allows comparison of the sample distribution being described by the empirical $Q_E(P_i)$ quantile function with the given theoretical one, with the theoretical $Q_T(P_i)$ quantile function. The values of empirical $Q_E(P_i)$ function are approximated by the sample order statistic $x_{(i)}$. If there is close agreement between sample and theoretical distributions, it must be true that

$$x_{(i)} \simeq Q_T(P_i) \qquad (16)$$

where P_i is the cumulative probability chosen as $P_i = (i - 1/3)/(n + 1/3)$. When the empirical sample distribution is the same as the theoretical one, the resulting *Q-Q* plot is represented by the straight line. Utilization of eqn (*16*) needs knowledge of all parameters for theoretical quantile functions $Q_T(P_i)$. Some theoretical distributions can be rewritten to the form

$$Q_T(P_i) = \mu + \sigma Q_{TS}(P_i) \qquad (17)$$

Here μ is usually the location parameter, σ is the spread parameter, and $Q_{TS}(P_i)$ is the standardized quantile function. For most two-parametric distributions the $Q_{TS}(P_i)$ is free of adjustable parameters. For some three-parametric distributions the shape factor is usually a parameter of the plot. The standardized quantile function $Q_T(P_i)$ is therefore used for practical construction of the quantile-quantile plot. For selected theoretical distributions the *x* and *y* coordinates of *Q-Q* graph are given in Table 3.

In Table 3 the normal distribution function $\Phi(s)$ is defined as

$$\Phi(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} \exp(-0.5\, u^2)\, du$$

For calculation of inverse function $\Phi^{-1}(P_i)$ the simple approximate relation may be used

$$\Phi^{-1}(P_i) = \frac{-9.4 \ln\left(\frac{1}{P_i} - 1\right)}{\mathrm{abs}\left(\ln\left(\frac{1}{P_i} - 1\right)\right) + 14}$$

Due to the strong dependence among order statistics $x_{(i)}$ and their nonconstant variance, the quantile-quantile plot for small samples has a very patterned appearance. Construction of improved quantile-quantile plot is described by *Kafander* and *Spiegelman* [9].

Rankit plot or the **normal-probability plot** enables classification of sample distribution according to its skewness, peakedness and tails length. A convex or concave shape indicates a skewed sample distribution. A sigmoid shape indicates that the tails length of the sample distribution differs from those of the normal one.

PROCEDURE

For creation of EDA diagnostic graphs and computation of quantile-based characteristics of sample distribution the module basic statistics of ADSTAT package is used, *cf.* Ref. [10]. For graphical visualization of data, five diagrams and simple plots, *i.e.* the quantile plot, the dot and jittered dot diagrams, box-and-whisker plot and notched box-and-whisker plot are supported. Sample distribution represented by a symmetry and tail lengths, skewness and peakedness is investigated by two plots, *i.e.* by the midsum plot and the quantile-box plot. Construction of sample distribution, *i.e.* the estimation of probability density function is carried out by the kernel estimation of probability density function, and by the histogram. The quantile-quantile plot and the rankit plot are used for comparison of sample distribution with the theoretical ones.

Table 3. Standardized Probability Density $f_T(s)$ and Distribution $F_T(s)$ Functions, and Corresponding Coordinates (*x*, *y*) of the *Q-Q* Plot

Distribution	$F_T(s)$	$f_T(s)$	*y*	*x*
Rectangular	s	1	$x_{(i)}$	P_i
Exponential	$1 - \exp(-s)$	$\exp(-s)$	$x_{(i)}$	$-\ln(1 - P_i)$
Normal	$\Phi(s)$	$(2\pi)^{-1/2} \exp(-0.5\, s^2)$	$x_{(i)}$	$\Phi^{-1}(P_i)$
Laplace, $x \le 0$	$0.5 \exp(s)$	$0.5 \exp(s)$	$x_{(i)}$	$\ln(2P_i)$, for $P_i \le 0.5$
Laplace, $x > 0$	$0.5(2 - \exp(-s))$	$0.5 \exp(-s)$	$x_{(i)}$	$-\ln(2(1 - P_i))$, for $P_i > 0.5$
Log-normal	$\Phi[\ln(s)]$	$(2\pi)^{-1/2} \exp(-0.5 \ln s^2)$	$x_{(i)}$	$\exp[\Phi^{-1}(P_i)]$

RESULTS

Study Case 1. *Use of EDA in the determination of phosphorus in blood*

A random sample of fifty milk cows was taken from of a herd of 2900 cows, and the blood of each was analyzed for phosphorus content in mmoles per dm^3. EDA estimates the sample distribution, and proposes whether the measures of location and spread should be computed from classical moment or robust quantile estimators.

Data: phosphorus content (c/(mmol dm^{-3})) for sample size n = 50; 2.17, 2.28, 2.17, 1.92, 2.21, 1.43,

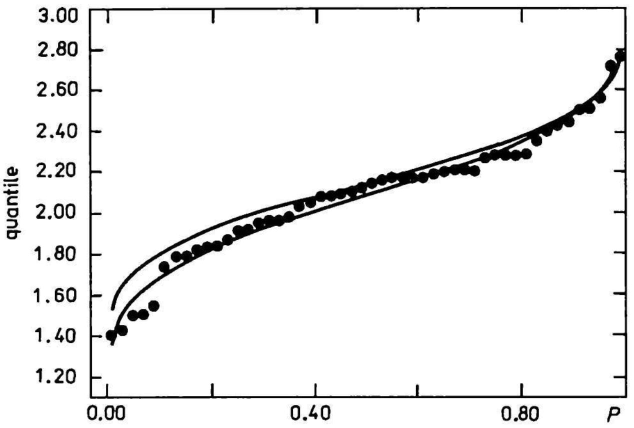

Fig. 2. The quantile plot for Study Case 1.

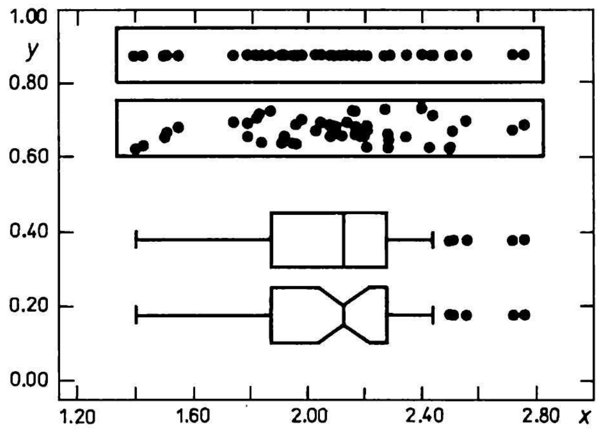

Fig. 3. The dot and jittered dot diagram, the box-and-whisker plot, and the notched box-and-whisker plot for Study Case 1.

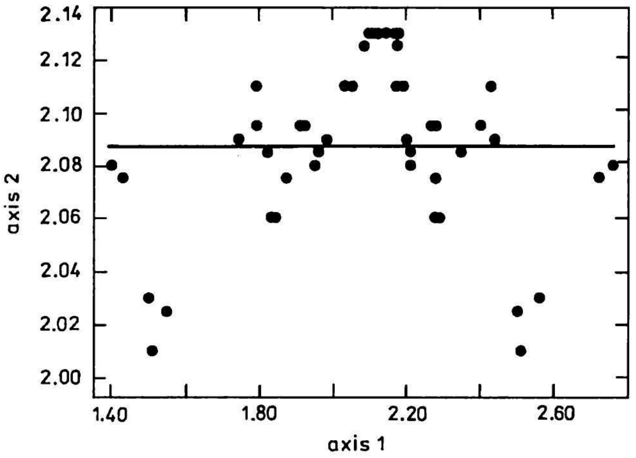

Fig. 4. The midsum plot for Study Case 1.

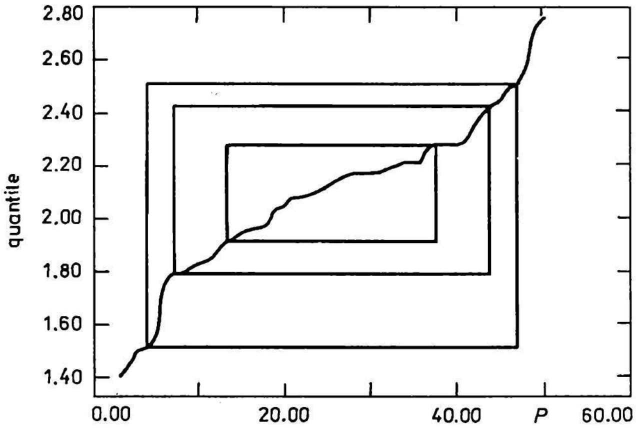

Fig. 5. The quantile-box plot for Study Case 1.

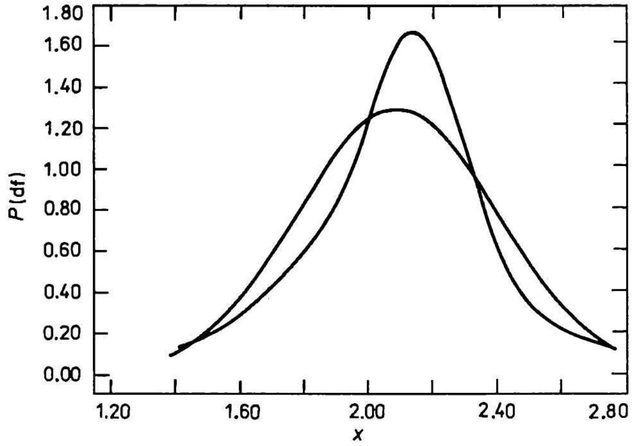

Fig. 6. The kernel estimate of probability density function for Study Case 1.

2.21, 2.17, 2.72, 2.03, 2.56, 2.16, 1.40, 1.83, 2.08, 2.76, 1.91, 2.43, 1.96, 2.28, 2.51, 1.87, 1.55, 2.40, 1.95, 2.50, 1.82, 1.74, 2.19, 2.08, 2.27, 2.05, 2.20, 2.10, 1.96, 2.35, 1.51, 2.17, 2.14, 2.21, 2.29, 1.98, 2.28, 1.50, 2.09, 2.44, 1.79, 1.84, 2.12, 1.79.

Solution: The diagnostic graphs of EDA are used as follows. The quantile plot (Fig. 2) shows small deviations from the normal distribution, especially at low values. The jittered diagrams and the box-and-whisker plots (Fig. 3) indicate that five low measurements and two high measurements differ from the rest of the sample. The midsum plot (Fig. 4) indicates that distribution is skewed to lower values, and both skewness and peakedness (kurtosis) differ from the expected values.

The quantile-box plot (Fig. 5) shows asymmetry of distribution and two lowest and two highest values as outliers.

The nonparametric kernel estimate of probability density function (Fig. 6) indicates that the distribution is skewed to lower values in comparison with the normal distribution. Therefore the mode $\tilde{x}_{\mathrm{mod}}$ is also shifted from the arithmetic mean $\bar{x}$. The quantile-quantile plot (Fig. 7) proves that there is a significant separation of five lowest and two highest values of the sample.

Numerical values of quantile measures of location and spread in Table 4 show that *a*) there is asymmetric skewing which is largest in the quartile range, *b*) the five lowest and two highest values are outliers.

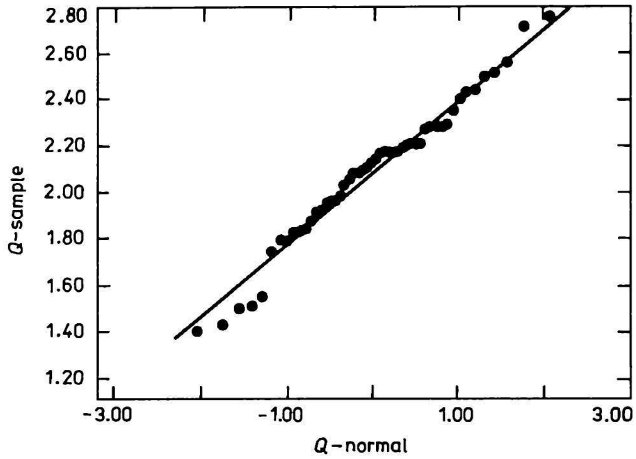

Fig. 7. The quantile-quantile (*Q*-*Q*) plot for Study Case 1.

Table 4. *a*) The Quantile Measures of Location

Quantile	*P*	Lower quantile	Upper quantile	Range
Median	0.5	2.1500	2.1500	–
Quartile	0.25	1.9125	2.2775	0.3650
Octile	0.125	1.7900	2.4263	0.6363
Sedecile	0.0625	1.5125	2.5094	0.9969

b) The Quantile Measures of Spread and Shape

Quantile	*P*	Midsum	Skewness	Tails length of the sample distribution	Tails length of the normal distribution
Quartile	0.25	2.0950	5.09	0.000	0.000
Octile	0.125	2.1081	2.91	1.025	0.556
Sedecile	0.0625	2.0109	1.81	0.125	1.005

The sample batch exhibits deviations from normality and has significant influence on the measures of location and spread. The robust quantile estimators are more suitable for these data.

CONCLUSION

The first step in effective analysis of the univariate data is an *exploratory data analysis* (*EDA*) when the data for uncovering typical features and patterns are surveyed and treated. The second step is a *confirmatory data analysis* (*CDA*) where probability models are created and tested. EDA being also called a detective work uses tools of descriptive and graphically oriented techniques known as the distribution-free methods. EDA is very effective for an investigation of statistical behaviour of experimental data coming from new or nonstandard analytical techniques.

REFERENCES

1. Tukey, J. W., *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts, 1977.
2. Chambers, J., Cleveland, W., Kleiner, W., and Tukey, P., *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*. Wiley, New York, 1985.
4. Stoodley, K., *Applied and Computational Statistics*. Ellis Horwood, Chichester, 1984.
5. Hoaglin, D. C., Mosteler, F., and Tukey, J. W. (Editors), *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, 1983.
6. Parzen, E,, *J. Am. Statist. Assoc. 74*, 105 (1985).
7. Scott, D. W. and Sheater, S. J., *Commun. Statist. 14*, 1353 (1985).
8. Lejenne, M., Dodge, Y., and Koelin, E., *Proceedings of the Conference COMSTAT '82, Toulouse*. P. 173 (Vol. *III*).
9. Kafander, K. and Spiegelman, C. H., *Comput. Stat. Data Anal. 4*, 167 (1986).
10. Meloun, M., Militký, J., and Forina, M., *Chemometrics for Analytical Chemistry,* Vol. *1. PC-Aided Statistical Data Analysis*. Ellis Horwood, Chichester, 1991.Vol. *2. PC-Aided Regression and Related Methods.* Ellis Horwood, Chichester, 1994.

Translated by M. Meloun