Analytica Chimica Acta, 277 (1993) 215–221
Elsevier Science Publishers B.V., Amsterdam

Some graphical aids for univariate exploratory data analysis

Jiří Militký

Department of Textile Materials, Technical University, 46117 Liberec (Czech Republic)

Milan Meloun

Department of Analytical Chemistry, Technical University, 53210 Pardubice (Czech Republic)

(Received 30th June 1992)

Abstract

The main parts of exploratory data analysis (EDA) are discussed. For data presentation the quantile plot and quantile-box plot are proposed. Special techniques for empirical probability density construction and empirical quantile–quantile plot creation are described. Some graphically oriented methods for selection of optimum power transformations are presented. These graphical aids in EDA are demonstrated on Hinkley's well known data.

Keywords: Exploratory data analysis; Power transformation; Probability density; Univariate data analysis

The classical approach to statistical data analysis is based on some strong assumptions about their statistical nature such as independence, normality and homogeneity. Frequently the data are less than ideal and their effective analysis can be realized in two stages. The first stage is exploratory data analysis (EDA), where data for uncovering typical relationships and patterns are surveyed and treated. The second stage is confirmatory data analysis (CDA), where probability models are created and tested.

According to Tukey [1], EDA is "detective work". It uses as its tools various descriptive and graphically oriented techniques that are typically free from strict statistical assumptions about data. These techniques are often called "distribution free" and are based only on basic assumptions such as continuity and differentiability of underlying density. EDA techniques are especially effective for the investigation of the statistical behaviour of data from new or non-standard measurements or for the creation of probability models. A typical chemometric example is trace analysis.

One of most frequent tasks in statistical data analysis is the one-sample problem based on a sample $x_1, \ldots, x_n$ representing the behaviour of a univariate (random) variable x. For this case EDA has three main goals: visualization of statistical features of the sample; construction of an empirical sample distribution and comparison of this distribution with theoretical ones; and data transformation for improving their distribution such as symmetrizing or normalization.

The realization of these goals involves the utilization of techniques well suited especially for

Correspondence to: J. Militký, Department of Textile Materials, Technical University, 46117 Liberec (Czech Republic).

small and moderate sample sizes [2]. The most popular EDA methods for one-sample problems with applications in chemistry have been surveyed [3].

In this paper, selected simple EDA techniques are discussed. The full set of EDA analysis techniques used in the module Basic Statistics in the package ADSTAT has been described elsewhere [3]. Some EDA techniques are demonstrated on Hinkley's well known data (sample of 30 values) [4].

SOME BASIC CONCEPTS

The EDA techniques for small and moderate samples are based on order statistics:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

which are the sample values (assumed to be distinct) arranged in increasing order. Let $F_e(x)$ be the distribution function from which values x_i are sampled. It is well known that the transformed random variable

$$Z_{(i)} = F_e[x_{(i)}] \tag{1}$$

has independently of the distribution function F_e the beta distribution $Be\ [i, n-i+1]$. Its mean value is given by

$$E[Z_{(i)}] = i/(n+1) \tag{2}$$

The elements V_{ij} of the covariance matrix V for all $Z_{(i)}(i = 1, \ldots, n)$ are simple functions of i, j and n only. Using back transformations of $E[Z_{(i)}]$, we obtain the relationship

$$E[x_{(i)}] = F_e^{-1}[Z_{(i)}] = Q_e(P_i) \tag{3}$$

where $Q_e(P_i)$ is a quantile function and $P_i = i/(n+1)$ is the cumulative probability. A detailed description of the quantile function properties and its advantages for constructing empirical sample distributions was given by Parzen [5].

From Eqn. 3 we can deduce that the order statistic $x_{(i)}$ is a raw estimate of the quantile function $Q_e(P_i)$ in the position of P_i. For estimation of the quantile $x_P \equiv Q_e(P)$ at a value $i/(n+1) < P < (i+1)/(n+1)$, the piecewise linear interpolation

$$x_P = (n+1)(P - i/(n+1))[x_{(i+1)} - x_{(i)}] + x_{(i)} \tag{4}$$

can be used. The variance $D(x_P)$ can be calculated from the equation

$$D(x_P) = P(1-P)/[nf_e^2(x_P)] \tag{5}$$

where $f_e(x_P)$ is a probability density function corresponding to the distribution function F_e.

Equation 4 can be used for estimation of sample quantiles x_{P_i} or x_{1-P_i} for $P_i = 2^{-i}$ $(i = 1, \ldots, n)$. These quantiles are called letter values [6]. All letter values except $i = 1$ (median) are in pairs. For example, we can estimate the lower quartile $x_{0.25}$ $(P_i = 0.25)$ and upper quartile $x_{0.75}$ $(P_i = 0.75)$, etc.

For EDA purposes the modified definition of cumulative probability

$$P_i = (i - 0.375)/(n + 0.25) \tag{6}$$

is often used. A discussion of a suitable definition of P_i was presented by Looney and Gulledge [7].

TECHNIQUES FOR DATA PRESENTATION

For the graphical representation of data, many simple techniques such as the stem-leaf plot, box plot, dot plot [1] and digdot plot [8] have been proposed. The quantile-box plot (QBP) and quantile plot (QP) are selected here. Symmetry and tail length can be characterized by use of the g–h distribution system [3,10].

Quantile plot

An empirical sample quantile plot $Q(P)$ is constructed as a dependence of $x_{(i)}$ on P_i. From patterns of points some statistical features of data such as symmetry, local concentration and rough normality can be simply recovered. A detailed interpretation of QP has been given elsewhere [9].

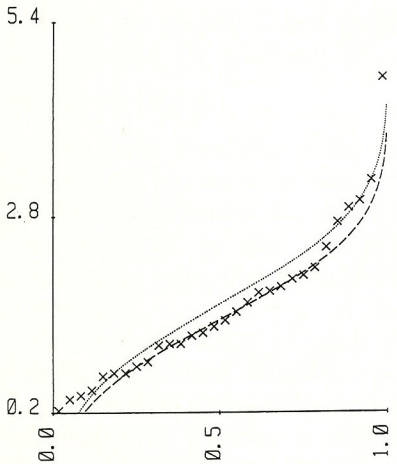

Fig. 1. Quantile plot for Hinkley's data. ———, Robust estimate; ·····, classical estimate.

For comparative purposes, the quantile functions of a normal distribution

$$Q_N(P) = \mu + \sigma u_P \tag{7}$$

where u_P are quantiles of the standard normal distribution $N(0, 1)$ and μ and σ are estimators of location and scale, are superimposed on QP.

Two different normal quantile functions are plotted. One is based on the sample mean x_M and sample standard deviation s. The other uses robust quantile estimators $\mu = x_{0.5}$ and $\sigma = (x_{0.75} - x_{0.25})/1.349$.

Figure 1 shows the quantile plot for Hinkley's data. It is clear that these data are positively skewed and can be approximated in the middle part by a normal distribution (with robust estimators of location and scale).

Quantile-box plot

A quantile-box plot (QBP), proposed by Parzen [5], is an extension of the idea of a box plot introduced by Tukey [1]. A QBP consists of quantile function graph (see previous section) on which various boxes with vertices

$$[(P, Q(P)); (P, Q(1-P)); (1-P, Q(P)); (1-P, Q(1-P))]$$

are superposed. One usually chooses quartile ($P = 1/4$), octile ($P = 1/8$) and sedecile ($P = 1/16$) boxes.

TABLE 1

Theoretical tail lengths

Distribution	T_3	T_4
Normal	0.534	0.822
Rectangular	0.405	0.559
Laplacian	0.693	1.098

Within the quartile box one draws a median line with the vertices

$$[(0.25, Q(0.5)), (0.75, Q(0.5))]$$

An approximate confidence interval for the median $x_{0.5} \equiv Q(0.5)$ is indicated by a vertical line with vertices

$$[(0.5, x_{0.5} \pm 1.57(x_{0.75} - x_{0.25})/\sigma)]$$

The QBP with some quantile measures of symmetry such as [5]

$$S_i = [x_{0.5} - 0.5(x_{P_i} + x_{1-P_i})]/(x_{1-P_i} - x_{P_i}) \tag{8}$$

($i = 2, 3, 4$) and tail length [5]

$$T_i = \ln[(x_{1-P_i} - x_{P_i})/(x_{0.75} - x_{0.25})] \tag{9}$$

($i = 3, 4$) can be used for the description of data peculiarities at various distances from median. The theoretical T_i values for octiles ($i = 3$) and sedeciles ($i = 4$) are presented in Table 1.

As described [5], the QBP can also be used for the identification of polymodal distributions and outliers.

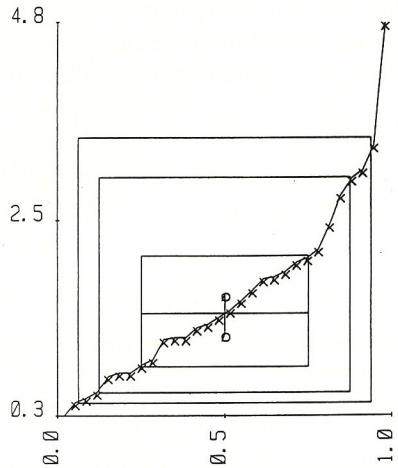

Fig. 2. Quatile-box plot for Hinkley's data.

Figure 2 shows the QBP for Hinkley's data. Corresponding values of S_i and T_i are presented in Table 2. It is evident that the data are positively skewed and contain one outlying observation.

CONSTRUCTION AND COMPARISON OF SAMPLE DISTRIBUTION

As an estimator of the empirical probability density function, histograms with variable bins and kernel type density are constructed. Comparison of the sample distribution with theoretical distribution is based on variants of the Quantile–Quantile (Q–Q) plot. The probability–probability and transformed distribution function plots [5] can also be used [11].

Empirical probability density

A histogram is a piecewise constant estimator of sample probability density (PDF). The histogram height in the jth class bounded by values (t_{j-1}, t_j) is calculated from the relationship

$$f_H(x) = (nh_j)^{-1} C_n(t_{j-1}, t_j) \qquad (10)$$

where the function $C_n(a, b)$ denote the number of sample elements within $\langle a, b \rangle$ and $h_j = t_j - t_{j-1}$ is the length of the jth interval. Now, the problem encountered is the choice of boundary values $\{t_j\}(j = 1, \ldots, M)$, the number of class intervals M and their lengths h_j with respect to the histogram quality. In our programs the simple data-based two-stage technique is used.

In the first stage, the number of class intervals

$$M_0 \approx \text{int}\left[2.46(n-1)^{0.4}\right]$$

is assessed. In the second stage, the individual lengths h_j are determined. The estimation of h_j is based on the requirement of equal probability in all classes. For this purpose the empirical quantile function $Q(P)$ based on order statistics $x_{(i)}$ is implemented.

In practice, the P-axis is divided into identical intervals of size $1/M_0$. For these intervals the corresponding quantile estimates $t_j = x(j/M_0)$ are constructed by using Eqn. 4, where $P = j/M_0$. Practical experiences has hitherto demonstrated that this construction is suitable even for strongly skewed sample distributions. Figure 3 shows the histogram with a normal probability density superimposed for Hinkley's data.

The kernel-type non-parametric estimator of sample probability density $f(x)$ can be constructed on the basis of the Lejenne–Dodge–Kaelin procedure [12].

TABLE 2

Characteristics of symmetry and tail length for Hinkley's data

Quantile	i	S_i	T_i
Quartile	2	−0.025	0.000
Octile	3	−0.134	0.712
Sedecile	4	−0.163	0.879

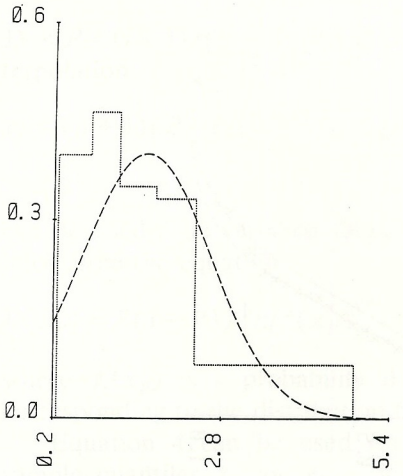

Fig. 3. Histogram for Hinkley's data. — — —, Normal PDF; · · · · · ·, histogram.

Comparison of sample distribution

For the purpose of comparison of empirical sample distributions with theoretical distributions, variants of the Q–Q plot are suitable. The classical Q–Q plot is based on a comparison of the empirical quantile function $Q(P_i) \equiv x_{(i)}$ with a chosen theoretical quantile function $Q_T(P_i)$. For theoretical distribution functions of the type $F_T[(x-\mu)/\sigma]$ it is advantageous to use the standardized quantile function $Q_{TS}(P_i)$.

When the empirical and theoretical distributions coincide, the relationship

$$x_{(i)} = \mu + \sigma Q_{TS}(P_i) \qquad (11)$$

is valid. Here usually μ is the shape parameter and σ is the parameter of scale. For some three-parameter distribution the shape factor is usually a parameter of the plot. Our programs select a shape factor value that straightens the individual points best.

Owing to the strong dependence among order statistics and their non-constant variance, the Q–Q plot gives a very patterned appearance and the degree of linearity is difficult to quantify. Michael [13] introduced the stabilized probability plot and Kafander and Spiegelman [14] proposed the conditional Q–Q plot.

For EDA purposes we use the empirical probability plot (EPP) [2]. In EPP the quantiles $Q_{TS}(P_i)$ are replaced with simulated ones T_i generated from chosen theoretical distribution.

The process of the computation of T_i can be divided into three main steps: from an assumed theoretical distribution the simulated samples $\{x_i^j\}(i = 1, \ldots, n;\ j = 1, \ldots, 25)$ are generated; from all samples $(j = 1, \ldots, 25)$ the order statistics $x_{(i)}^j$ are computed; and the simulated quantiles T_i are medians from corresponding order statistics of all simulated samples:

$$T_i = \text{med}\{x_{(i)}^1, \ldots, x_{(i)}^{25}\}$$

Based on the second largest values and second lowest values in the sequence $\{x_{(i)}^1, \ldots, x_{(i)}^{25}\}$, the boundary of the 85% confidence intervals can be constructed. An analogous procedure for the case of logistic regression has been described [15].

Figure 4 shows the classical Q–Q plot and Fig. 5 shows the EPP plot for Hinkley's data. In both plots the normal distribution is selected as a theoretical distribution. The systematic deviation from linearity indicated a non-normal sample.

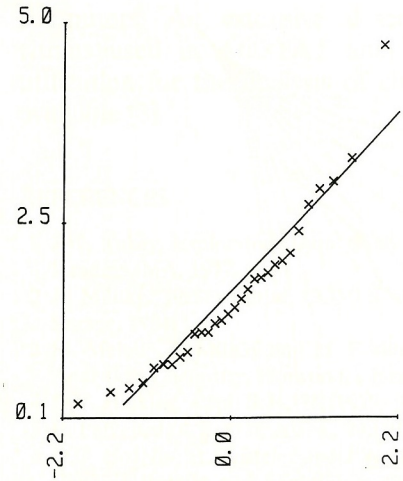

Fig. 4. Q–Q plot for Hinkley's data (theoretical is normal distribution).

POWER TRANSFORMATION OF DATA

Power transformation is used in the context of EDA as a tool for simplifying the data distribution. Suitable power–law transformations may result in a distribution that is nearly symmetrical and perhaps more nearly normal.

In many instances the symmetrizing of the data distribution by using a simple power transformation:

$$\begin{aligned} y = g(x) &= x\hat{}(\beta) && \text{for } \beta > 0 \\ y = g(x) &= \ln x && \text{for } \beta = 0 \\ y = g(x) &= -x\hat{}(-\beta) && \text{for } \beta < 0 \end{aligned} \qquad (12)$$

can be obtained. This transformation is not scale invariant and is not a continuous function of β. It

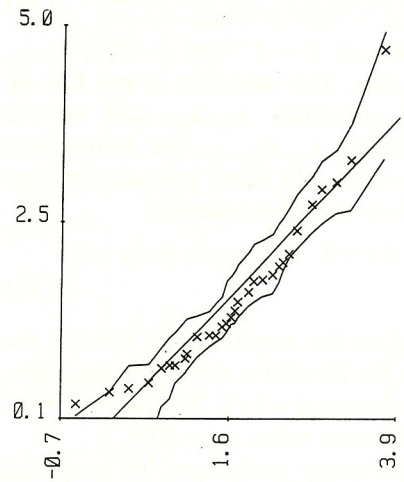

Fig. 5. EPP plot for Hinkley's data (theoretical is normal distribution).

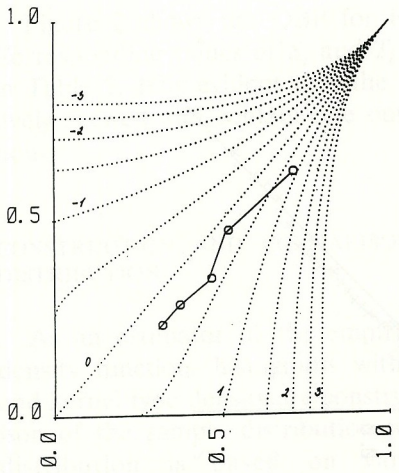

Fig. 6. Selection graph for Hinkley's data.

is suitable for positive data only. Optimum transformation can be selected by minimizing some robust measures of skewness:

$$g_R = [(y_{0.75} - y_{0.5}) - (y_{0.5} - y_{0.25})] / (y_{0.75} - y_{0.25}) \quad (13)$$

As a diagnostic tool a selection graph can be simply constructed. This graph is based on the requirement of symmetry of quantiles about the median. This requirement can be mathematically described by the relationship [16]

$$(x_{P_i}/x_{0.5})^{\wedge}(\beta) + (x_{0.5}/x_{1-P_i})^{\wedge}(-\beta) = 2 \quad (14)$$

Letter values, where $P_i = 2^{-i}$ for $i = 2, 3, 4, \ldots$, are usually chosen. The selection graph has on the y-axis the quantities $x_{P_i}/x_{0.5}$ and on the x-axis the quantities $x_{0.5}/x_{1-P_i}$. For comparison of computed points with ideal courses for constant β, the solution of the equation

$$y^{\wedge}(\beta) + x^{\wedge}(-\beta) = 2 \quad 0 \le x \le 1; 0 \le y \le 1 \quad (15)$$

is superimposed on the graph. Figure 6 shows the selection graph for Hinkley's data. It is clear that it is suitable to choose $\beta \approx 1/3$. The value $\beta = 0.2$ leads to the minimum value $g_R = 0.06$.

Another exploratory technique for graphical estimation of power β was described by Emerson and Stoto [17].

The Box–Cox power transformation family, which is a continuous function of β, can be defined by

$$y = g(x) = [x^{\wedge}(\beta) - 1]/\beta \quad \text{for } \beta \neq 0$$
$$y = g(x) = \ln x \quad \text{for } \beta = 0 \quad (16)$$

This transformation can be used for positive data only. After slight modification the range of applicability can be arbitrarily extended.

The properties of this transformation family have been studied in depth (e.g., [6]). Based on the assumption that for some β, y is a normally distributed variable $N(\mu_y, \sigma_y^2)$, the likelihood function can be constructed. The logarithm of likelihood function has the form

$$\ln L(\beta) = -n/2 \ln(s_y^2) + (\beta - 1) \sum_i \ln x_i \quad (17)$$

where s_y^2 is the sample variance of transformed data. The likelihood function can be plotted against β in a suitable range (the standard range is $-3 \le \beta \le 3$). On this plot the $100(1-\alpha)\%$ confidence interval of power β:

$$2[\ln L(\beta^*) - \ln L(\beta)] \le \chi^2(1) \quad (18)$$

is superimposed, where β^* is the maximum likelihood estimator of β. In the confidence interval defined by Eqn. 18 are all values of β for which $\ln L(\beta) \in \ln L(\beta^*) - 0.5\chi^2(1)$, where χ^2 is a quantile of the χ-squared distribution. From the

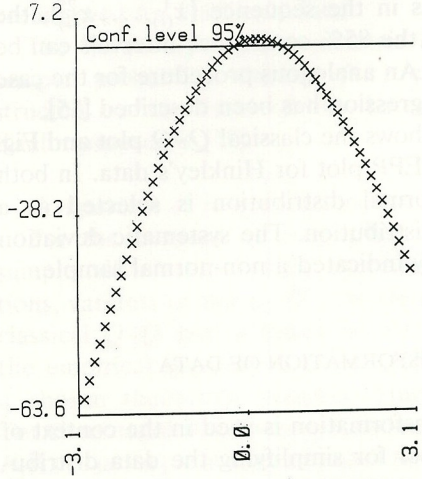

Fig. 7. Likelihood function plot for Hinkley's data.

width of the confidence interval the quality of power transformation can be indicated.

Figure 7 shows the likelihood function plot for Hinkley's data. The optimum power maximizing ln $L(\beta)$ is $\beta = 0.2$.

The quality of power transformation can also be described by using the above-discussed graphical techniques.

PROGRAM SYSTEM ADSTAT

ADSTAT contains eight modules of statistical methods for univariate and multivariate data [3]. Manipulations with ADSTAT are very simple by using pull-down menus and panes. Individual program modules are built in a window-like environment. This environment includes the powerful block-oriented data editor, context-sensitive help system and unified graphical presentation. Exploratory methods included in the module Basic Statistics can be divided to three main parts: techniques for presentation of data; construction of empirical sample distribution and its comparison with twelve theoretical distributions; and power transformation of data by using Eqns. 12 and 16.

The above-mentioned and more complex EDA techniques described elsewhere [3] are used.

Conclusion

The program system ADSTAT is well suited for EDA of one-sample problems on personal computers. An extensive description of algorithms used in ADSTAT and examples of its utilization for the analysis of chemical data are available [3].

REFERENCES

1 J.W. Tukey, Exploratory Data Analysis, Addison Wesley, Reading, MA, 1977.
2 J. Militký, presented at COMPSTAT '84 Conference, Prague, 1984.
3 M. Meloun, J. Militký and M. Forina, Chemometrics for Analytical Chemistry, Horwood, Chichester, 1992.
4 D.V. Hinkley, Appl. Stat., 26 (1977) 67.
5 E. Parzen, J. Am. Stat. Assoc., 74 (1985) 105.
6 D.C. Hoaglin, F. Mosteler and J.W. Tukey (Eds.), Understanding Robust and Exploratory Data Analysis, Wiley, New York, 1983.
7 S.W. Looney and T.R. Gulledge, Staistician, 34 (1985) 297.
8 S. Hunter, Am. Stat., 42 (1988) 54.
9 J. Chambers, W.S. Cleveland, B. Kleiner and P.A. Tukey, Graphical Methods for Data Analysis, Duxbury Press, Boston, 1983.
10 D.C. Hoaglin, F. Mosteler and J.W. Tukey (Eds.), Exploring Data, Tables, Trends and Shapes, Wiley, New York, 1985.
11 J. Militký, presented at COMPSTAT '88 Conference, Copenhagen, 1988.
12 M. Lejenne, Y. Dodge and E. Kaelin, presented at COMPSTAT '82 Conference, Toulouse, 1982.
13 J.R. Michael, Biometrika, 70 (1983) 11.
14 K. Kafander and C.H. Spiegelman, Comput. Stat. Data Anal., 4 (1986) 167.
15 J.M. Landwehr, D. Pregibon and A.C. Shoemaker, J. Am. Stat. Assoc., 79 (1984) 61.
16 W.G.S. Hines and R.J.H. Hines, Am. Stat., 41 (1987) 21.
17 J.D. Emerson and M.A. Stoto, J. Am. Stat. Assoc., 77 (1982) 103.