ZESZYTY NAUKOWE POLITECHNIKI ŁÓDZKIEJ
SCIENTIFIC BULLETIN OF ŁÓDŹ TECHNICAL UNIVERSITY
Nr 845 WŁÓKIENNICTWO, z. 58 2000

EXPLORATION OF UNIVARIATE DATA IN THE TEXTILE METROLOGY

Jiří Militký
Dept. of Textile Materials, Technical University of Liberec,
Hálkova 6, 461 17 Liberec, Czech Republic

Milan Meloun
Dept. of Analytical Chemistry, Uiversity of Pardubice
Pardubice, Czech Republic
e-mail: jiri.militky@vslib.cz

The main aim of data analysis in textile metrology is the extraction of relevant information from measurements. This contribution is focused to most frequent task, i.e univariate data treatment based on the analysis of experimentally determined values x_i, $i = 1,...N$. The system of exploratory data analysis based on the concept of quantile estimation is proposed. The main part of this contribution is devoted to the description of several graphical tools for data summarization and exploration. Selected techniques are used for analysis of the strength of basalt fibers at small gauge length. The program package ADSTAT for realization of the above-mentioned techniques is briefly described.

1. INTRODUCTION

Metrology is one of the very important disciplines enabling evaluation of products and processes quality. Traditionally, the core of metrology is realization of measurements. With the advent of computers and modern sophisticated measuring instruments the evaluation and interpretation of results is the main problem. In the textile branch is metrology connected with textiles production and investigation of product properties. For univariate data, the following characteristics are routinely computed [1,2]:

a) Estimation of location $\bar{x}$, and variability s^2,
b) Creation of confidence interval for expected value μ

confidence interval: $P(x_D \le \mu \le x_H) = 1 - \alpha$

The limits of this interval are

$$x_D = \bar{x} - t_{1-\alpha/2}(n-1) \,.\, s / \sqrt{n}, \qquad x_H = \bar{x} + t_{1-\alpha/2}(n-1) \,.\, s / \sqrt{n}$$

For large samples is $t_{0.975} \approx 2$. (it is equal to 95 % confidence interval) Modern approach is based on the uncertainty evaluation (see [1]).
These calculations are very simple and can be realized on the pocked calculators. Application of computer leads to the shortening of time about 80%. The main problem is that this confidence interval is correct only if data are independent realizations of random variable from normal distribution [2]. It is therefore necessary to check these assumptions and select suitable procedure of data treatment for non-normal cases. Moreover it is possible to [1]:

- Check assumption about data,
- Create probability models,
- Apply computer-oriented methods (robust, adaptive).

The main part of this contribution is devoted to extraction of probabilistic information from data x_i, i = 1,...N (**data mining**). The system of exploratory data analysis based on the concept of quantile estimation is proposed. The program package ADSTAT for realization of the above-mentioned techniques is briefly described.

2. EXPLORATORY DATA ANALYSIS

From statistical point of view leads the analysis of measurements results to the identification of probability model and estimation of corresponding parameters. Due to well-known fact that a lot of experimental data have non normal distribution, the classical analysis based on the normality assumption cannot be used. Frequently., the textiles are strongly non-homogeneous and technological process is influenced by many random events. The results of measurements are therefore often corrupted by the outliers (dirty data). Techniques that allow isolating certain basic statistical features and patterns of data are therefore necessary.

Special distribution free robust methods of this type are collected under the name exploratory data analysis (EDA) According to Tukey [3] the EDA is a "detective work". It uses as tools various descriptive and graphically oriented techniques that are free of strict statistical assumptions. These techniques are based on the assumptions of the continuity and differentiability of underlying density only.

In this contribution the set of selected computationally assisted EDA methods are discussed. The computationally assisted exploratory data analysis system is described in the book [1]. The EDA techniques are one of main parts of „statistical methods mining" which is collection of classical and modern parametric, non-parametric and function estimation methods for data treatment [4, 5].
The special variants of the *quantile plot* are proposed for graphical visualization of data and evaluation of dirty data

2.1. Some Basic Concepts

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

$$x_{(1)} < x_{(2)} < \ldots < x_{(N)}$$

which are the sample values (assumed to be distinct) arranged the in increasing order.

Let $F_e(x)$ is the distribution function from which values x_i have been sampled. It is well known that the transformed random variable

$$z_{(i)} = F_e(x_{(i)}) \tag{1}$$

has independently on distribution function F_e the Beta distribution Be [i, N-i+1]. Corresponding mean value is

$$E(z_{(i)}) = \frac{i}{N+1} \tag{2}$$

where E(.) is operator of mathematical expectations. The elements V_{ij} of covariance matrix V for all pairs $z_{(i)}$, $z_{(j)}$ i, j=1,...N are simple functions of i,j and N only. Using back transformations of $E[z_{(i)}]$ the relation

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

is obtained. In eqn. (3) the $Q_e(P_i)$ denotes quantile function and

$$P_i = \frac{i}{N+1}$$

is cumulative probability.

Description of quantile function properties and its advantages for constructing of empirical sample distribution contains paper of Parzen [5,6]. Generally, the quantile function is inverse of distribution function. When $F(x) = P$ is continuous distribution function and $f(x)$ is corresponding density is $Q(P) = x$ quantile function. It is simple to prove that $F(Q(P)) = P$ for all $0 \leq P \leq 1$ and $f(Q(P))*q(P) = 1$. The $q(P) = dQ(P)/dP$ is so called quantile density function and $f(Q(P) = 1/q(P)$ is density quantile function.

From eqn. (3) is obvious that the order statistic $x_{(i)}$ is raw estimate of the quantile function $Q_e(P_i)$ in position of P_i. For estimation of quantile $x_P=Q_e(P)$ at value $i/(n+1) < P <(i+1)/(n+1)$ the piecewise linear interpolation

$$x_{(P)} = (N+1)(\frac{PN+P-i}{N+1})(x_{(i+1)} - x_{(i)}) + x_{(i)} \tag{4}$$

can be used. Variance $D(x_P)$ can be computed from equation

$$D(x_P) = \frac{P(1-P)}{N f^2(x_P)} \tag{5}$$

Symbol $f_e(x_P)$ means the probability density function corresponding to distribution function F_e. The asymptotic distribution of x_P is normal with mean $Q(P)$ and variance defined by eqn. (5).

The interpolation (4) can be used for estimation of sample quantiles x_{Pi} or x_{1-Pi} for $P_i=2^{-i}$ $i=1,...n$. These quantiles are called letter values [7]. All letter values except for i=1 (median) are in pairs. For example we can estimate 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

$$P_i = \frac{i-0.375}{N+0.25} \tag{6}$$

proposed by Blom is often used. Some proposals for definitions of P_i are presented in paper [8]. For characterization of <u>location</u> the median $x_{0.5}$ can be evaluated as the middle of order statistics. <u>Spread</u> is expressed as difference between upper quartile $x_{0.75}$ ($P_i= 0.75$) and lower quartile $x_{0.25}$ ($P_i= 0.25$). Parzen [5] proposed so called quantile deviation

$$DQ = 2*(x_{0.75} - x_{0.25}) \tag{7}$$

DQ is justified as numerical derivative of Q(P) at P=0.5 which roughly approximate so called density quantile deviation

$$DfQ = 1/f(Q(0.5)) = 1/f(x_{0.5}) = q(0.5) \tag{8}$$

By using of these estimators of spread and location is possible to standardize data probability distribution to have quantile function to make DQ = 1 and median $x_{0.5}$ = 0. Standardized quantile function is called shape identification quantile function QI(P). The sample form of QI(P) is defined as

$$QI_e(P) = \frac{(x_{(i)} - x_{0.5}}{2*(x_{0.75} - x_{0.25})} \tag{9}$$

Values $|QI_e(P)| \geq 1$ are identified as outliers (for normal distribution) or as indicators of long tailed distribution. Values QI$_e$(P) can be easily used for specification of skewness and tail length. As a measures of skewness the SQ = QI$_e$(0.25) + QI$_e$(0.75) can be used (for symmetric distribution is this quantity equal to zero). Measures of tail length are QI$_e$(0.05) and QI$_e$(0.95):

short tail is for QI$_e$(0.95) < 0.5,
long tail is for QI$_e$(0.95) > 1
medium tail is for 0.5 < QI$_e$(0.95)< 1.

These diagnostics are very simple and can be used for crude evaluation of data distribution type.

2.2. Data Visualization

For graphical visualization of data many simple techniques as stem-leaf plot, box plot, dot plot [1] and dig-dot plot [9] have been proposed. Only the simple quantile plot (QP) and its variant are described here.

Symmetry and tail length can be characterized by using of g - h distribution system (see [4]).

Empirical (sample) quantile plot Q(P) is constructed as dependence of $x_{(i)}$ on P_i. From patterns of points some statistical features of data as a symmetry, local concentration and rough normality can be simply recovered. Detailed interpretation of QP is described in the book [1].

For better interpretation the quantile functions of normal distribution

$$Q_N(P) = \mu + \sigma u_P \tag{10}$$

are superimposed to QP. In eqn. (10) the u_p are quantiles of standard normal distribution N(0, 1). Parameters μ and s are estimators of location and scale.

Two various normal quantile functions are graphed. The first one is based on the moment estimators i.e. sample mean x_M and sample standard deviation s. The second one uses robust quantile estimators median $x_{0.5}$ and quartile based standard deviation

$$s_M = \frac{x_{0.75} - x_{0.25}}{1.349} \qquad (11)$$

This variant of QP enables to compare deviation of sample values from assumed normal distribution. For data from skewed distribution the normal quantile function can be replaced by assumed distribution. For frequent exponential distribution is quantile function in the form

$$EX(P) = A + B\ln(1 - P) \qquad (12)$$

The parameters of threshold A and scale B can be estimated as A ~ $x_{(1)}$ and B ~ x_M - $x_{(1)}$. Quantile functions for other types of positively skewed distributions are summarized in book [1]. QP graphs with superimposed theoretical quantile function enables to identify the outliers (dirty data) as well. For complex visualization of data the quantile box plot (QBP) proposed by Parzen [5] is useful.

3. PROGRAM SYSTEM ADSTAT

System ADSTAT contains 8 independent modules of statistical methods for univariate and multivariate data [4, 12]. The manipulation with ADSTAT is very simple by using of pull down menu and panes. Individual program modules are built windows like environment. This environment includes the powerful block oriented data editor, context sensitive help and unified graphical presentation. Exploratory methods included in module "Basic Statistics" can be divided to three main parts.

A. Techniques for presentation of data.

B. Construction of empirical sample distribution and comparison with 12 theoretical ones.

C. Power transformation of data

The above-mentioned and more complex EDA techniques described in [4] are used. This program realized the computations in this contribution.

4. EXPERIMENTAL PART

The above-mentioned methods were used for identification of strength type distribution of basalt fibers. Strength of fibers was measured on the tensile testing machine at gauge length 1 cm. These values were converted to the stress at break values by dividing of strength by fiber cross-section area. The 49 values S_i of stress at break [GPa] were used for further analysis.

5. RESULTS AND DISCUSSION

The exploratory data analysis of sample values S_i, i = 1,...49) consists from investigation of data peculiarities by the combination of graphs and numerical indicators.

The basic characteristics of data are summarized in the table I.

Table 1

Basic data characteristics

Median:	**2.3400**	**Mean:**	**2.7543**
Variance:	**2.3860**	**Standard deviation:**	**1.5447**
Skewness:	**2.3111**	**Kurtosis:**	**9.2342**

Selected quantiles are in the table 2.

Table 2

Selected quantile from data

Quantile	**Probablity P_i**	**Lower value**	**Upper value**
Sedecile	**0.0625**	**1.30**	**5.27**
Octile	**0.1250**	**1.59**	**4.05**
Quartile	**0.2500**	**1.91**	**3.01**

Estimated quantile for $P_i = 0.95$ is $x_{0.95} = 5.918$. The DQ = 2.2, $QI_e(0.25) = -0.195$, $QI_e(0.75) = 0.304$, SQ = 0.109 and $QI_e(0.95) = 1.626$. These values indicate that data distribution is skewed with possible outliers Simple exploratory graphs i.e. dot plots and box plots (see [4]) presented on the fig. 1 support these findings.

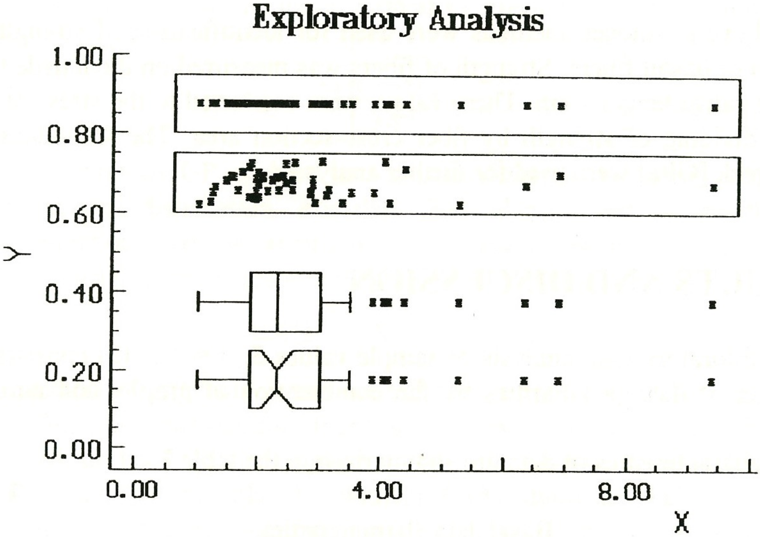

Fig. 1. Simple exploratory graphs

The quantile plot on the fig, 2 shows systematic deviation form normality and one possible outlier

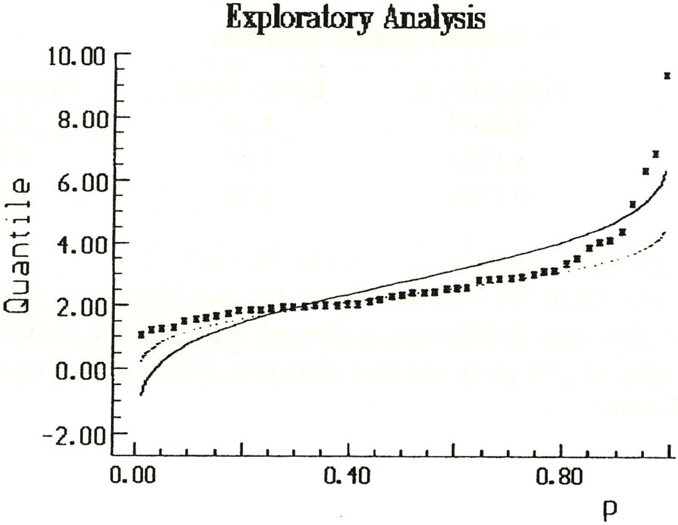

Fig 2. Quantile graph (solid line-moment estimators, dotted line-quantile estimators)

The main result of exploratory analysis is that the data are skewed to the right and normality assumption cannot be used.

7. CONCLUSION

The methodologies for statistical analysis based on the exploratory data analysis principles are shown. Program system ADSTAT is well suited for EDA of one-sample problems on personal computers. Extensive description of algorithms used in ADSTAT and examples of its utilization for analysis of chemical data is given in the book [1].

Acknowledgment: This work was supported by the Czech Grant Agency; grant GACR No. 106/99/1184 and Czech Ministry of Education Grant VS 97084.

REFERENCES

[1] **Meloun M., Militký J., .,Forina M.,** Chemometrics for Analytical Chemistry, vol1 PC Aided Statistical Data Analysis, Ellis Horwood, Chichester 1992.

[2] **Burry K.V.,** Statistical Models in Applied Science, J. Wiley, New York 1975.

[3] **Tukey J.W.,** Exploratory Data Analysis, Addison Wesley, Reading, Mass. 1977.

[4] **Militký J.,** System EXDAB, Proc. Int. Conf. COMPSTAT' 84, Prague, 1984, Poster section.

[5] **Parzen E.,** Statistical methods mining and non parametric quantile domain data analysis, Proc Ninth int. conf. on quantitative methods for environmental science, July 1988, Melbourne.

[6] **Parzen E.,** J. Amer. Statist. Assoc. 74, 105 (1985).

[7] **Hoaglin D. C., Mosteler F. Tukey J. W.,** Eds. Understanding Robust and Exploratory Data Analysis, J. Wiley, New York, 1983.