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

DATA BASED PROBABILITY MODELS 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 and creation of probability models. 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 main part of this contribution is devoted to the description of several graphical tools for construction of data based distribution and comparison of data distribution with theoretical ones. Selected techniques are used for analysis of the strength of basalt fibers at small gauge length.

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.

The main problem of data analysis is selection of data distribution. Classical methods are valid under strict assumption. For example the results uncertainty expressed by the confidence interval is correct only if data are independent realizations of random variable from normal distribution [2]. Assumption of normality cannot be generally accepted and in some situations the non normal distribution arises from pure theoretical ideas. It is therefore necessary to check this assumption and select suitable procedure for creation of data based probability models.

The main part of this contribution is devoted to creation of probabilistic models from data x_i, i = 1,...N. The system of exploratory data analysis based on the concept of quantile estimation is proposed

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 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 for creation of probability models 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 [5].

The construction of sample distribution i.e. the estimation of probability density function will be carried out by the *kernel estimation* of probability density function and by the *quantile-quantile* plot [4].

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]. 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

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

can be used. The interpolation (4) is useful 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. Some proposals for definitions of P_i are presented in paper [8].

2.2. Building of Sample Distribution

As an estimator of the empirical probability density function histogram with variable bins is often constructed. Smooth kernel type density estimator is natural

generalization of histogram. Histogram is piecewise constant estimator of sample probability density. Histogram height in j-th class bounded by values (t_{j-1}, t_j) is calculated from the relationship

$$f_H(x) = \frac{C_N(t_{j-1}, t_j)}{N h_j} \tag{5}$$

where the function $C_N(a, b)$ denotes the number of sample elements within interval <a, b> and

$h_j = t_j - t_{j-1}$ is the length of the j-th interval. Now, the problem encountered is the choice of boundary values $\{t_j\}$ j=1,...M, the number of class intervals M and their lengths h_j with respect to the histogram quality. In our ADSTAT programs the simple data based two-stage technique is used. In the first stage the number of class intervals

$$M = \text{int}[2.46\,(N-1)^{0.4}] \tag{6}$$

is computed Here int[x] is integer part of number x. 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 the order statistics $x_{(i)}$ is used. In practice the P-axis is divided into identical intervals having the size of 1/M. For these intervals the corresponding quantile estimates $tj = x_{(j/M)}$ are constructed by using of eqn. (4) where P = j/M. Practical experiences have hitherto proven that this construction be suitable even for strongly skewed sample distributions.

The kernel type nonparametric estimator of sample probability density f(x) can be constructed on the basis of Lejenne-Dodge-Kaelin procedure [11]. The final estimator has the form

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K\left[\frac{x - x_i}{h_i}\right] \tag{7}$$

Selection of kernel function K[x] and computation of bandwidths h_i is described in [12].

2.3. Selection of Sample Distribution

The main goal is to approximate the empirical sample distribution by suitable theoretical one. The comparison of these distributions can be made by the variants of Q - Q or P - P plot.

Classical Q - Q plot is based on comparison of empirical quantile function $Q(P_i) \sim x_{(i)}$ with chosen theoretical quantile function $Q_T(Pi)$. For theoretical distribution functions of type $F_T((x-T)/S)$ is attractive to use standardized quantile function $Q_{TS}(P_i)$ (see[4]) or shape identification quantile function $QI_T(P_i)$. When empirical and theoretical distributions are in coincidence, the relationship

$$x_{(i)} = T + S\,Q_{TS}(P_i) \tag{8}$$

is valid. Here usually T is the location parameter and S is the parameter of scale. For some three-parameter distribution the shape factor is usually a parameter of the plot. Our programs (ADSTAT) select such shape factor value that straightens the individual points best. Due to 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 often hard to quantify.

In the work of Michael [13] the stabilized probability plot is introduced. Kafander and Spiegelman [14] propose the conditional Q - Q plot. For EDA purposes we use the empirical probability plot (EPP) (see also [4]).

The P-P (or Probability-Probability) plot is useful for determining how well a specific theoretical distribution fits the observed data. In the P-P plot, the observed cumulative distribution function (estimated by probability P_i) is plotted against a theoretical cumulative distribution function $F_T(x_{(i)})$ in order to assess the fit of the theoretical distribution to the observed data. If all points in this plot fall onto a diagonal line (with intercept 0 and slope 1), then can be concluded that the theoretical cumulative distribution approximates the observed distribution well. If the data points do not all fall on the diagonal line, then can be used this plot to visually assess where the data do and do not follow the distribution (e.g., if the points form an S shape along the diagonal line, then the data may need to be transformed in order to bring them to the desired distribution pattern).

In order to create this plot, the theoretical distribution function must be completely specified. Therefore, the parameters for the distribution must either be defined by the user or computed from the data (see the specified distribution for more information on the respective parameters). Parzen [5] proposed for this purpose so called comparison distribution function

$$CDD = F_T(Q_e(P_i)) \tag{9}$$

The CDD is roughly equal to the theoretical distribution function in point $x_{(i)}$. plot of CDD against P_i is therefore for sample data equal to the P -P plot. For comparative purposes is better to use comparative P - P plot where CDD is replaced by the difference CDD-P_i. Combining of Q - Q and P - P plots leads to the best diagnostics.

3. 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.

4. RESULTS AND DISCUSSION

The main aim of data analysis of sample values S_i, i = 1,...49) is identification of suitable probability model for further detailed statistical analysis

4.1. Non-parametric density estimation

It is well known, that for full description of random variables the corresponding probability density function is required. From values S_i, i = 1,...N sample density estimator was constructed. For creation of density trace the kernel type non parametric estimator has been used.. Typical kernel type estimator (dotted line) is compared with density of normal distribution (solid line) on the fig 1.

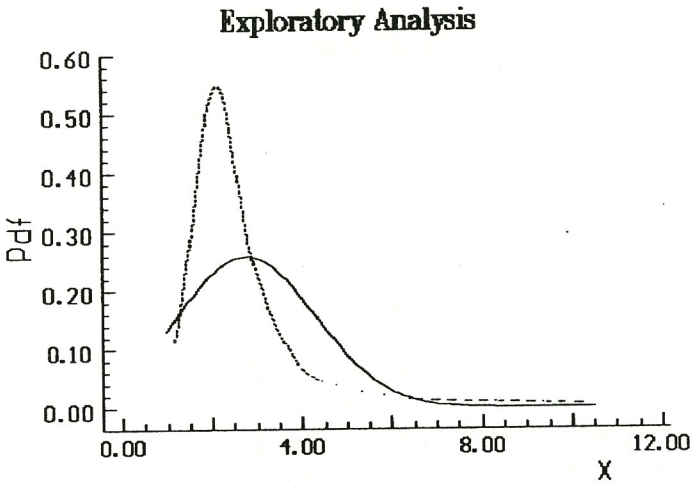

Fig. 1. Non-parametric density trace (dotted) and normal one (solid)

From these kernel type non-parametric estimator is evident, that the sample distribution is skewed to the right. On the fig. 2 is shown histogram and best lognormal distribution and on the fig. 3 is the same for best Weibull type distribution. If is evident that the better results gives lognormal distribution

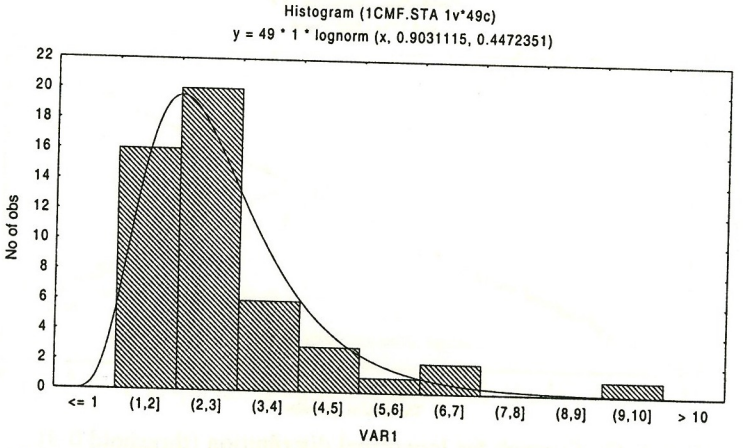

Fig. 2. Histogram and best density trace of lognormal distribution

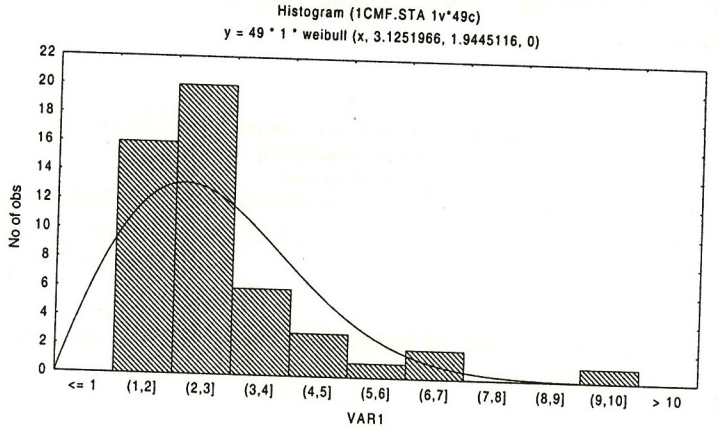

Fig. 3. Histogram and best density trace of Weibull distribution

4.2. Selection of theoretical distribution

The graphical tools for selection of theoretical distribution well approximating the sample one are the Q-Q graphs (see chap. 2). Empirical quantiles $Q_e(P_i)$ are approximated by the sample order statistics $x_{(i)}$. For experimental data the Q-Q graphs for normal, log-normal,, rectangular

exponential, Weibull, gamma, Pareto and Gumbell distribution were created. The Q-Q graph for lognormal distribution presented on the fig 4 was the best one.

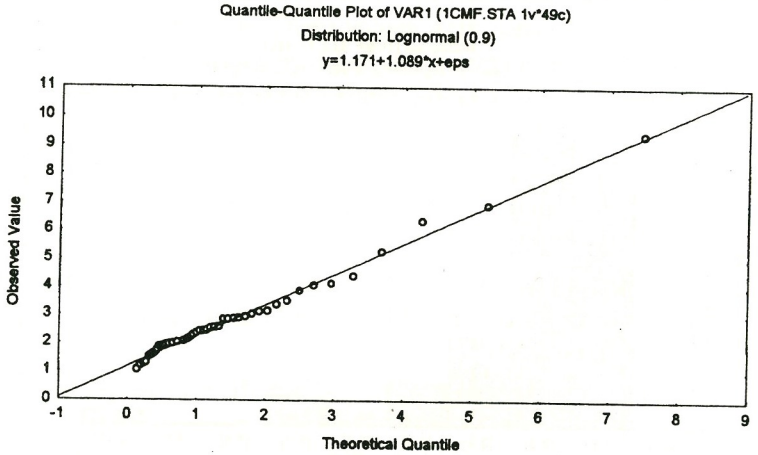

Fig. 4. Q - Q graph for lognormal distribution (threshold 0.9)

For comparison is on the fig. 5 the Q - Q graph for Weibull distribution with o threshold equal to 1.

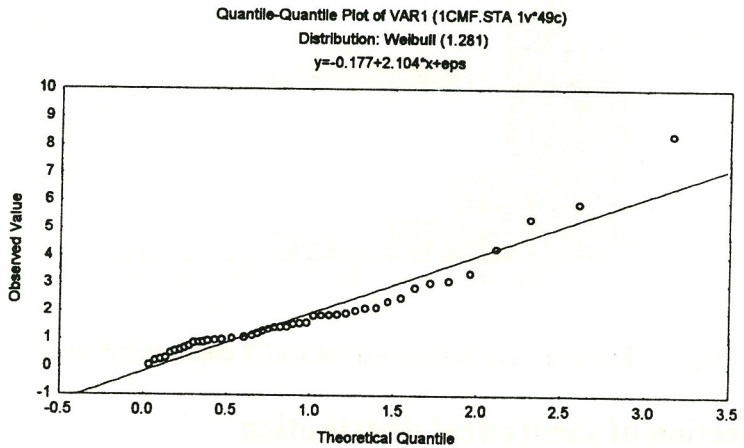

Fig. 5. Q - Q graph for Weibull distribution (optimal threshold = 1)

The P - P plot for optimal lognormal distribution is shown on the fig. 6

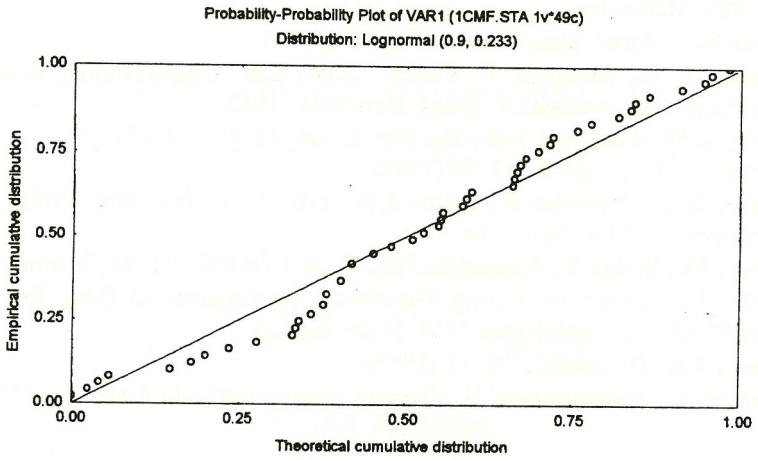

Fig. 6. P - P graph for optimal lognormal distribution (threshold 1)

The P - P plot is here more decisive from point of view of selection of suitable theoretical model approximating of data distribution. Under the validity of lognormal distribution no outliers are detected.

7. CONCLUSION

The methodologies for data based probability models creation are shown. By using of this methodology the statistical model of stress at break is identified as lognormal distribution

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.

[8] **Looney S. W., Gulledge T.R.,** The Statistician, 34, 297, (1985).

[9] **Hunter S.,** Amer. Statist., 42, 54 (1988).

[10] **Hoaglin D. C., Mosteler F., Tukey J.W.,** Eds.: Exploring Data, Tables, Trends and Shapes, J. Wiley, New York 1985.

[11] **Lejenne M., Dodge Y., Kaelin E.,** Proc. Conf. COMPSTAT' 82, Toulouse, 1982.

[12] **Militký J.,** Package for Fitting Theoretical Distributions to Data, Proc. Conf. COMPSTAT '88, Copenhagen 1988, Poster section.

[13] **Michael J.R.,** Biometrika, 70, 11 (1983).

[14] **Kafander K., Spiegelman C.H.,** Comput. Statist. and Data Anal., 4, 167 (1986).