Computer-Assisted Data Treatment in Analytical Chemometrics II. Analysis of Sample Assumptions

[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 23 April 1993

The analysis of basic assumptions about the data set examines an independence of sample elements, normality of sample distribution, minimum sample size, and sample homogeneity. These procedures are illustrated on the trace analysis of a DDT content in 144 fish specimens and on the quantitative determination of calcium oxide in glass reference material.

Statistical tests and interval estimates are designed to yield reliable results with data which meet certain requirements [1—11]. These requirements are certain assumptions about the nature of the data, observations. If our data do not meet the assumptions, the results may give incorrect answers. Assumptions are usually made, the observations are normally distributed and the errors associated with the observations are independent and random in nature. If the error associated with the one observation does not affect the other observations, the errors are said to be independent. In the usual course of events some of these errors will be relatively large, some relatively small, some positive, and some negative. If these errors have no particular pattern with respect to the size and sign, they are considered to be random.

This paper brings a description of some statistical procedures applied in original statistical package ADSTAT for examination of all above assumptions about data.

THEORETICAL

Statistical treatment of experimental data supposes that the data are independent random variables coming from the same distribution, obviously normal one, and that the sample size is sufficient for precise estimates of location and spread to be obtained. When some of these assumptions are not fulfilled, the data analysis is rather complicated. These assumptions must be examined before interval estimation and testing.

Examination for Independence of Sample Elements

The basic assumption of good measurement is that the individual measurements, observations in the sample set are independent. Interdependence of measurements is obviously caused by

1. instability of the measurement device, for example, a shift with temperature;

2. variable conditions of measurements, which could be suddenly changed;

3. neglection of important factor(s) which have a great influence on measurement, for example, the sample volume, temperature, purity of chemicals, *etc.*;

4. false and nonrandom (stratified) sampling.

When some experimental conditions change over time, a time dependence in the observations may be indicated. When there is a sudden change in observations, a heterogeneous sample is formed. In both the above cases, a higher value for the variance is found than for a homogeneous sample.

Time dependence or dependence on the order of observations can be tested for by examining the significance of the first-order autocorrelation coefficient ρ_a according to

$$t_n = \frac{T_1\sqrt{(n+1)}}{\sqrt{(1-T_1)}} \qquad (1)$$

where

$$T_1 = \left(1-\frac{T}{2}\right)\sqrt{\frac{n^2-1}{n^2-4}} \qquad (2)$$

and T is the von Neumann ratio defined by

$$T = \frac{\sum_{i=1}^{n-1}(x_{(i+1)}-x_{(i)})^2}{\sum_{i=1}^{n}(x_{(i)}-\bar{x})^2} \qquad (3)$$

When the null hypothesis H_0: $\rho_a = 0$ is valid, the

test criterion t_n has the Student distribution with $(n + 1)$ degrees of freedom. The alternative hypothesis is H_A: $\rho_a \neq 0$. When $|t_n| > t_{1-\alpha/2}(n_1 + 1)$, the null hypothesis about the independence of sample observations is rejected at the significance level α.

Examination for Normality of Sample Distribution

Normality of a sample distribution is the basic assumption of most statistical data treatment, because many statistical tests require normality. When the type of deviation from normality of the sample is known before statistical inference, the *directional tests* are used; when the type of deviation from normality is unknown, the *omnibus tests* are used.

Generally, the statistical tests are less sensitive to deviations from normality than diagnostic graphs. Moreover, the deviation from normality can be also caused by the presence of outliers. When the normality of sample distribution is not proved, the data should be analyzed with great care. For testing normality of a sample distribution the rankit plot is one of the most useful tools, but other useful tests are available.

1. Test of combined sample skewness and curtosis

The testing criterion used in ADSTAT is defined as

$$C_1 = \frac{\hat{g}_1^2(x)}{D(\hat{g}_1(x))} + \frac{(\hat{g}_2(x) - 3)^2}{D(\hat{g}_2(x))} \qquad (4)$$

where $\hat{g}_1(x)$ is the sample skewness and $D(\hat{g}_1(x))$ is its variance, $\hat{g}_2(x)$ is the sample curtosis and $D(\hat{g}_2(x))$ is its variance. For a normal distribution, the test criterion C_1 has approximately the χ^2 distribution, so that when $C_1 > \chi^2_{1-\alpha}(2)$, the null hypothesis about normality of sample distribution is rejected.

2. Anderson—Darling test

This test is based on the empirical distribution function $F_E(x)$. The null hypothesis, H_0: $F_E(x) = F_T(x)$ is tested *vs.* H_A: $F_E(x) \neq F_T(x)$ where $F_T(x)$ is the distribution function of the fully specified distribution. The test criterion is defined as

$$AD = n - \frac{\left[\sum_{i=1}^{n} (2i-1)(\ln Z_i + \ln(1 - Z_{n-i+1}))\right]}{n} \qquad (5)$$

where Z_i is the standardized variable $Z_i = F_T(x_{(i)})$. When testing for normality of a sample distribution, the null hypothesis is formulated H_0: $F_E = N(\bar{x}; s^2)$ and the variable $Z_i = \Phi[(x_{(i)} - \bar{x})/s]$ represents the values of the normal distribution function. When $AD > D_{1-\alpha}$, the null hypothesis about normality is rejected. The quantile $D_{1-\alpha}$ may, for large samples, be approximated by

$$D_{0.95} = 1.0348\left(1 - \frac{1.013}{n} - \frac{0.93}{n^2}\right) \qquad (6)$$

Examination for Minimum Sample Size

The sample size has an influence on the precision of estimates and controls the size of confidence intervals. For very small sample sizes it may happen that hypothesis tests are affected more by the sample size n than by the variability of data. The procedure for finding sufficient sample size is in ADSTAT as follows:

1. From n_1 starting values the sample variance $s_0^2(x)$ is calculated. The minimum size n_{min} of a sample taken from a normal distribution is calculated in such a way that for an optioned probability $(1 - \alpha)$ and value of d, the confidence interval will be $\mu - d \leq \bar{x} \leq \mu + d$. Then n_{min} is given by

$$n_{min} = s_0^2(x)\left[\frac{t_{1-\alpha/2}(n_1 - 1)}{d}\right]^2 \qquad (7)$$

where $t_{1-\alpha/2}(n_1 - 1)$ is the quantile of the Student distribution with $(n_1 - 1)$ degrees of freedom.

2. The minimum sample size n_{min} may be chosen so that the relative error of the standard deviation $\delta(s)$ has a selected value. Then n_{min} is given by

$$n_{min} = 1 + \frac{\hat{g}_2(x) - 1}{4\,\delta^2(s)} \qquad (8)$$

where $\hat{g}_2(x)$ is the estimate of the curtosis of the sample distribution. The value of $\delta(s)$ in % usually chosen is 10, *i.e.* $\delta(s) = 0.1$. The minimum size n_{min} is several tens, so typical sample sizes used in chemical laboratories $n = 5, 10, \ldots$ are too small from the statistical point of view.

Examination of Sample Homogeneity

Sample heterogeneity becomes evident when a sample contains outliers or when the sample can be logically divided into several subsamples and each of them can be analyzed separately. Testing the difference of subsample averages can indicate whether the separation into subsamples can be taken as significant or not. We limit ourselves here to the situation when outliers exist in a data batch. Outliers significantly differ from all other values and can be read-

ily identified by EDA plots [12]. Outliers cause distortion of the estimates $\bar{x}$ and s^2 and may impair the subsequent statistical testing.

There are many different techniques, *e.g. cf.* Ref. [10] for the identifying outliers, when a normal distribution of data can be assumed. One of the simplest and most efficient methods seems to be *Hoaglin's modification of inner bounds* B_L^* and B_U^*

$$B_L^* = \tilde{x}_{0.25} - K(\tilde{x}_{0.75} - \tilde{x}_{0.25}) \qquad (9)$$

and

$$B_U^* = \tilde{x}_{0.75} + K(\tilde{x}_{0.75} - \tilde{x}_{0.25}) \qquad (10)$$

where $\tilde{x}_{0.25}$ is the lower quartile, $\tilde{x}_{0.75}$ is the upper quartile and the parameter K is selected so that the probability $P(n, K)$ that no observation from a sample of size n will lie outside the modified inner bounds $[B_L^*, B_U^*]$, is sufficiently high, for example, $P(n, K) = 0.95$. For $P(n, K) = 0.95$ and $8 \leq n \leq 100$, *Hoaglin* [9] derived the following equation for calculation of parameter K

$$K = 2.25 - \frac{3.6}{n} \qquad (11)$$

All elements lying outside the modified inner bounds $[B_L^*, B_U^*]$ are considered to be potential outliers.

COMPUTATION

Procedure of EDA of Univariate Data

The extent of exploratory data analysis (EDA) of univariate data is best chosen according to experience from the previous data analysis. We consider here two common situations: *a*) the treatment of routine data and *b*) the treatment of new data when no preliminary information is available.

a) The Analysis of Routine Data

With routine data, some knowledge of the sample distribution is assumed — it is usually normal, and the data elements are homogeneous and independent. Tests for examining all assumptions about data should include: *i*) a test for minimal sample size; *ii*) a test for independence of sample elements; *iii*) a test for normality; *iv*) a test for homogeneity of sample. Graphical EDA techniques such as the rankit plot and quantile-box plot are often used.

When no preliminary information about the data is available, the full range of EDA plots should be followed by determination and construction of the sample distribution. When no suitable distribution has been found, a power transformation of data is recommended [13]. To summarize a batch of experimental data, the quantile-box plot is always used.

b) The Analysis of New Data

Analyzing new data, there are several cases that require different strategies for the EDA and CDA procedures [12—14].

Case I. *No independence of sample elements*

When the sample elements are not proved to be independent a danger of systematically biased and overevaluated estimates for a positive value of ρ_a arises. Therefore, a new logical analysis of the experimental equipment and data measurement procedures is necessary: after an improvement in the experimental strategy, the new data should be examined again.

Case II. *No normality of sample distribution*

The actual distribution of sample is not normal in nature, or outliers are present in data. When the distribution is not normal, the deviation can be in the length of tails or in skewing. When tails differ in length, robust estimates may be used, or a power transformation [13] should be chosen. For skewed distributions, a power transformation should be always used. When a power transformation is successful and the optimal value λ is found, the estimates of the parameters of location and spread can be calculated and reexpressed in the measure of the original variables. If the power transformation is not successful, exploratory data analysis [12] can be used to find a suitable approximate theoretical distribution.

When the actual distribution is strongly skewed, with the skewness $\hat{g}_1$, the random variable t_c can be defined

$$t_c = \left[(\bar{x} - \mu) + \frac{\hat{g}_1}{6\sigma^2 n} + \frac{\hat{g}_1}{3\sigma^4}(\bar{x} - \mu)^2\right]\frac{\sqrt{n}}{s} \qquad (12)$$

Variable t_c has the Student distribution with $(n - 1)$ degrees of freedom and can be used for confidence interval construction. In practical calculations the variance σ^2 is replaced by its unbiased estimate s^2 and the skewness $\hat{g}_1$ by its unbiased estimate

$$\hat{g}_1 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(x_i - \bar{x})^3 \qquad (13)$$

For a construction of the confidence intervals $H_L \leq \mu \leq H_U$, the quadratic equation to μ defined by eqn

(*12*) should be solved. The limits H_L and H_U will then be

$$H_L = \bar{x} + \frac{1 - \sqrt{d_1}}{2\,C_2} \qquad (14)$$

$$H_U = \bar{x} + \frac{1 - \sqrt{d_2}}{2\,C_2} \qquad (15)$$

where

$$C_1 = \frac{\hat{g}_1}{6\,s^2 n}$$

$$C_2 = \frac{\hat{g}_1}{3\,s^4}$$

$$d_1 = 1 - 4C_2(C_1 - C)$$

$$d_2 = 1 - 4C_2(C_1 + C)$$

$$C = t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

The confidence interval of the mean $H_L \leq \mu \leq H_U$ can be also used for statistical inference about this parameter of location.

Case III. *Sample not homogeneous*

It should first be considered whether the distribution is skewed or not, because some points would appear to be outliers for a symmetrical (normal) distribution, but would be accepted in a skewed distribution.

When some points may be extremes or outliers there are two alternatives: *a*) Exclude the outliers from the data batch. For a small sample size, this may lead to loss of valuable information; *b*) Apply robust methods. In both cases the experimenter should be consulted about the suspect points from the physical point of view, in order to consider the possibility of gross errors.

Case IV. *The sample size is not sufficient*

The best solution is to carry out new experimental measurements. As a general rule, when the variance of the data is small, a relatively smaller size will be required for any given precision of estimate. When no extra experiments can be carried out, the technique for small sample sizes should be applied. This is convenient for routine data analysis, but for new data exploratory data analysis should be used first, so that any statistical peculiarities of the sample are determined.

SOFTWARE

Procedure BASIC ASSUMPTIONS in package ADSTAT [15] computes minimal sample size for normally distributed data. It enables a test of sample independence based on autocorrelation coefficient, test of normality, and test of homogeneity based on normality assumption, too.

RESULTS

Study Case 1. *Trace analysis sample contains many outliers*

Chemical plants often discharge toxic waste material into nearly rivers and streams. One type of pollutant, commonly known as DDT, is especially harmful to fish and, indirectly, to people. There is the limit for DDT content (w) in individual fish at 5 parts per million (5 ppm). Fish with DDT content exceeding this limit are considered potentially hazardous to people if consumed. A study was undertaken to examine the DDT content of fish inhabiting the river near the chemical plant [16]. A sample of 144 fish specimens was analyzed on the DDT content to estimate the measures of location and to test the allowed content w = 5 ppm.

Data (w/ppm): 10.00, 19.00, 1.30, 12.00, 16.00, 7.20, 4.80, 33.00, 23.00, 6.00, 5.10, 48.00, 21.00, 10.00, 5.10, 10.00, 50.00, 12.00, 4.00, 44.00, 150.00, 2.80, 10.00, 0.43, 28.00, 0.48, 12.00, 1100.00, 7.70, 0.18, 22.00, 9.40, 2.00, 0.34, 10.00, 4.10, 19.00, 0.11, 11.00, 2.80, 16.00, 0.22, 17.00, 0.74, 5.40, 0.80, 9.70, 14.00, 2.60, 8.70, 12.00, 22.00, 3.10, 22.00, 4.70, 9.10, 3.50, 13.00, 6.00, 140.00, 9.10, 3.50, 3.80, 4.20, 7.80, 9.30, 17.00, 12.00, 4.10, 21.00, 12.00, 2.00, 8.40, 3.40, 1.40, 0.30, 15.00, 13.00, 6.10, 1.20, 25.00, 5.60, 2.80, 7.10, 5.60, 12.00, 4.80, 180.00, 4.60, 21.00, 5.70, 1.50, 8.20, 8.00, 3.30, 2.40, 6.10, 12.00, 3.30, 4.30, 13.00, 6.00, 3.70, 3.90, 6.00, 4.70, 9.90, 99.00, 6.60, 31.00, 6.80, 0.45, 5.50, 5.20, 13.00, 2.50, 11.00, 27.00, 8.80, 0.25, 4.50, 18.00, 57.00, 0.58, 4.20, 7.50, 96.00, 2.00, 3.00, 3.00, 360.00, 2.20, 2.30, 13.00, 130.00, 7.40, 2.50, 7.30, 13.00, 0.35, 6.80, 15.00, 61.00, 1.90.

Solution: Notice that the data set contains many outliers because in some river tributaries the DDT content is even 1000 times greater. For this reason some of the EDA graphs [12] are confused (Fig. 1). Statistical tests examining basic assumptions of sample are here suitable.

a) Test for independence of sample elements t_n = 0.3472 < $t_{0.975}$ (144 + 1) = 1.976 leads to conclusion that an independence of sample elements is accepted.

b) Test for normality of sample distribution C_1 = 65806 > $\chi^2_{1-\alpha}(2)$ = 5.992 leads to conclusion that the normality is rejected. It is perhaps for the presence of outliers in a sample.

c) Test for sample homogeneity leads to conclusion that a sample contains 13 outliers which should be excluded from the sample. Outliers are: 50, 150, 57, 96, 360, 130, 61, 48, 44, 1100, 140, 180, 99.

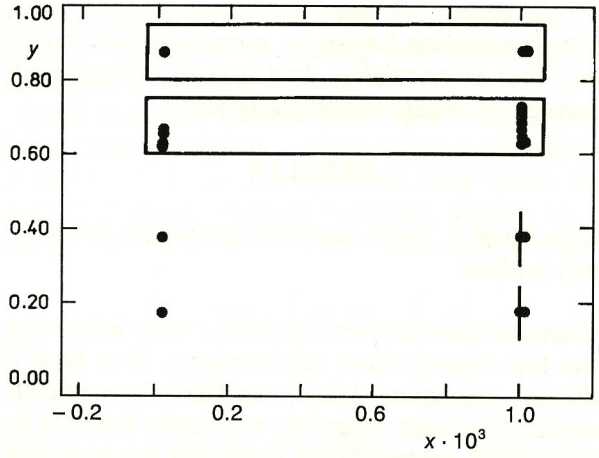

Fig. 1. Box-and-whisker plots of DDT content in fish.

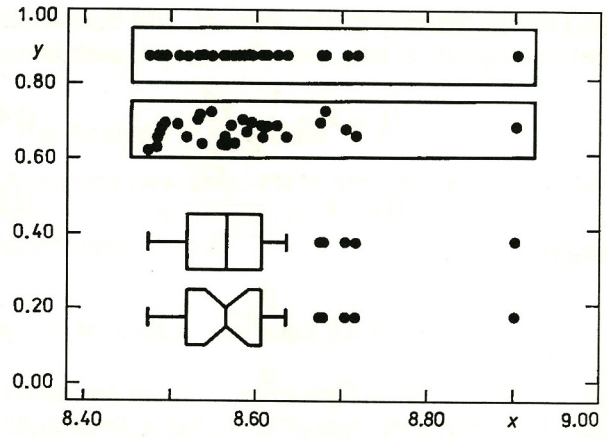

Fig. 2. Box-and-whisker plots of CaO content in glass.

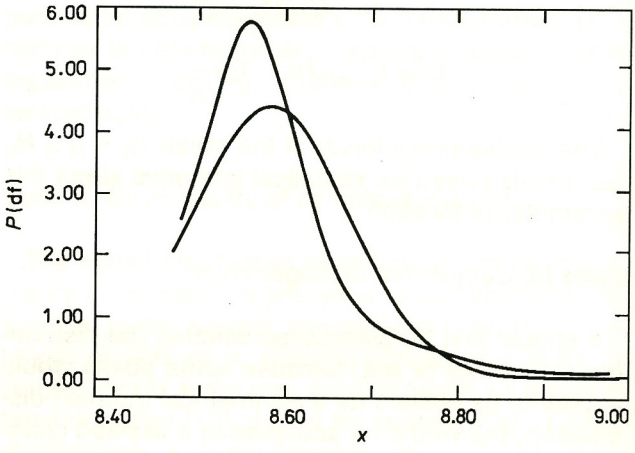

Fig. 3. Kernel estimate of the probability density function of original sample concerning CaO content in glass. Upper curve: the empirical curve of sample distribution. Lower curve: the curve of Gaussian distribution.

d) Interval estimate of location of DDT content for original sample represented by the mean is (25.04 ± 16.23) ppm. As normality of sample distribution is rejected and sample contains 13 outlying values, this value of sample mean is false and cannot be used. Excluding outliers the mean reaches more realistic value (8.32 ± 1.70) ppm.

When some basic assumptions are not fulfilled, robust estimates of location, *i.e.* the median and the trimmed mean should be used: median = (7.25 ± 1.79) ppm, 5 % trimmed mean = (11.00 ± 4.09) ppm, 10 % trimmed mean = (9.01 ± 1.89) ppm, and 40 % trimmed mean = (7.30 ± 1.63) ppm. Robust estimates lead to similar values as the mean of sample with excluding outliers and can be taken as acceptable estimates of location. Presence of outliers leads here to the unacceptable high estimate of location.

The DDT content of 7.25 ppm is higher than allowed 5 ppm so that the fish are considered toxic.

Study Case 2. *Comparison of graphical and test diagnostics*

To prepare the reference glass material the analytical content of calcium oxide CaO was evaluated by the AFS method and the following data set has been achieved [17].

Data (w(CaO)/%): 4.084, 4.043, 4.004, 4.048, 4.013, 3.993, 4.067, 4.019, 4.073, 4.041, 4.056, 3.985, 4.038, 4.007, 4.046, 3.996, 4.004, 4.050, 4.020, 4.082, 3.992, 4.039, 4.047, 4.024, 4.001, 4.004, 4.056, 4.048, 4.030, 4.015.

Solution: Graphical diagnostics of the exploratory data analysis are compared with statistical tests of basic assumptions about a sample:

a) Test for independence of CaO sample elements $t_n = 1.110 < t_{0.975}\,(30 + 1) = 2.039$ leads to conclusion that an independence is accepted. There is no time dependence of measured values of CaO content.

b) The test of combined sample skewness and curtosis for normality of sample distribution $C_1 =$ $40.434 > \chi^2_{1-\alpha}(2) = 5.992$ leads to conclusion that the normality is rejected while the Anderson—Darling test, $AD = 0.390 < D_{0.95} = 0.731$, proves a sample normality. The box-and-whisker plot (Fig. 2) and the kernel estimate of probability density function (Fig. 3) show that the sample contains one significant outlier of high value and distribution is rather skewed to lower values. The quantile-quantile plot (Fig. 4) shows that except one outlier the points fit with a straight line quite well.

c) Test for minimum sample size leads to conclusion that to reach 10 % relative error of standard deviation there should be measured 135 observations.

d) Test for sample homogeneity leads to conclusion that 1 significant outlier (8.902) should be excluded. This conclusion is indicated also by graphical diagnostics of exploratory data analysis. Excluding this outlier from the sample, the quantile-quantile plot proves normality (Fig. 5) and the curve of Gaussian distribution is closer to the empirical curve (Fig. 6).

e) Interval estimate of location for original sample

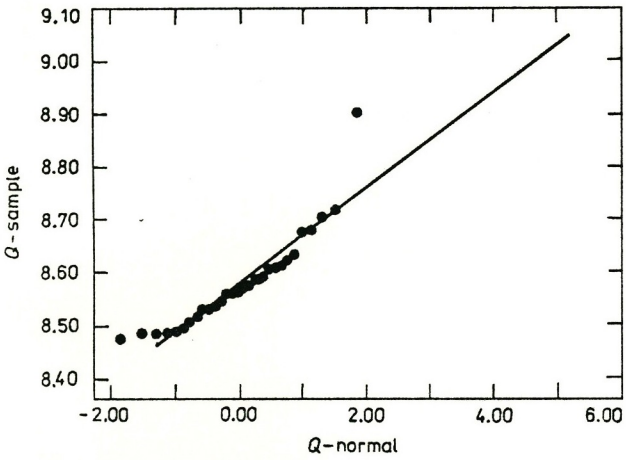

Fig. 4. Quantile-quantile plot of original sample concerning CaO content in glass.

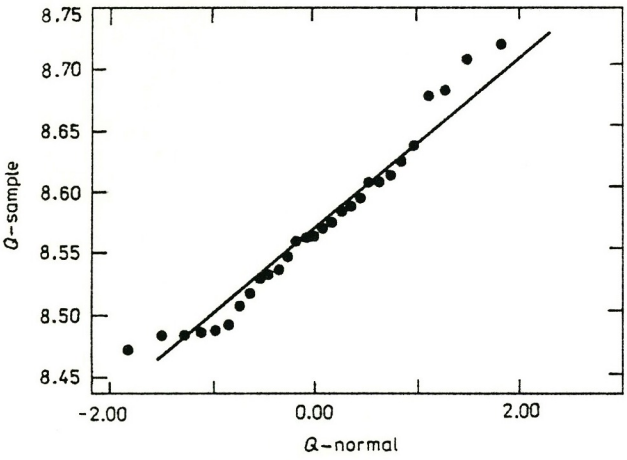

Fig. 5. Quantile-quantile plot of sample with excluded outlier concerning CaO content in glass.

represented by the mean is 8.581 ± 0.034. Excluding the outlier 8.902, the new mean is 8.570 ± 0.030. Robust estimates of original data lead to the values: 5 % trimmed mean is 8.572 ± 0.029, 10 % trimmed mean 8.569 ± 0.031, 40 % trimmed mean 8.568 ± 0.033, and the median 8.566 ± 0.033.

It is evident that the presence of one outlying point did not significantly corrupt the interval estimate of location.

CONCLUSION

Classical approach to instrumental data analysis in chemistry is based on some strong assumptions about statistical nature of data as an independence of sample elements, a sample normality, a sample homogeneity, and the minimal sample size. Besides statistical tests the graphical diagnostics of exploratory data analysis may be used. Often, the chemical data are less ideal and do not fulfil all these assumptions. The robust statistics are recommended mainly for a case of outliers in the sample.

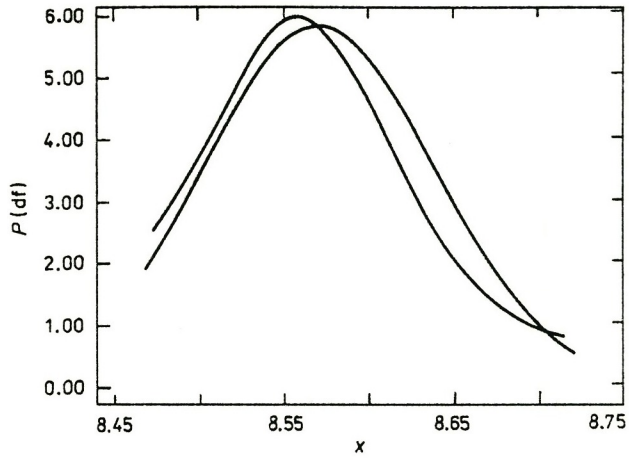

Fig. 6. Kernel estimate of the probability density function of sample concerning CaO content in glass when outlier is excluded. Upper curve: the empirical curve of sample distribution. Lower curve: the curve of Gaussian distribution.

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. Scott, D. W. and Sheater, S. J., *Commun. Statist. 14*, 1353 (1985).
5. Lejenne, M., Dodge, Y., and Koelin, E., *Proceedings of the Conference COMSTAT '82, Toulouse*. P. 173 (Vol. *III*).
6. Hoaglin, D. C., Mosteler, F., and Tukey, J. W. (Editors), *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, 1983.
7. Kafander, K. and Spiegelman, C. H., *Comput. Stat. Data Anal. 4*, 167 (1986).
8. Hines, W. G. S. and Hines, R. J. H., *Am. Statist. 41*, 21 (1987).
9. Hoaglin, D. C., *J. Am. Statist. Assoc. 81*, 991 (1986).
10. Stoodley, K., *Applied and Computational Statistics*. Ellis Horwood, Chichester, 1984.
11. Meloun, M., Militký, J., and Forina, M., *Chemometrics for Analytical Chemistry*, Part 1. *PC-Aided Statistical Data Analysis*. Ellis Horwood, Chichester, 1992. Part 2. *PC-Aided Regression and Related Methods*. Ellis Horwood, Chichester, 1994.
12. Meloun, M. and Militký, J., *Chem. Papers 48*, 151 (1994).
13. Meloun, M. and Militký, J., *Chem. Papers 48*, 164 (1994).
14. Meloun, M. and Militký, J., *Chem. Papers*, submitted for publication.
15. *Statistical package ADSTAT 2.0*. TriloByte, Pardubice, 1992.
16. Mendenhall, W. and Sincich, T., *Statistics for the Engineering and Computer Sciences*, p. 3. Dellen Publishing Company, San Francisco, 1988.
17. Špaček, M., *Thesis*. University Pardubice, 1990.

Translated by M. Meloun