Computer-Assisted Data Treatment in Analytical Chemometrics IV. Classical Estimates of Parameters of Location, Scale, and Shape

[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 24 June 1994

Classical point estimates of parameters of location, scale, and distribution shape are measures characterizing the sample distribution. More meaningful are interval estimates which for a specified degree of assurance cover an unknown value of the population parameter. Routine data treatment comprises the exploratory data analysis, the test of basic assumptions about a sample, and a calculation of the classical and robust estimates of the parameters of location, scale, and shape. For an analysis of small samples the Horn procedure of pivot measures is more suitable. Procedure of the univariate data treatment is demonstrated on the calibration of a pipette.

After the exploratory data analysis, the statistical analysis is a next step. With small samples the statistical characteristics are estimated directly, but with the large ones the data are divided into classes and the statistical characteristics of each class are estimated.

Univariate samples come from population with an unknown probability distribution. An univariate population is considered to be a set in which only one property is studied, and one quantity is measured. The population is characterized both by measures of the *location*, *i.e.* the level at which the quantity values vary, and by the degree of the *dispersion* (or *spread, scatter, scale, variability*) of the quantity of interest, and as such by the shape parameters of distribution. As the large population of all measured quantities is rarely available the *representative random sample* (or the *sample*) of few measurements is analyzed.

The sample is characterized by information about the *mean value* of the sample elements and their *variability* around this mean. Statistical characteristics of location, spread, and shape are called the *sample characteristics*. From these sample characteristics, the measures for the population are derived.

The main purpose of chemometrics experimentation is to draw inferences about a population from samples of the population. We can identify three different types of inferences, namely: *1.* the point estimation, *2.* the interval estimation, *3.* the hypothesis testing.

Estimation of a single value for a parameter is termed the *point estimation*. The *interval estimation* is concerned with estimation of interval that will include the population parameter with a specified probability. An interval estimate is more informative than a point estimate. Interval estimation is closely related to *hypothesis testing*.

This paper brings the classical point and interval estimates of parameters of location, scale, and distribution shape.

THEORETICAL

Point Estimates

Data samples are classically characterized by the *sample arithmetic mean* $\bar{x}$ and the *sample variance* s^2. These estimates can be used for characterization of data sampled from unknown distribution. If the sample comes from a symmetric population distribution with the mean μ, variance σ^2, and curtosis g_2, it can be proved that

$$E(\bar{x}) = \mu \qquad (1)$$

$$D(\bar{x}) = \frac{\sigma^2}{n} \qquad (2)$$

and

$$E(s^2) = \sigma^2 \qquad (3)$$

$$D(s^2) = \frac{\sigma^4}{n}\left[g_2 - \frac{n-3}{n-1}\right] \qquad (4)$$

In addition to the sample arithmetic mean and the sample variance, other parameters of location and scale can be used: the *sample modus* (or the *modus*

only) $\hat{x}_M$ is the most frequently found element value in the sample. The *sample quantiles* are descriptive statistics from the exploratory data analysis and are sometimes used to supplement the information obtained from the mean and the variance. Sample values $x_1, \ldots, x_n$ arranged in order of ascending magnitude, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ are called the *order statistics*. The *p-th quantile* (or the *percentile*) is defined to be the value of x below which p % of the sample value lie. The p-th quantile separates the order statistics into two parts so that each contains the required percent of the sample elements, p % and $(100 - p)$ %.

The *sample median* $\tilde{x}_{0.5}$ is the quantile that separates order statistics into two parts: 50 % of the elements lie below $\tilde{x}_{0.5}$ and 50 % of the elements lie above $\tilde{x}_{0.5}$. The sample median for odd sample size has the form

$$\tilde{x}_{0.5} = x_{(k)} \tag{5}$$

where $k = (n + 1)/2$. For an even sample size, it is

$$\tilde{x}_{0.5} = \frac{x_{(k)} + x_{(k+1)}}{2} \tag{6}$$

where $k = n/2$.

The 25th and 75th percentiles may be called the *first* (or *lower*) *quartile* and the *third* (or *upper*) *quartile* of the sample. The median represents the maximum likelihood estimate of location for the Laplace distribution. For this distribution the variance of median is expressed by

$$D_L(\tilde{x}_{0.5}) = \frac{\sigma^2}{2n} \tag{7}$$

For the normal distribution, however, the sample median is not efficient (see Table 1). For the rectangular distribution, the efficient estimate of location is the *midsum* $\hat{x}_P$ defined by

$$\hat{x}_P = \frac{x_{(1)} + x_{(n)}}{2} \tag{8}$$

where $x_{(1)}$ is the smallest and $x_{(n)}$ the largest element of an ordered sample. The variance of the midsum estimate for the rectangular distribution is defined by

$$D_R(\hat{x}_P) = \frac{6\sigma^2}{(n-1)(n-2)} \tag{9}$$

Index R denotes the rectangular distribution. The variance of $\hat{x}_P$ for normal distribution is much higher.

Often the condition of constant variance of all sample elements is not maintained. If each x_i has a normal distribution with variance σ_i^2, the statistical weight is calculated as $w_i = 1/\sigma_i^2$. Instead of sample mean $\bar{x}$, the weighted sample mean $\bar{x}_w$ is computed as follows

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} x_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2} \tag{10}$$

The variance of weighted mean is

$$D(\bar{x}_w) = \frac{1}{\sum_{i=1}^{n} 1/\sigma_i^2} \tag{11}$$

If the *relative error* has a constant value, $\delta = \sigma/x =$ constant, then $\sigma_i^2 = x_i^2\delta^2$. Then $w_i = 1/x_i$ and the sample mean can be calculated as

$$\bar{x}_w = \frac{\sum_{i=1}^{n} 1/x_i}{\sum_{i=1}^{n} 1/x_i^2} \tag{12}$$

with variance

$$D(\bar{x}_w) = \frac{\delta^2}{\sum_{i=1}^{n} 1/x_i} \tag{13}$$

The dispersion parameters describe the degree of dispersion (scale, spread, variability or scatter) of the population elements. The *range* is one of the measures of spread which represents the difference between the largest and the smallest value of sample. The *interquantile range* R_F is the quantile estimate of population standard deviation σ defined as

$$R_F = 0.7413\left(\tilde{x}_{0.75} - \tilde{x}_{0.25}\right) \tag{14}$$

where $\tilde{x}_{0.75}$ is the upper and $\tilde{x}_{0.25}$ the lower quartile.

Table 1 surveys the sample estimates of location and dispersion, with their variances, efficiency, and distribution. Sample estimates are for sample size n, and the sample comes from a population with normal distribution $N(\mu, \sigma^2)$.

Table 1. Estimates of Location and Dispersion for Sample of Size n from a Population with Normal Distribution $N(\mu, \sigma)$

Parameter	Estimate	Variance estimate	Efficiency	Estimate distribution
Mean μ	$\bar{x}$	σ^2/n	1	$N(\mu, \sigma^2)$
	$\tilde{x}_{0.5}$	$\sigma^2\pi/2n$	0.63	$N(\mu, \sigma^2)$
	$\hat{x}_P$	$\sigma^2\pi^2/(24 \ln n)$	$24 \ln n/(\pi^2 n)$	$N(\mu, \sigma^2)$
Variance σ^2	s^2	$2\sigma^4/(n-1)$	1	$N(\sigma^2, D(\sigma^2))$
Standard deviation σ	$\hat{\sigma}$	$\sigma^2/2n$	≈ 1*	
	s	$\sigma^2/(2(n-1))$	1	$N(\sigma, D(\sigma))$
	R	$\approx 1.36\ \sigma^2/n$	≈ 0.368	
	d	$\sigma^2/((\pi-2)n)$	≈ 0.876	

* Biased estimate.

Another measure of dispersion is the *mean deviation d* defined by the equation

$$d = \sqrt{\frac{\pi}{2}}\left[\frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|\right] \qquad (15)$$

where the factor $\sqrt{\pi/2}$ ensures that for normal distribution the value of d approaches that of the standard deviation σ.

The widely used *coefficient of variation* δ (denoted CV) also known as the *relative standard deviation* s_{rel} (denoted RSD) is given by 100 σ/μ and may be estimated by the relation

$$\hat{\delta} = \frac{s}{\bar{x}} \qquad (16)$$

Variance of $\hat{\delta}$ is approximately equal to

$$D(\hat{\delta}) = \hat{\sigma}^2\,\frac{n + \hat{\delta}^2(2n+1)}{2n(n-1)} \qquad (17)$$

The error $\hat{\delta}$, expressed in percents, is also called a *relative error*. Relative errors are frequently used in the comparison of the precision of results with different units or magnitudes, and are again important in calculations of error propagation.

To characterize the shape of a distribution, the skewness and curtosis are used. *Skewness* g_1 is a measure characterizing symmetry, which is equal to zero for a symmetrical distribution. The positive values of g_1 indicate smaller scattering of lower values of elements x_i than of the larger values and the negative values of g_1 indicate the opposite case. The *moment estimate of skewness* is defined by

$$\hat{g}_1 = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^3\right]^{3/2}} \qquad (18)$$

Its asymptotic variance is

$$D(\hat{g}_1) \approx \frac{(n-2)}{(n+1)(n+3)} \qquad (19)$$

The curtosis characterizes the peakedness of the distribution near a modal value and provides a picture of the shape of the distribution peak. For higher values of curtosis than 3, the distribution has a sharper peak than the normal distribution while a flat shape is indicated for curtosis lower than 3. The moment estimate of curtosis is defined by

$$\hat{g}_2 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \qquad (20)$$

Its asymptotic variance has the form

$$D(\hat{g}_2) \approx \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} \qquad (21)$$

When a point estimate of any parameter is determined, the variance of the parameter must also be calculated. To achieve the same "precision" of estimates when less effective estimates are used, a greater number of measurements n should be used. To achieve the same parameter precision for data of normal distribution, for example, the calculation of median $\hat{x}_{0.5}$ needs 1.6 times more measurements than the application of arithmetic mean $\bar{x}$

For samples coming from a population of normal distribution the random variable

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n} \qquad (22)$$

has the Student distribution with $(n-1)$ degrees of freedom. Also, the random variable

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \qquad (23)$$

has the χ^2-distribution with $(n-1)$ degrees of freedom. The random variable t and χ^2 are mutually independent. For sufficiently large samples ($n \geq 40$) from the normal distribution, for some estimate $\hat{\theta}$ of parameter θ, the random variable

$$U = \frac{\hat{\theta} - \theta}{D(\hat{\theta})} \qquad (24)$$

has an approximately standard normal distribution N (0, 1). Eqn (*24*) is asymptotically valid for any estimate $\hat{\theta}_i$ with variance $D(\hat{\theta}_i)$ determined by the maximum likelihood method, and for any theoretical distribution $f(x, \theta)$.

It is clear that the distribution of estimators is connected with sample distributions like the Student and χ^2 ones. The Student and χ^2-distributions are both among basic sample distributions which depend on degrees of freedom ν. For various values of degree of freedom ν, the quantiles of Student distribution and χ^2-distribution may be found in statistical tables.

Interval Estimates

More meaningful statement than the point estimate is the *confidence interval* which is calculated from the sample estimators. It includes the value of the population parameter within the interval limits, termed *confidence limits*, for a specified degree of assurance, called the *confidence coefficient*. Here, the confidence limits are random variables dependent on the sample.

The parameter of location is then described not by one value $\bar{x}$ but by two numerical values L_1 and L_2. It is expected that the confidence interval $\langle L_1, L_2 \rangle$ will include the unknown population parameter θ with preselected probability $(1-\alpha)$. The degree of trust associated with the confidence statement expresses the degree of certainty or reliability $(1-\alpha)$ about the unknown population parameter θ

$$P(L_1 < \theta < L_2) = 1 - \alpha \qquad (25)$$

where α is termed the significance level; the value chosen for α is usually 0.05 or 0.01. It is useful to know that *a*) the confidence interval is small if the variance of estimate $D(\theta)$ is small; *b*) a large sample size n gives a small confidence interval $\langle L_1, L_2 \rangle$; and *c*) higher degrees of certainty $(1-\alpha)$ give broader confidence intervals $\langle L_1, L_2 \rangle$.

Confidence interval $\langle L_1, L_2 \rangle$ is referred to as a two-tailed interval, but one-tailed intervals may be also used in chemical laboratory. One-tailed confidence intervals can be *a*) the left-side or lower-tail interval $\langle L_2, \infty)$, or *b*) the right-side or upper-tail interval $(-\infty, L_1 \rangle$.

Let us find the confidence interval of the population mean of the normal distribution $N(\mu, \sigma)$. Let $\bar{x}$ be the mean of a sample of n observations on a normally distributed random variable x with unknown mean μ and known variance σ^2. Then 100 $(1-\alpha)$ % confidence interval $L_{1,2}$ for μ may be found from

$$\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (26)$$

where $u_{1-\alpha/2}$ is 100 $(1-\alpha)$ % quantile of the standardized normal distribution (*e.g.* for $u_{0.975} = 1.96$ $L_{1,2} = x \pm 1.96\,\sigma/\sqrt{n}$).

In cases where the sample size n is not large enough and the variance σ^2 is not known, the confidence limit for μ may be found from eqn (*26*), but using quantiles for Student t-distribution instead of those for the normal one. The 100 $(1-\alpha)$ % confidence limits $L_{1,2}$ are then given by

$$\bar{x} - t_{1-\alpha/2}(\nu)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}(\nu)\frac{s}{\sqrt{n}} \qquad (27)$$

where $\nu = n-1$ is the number of degrees of freedom, $t_{1-\alpha/2}(\nu)$ is the 100 $(1-\alpha/2)$ % quantile of the Student distribution. For large sample sizes ($n > 30$) instead of $t_{1-\alpha/2}(\nu)$ the quantile $u_{1-\alpha/2}$ can be used.

According to eqn (*24*), the 100 $(1-\alpha)$ % asymptotic confidence interval of any parameter θ may be expressed by

$$\hat{\theta} - u_{1-\alpha/2}\sqrt{D(\hat{\theta})} \leq \theta \leq \hat{\theta} + u_{1-\alpha/2}\sqrt{D(\hat{\theta})} \qquad (28)$$

The 100 $(1-\alpha)$ % two-tailed confidence interval of the variance σ^2 is given by

$$\frac{\nu s^2}{\chi^2_{1-\alpha/2}(\nu)} \leq \sigma^2 \leq \frac{\nu s^2}{\chi^2_{\alpha/2}(\nu)} \qquad (29)$$

where $\chi^2_{1-\alpha/2}(\nu)$ is upper and $\chi^2_{\alpha/2}(\nu)$ lower quantile of χ^2-distribution, and $\nu = n-1$ is the number of degrees of freedom.

Construction of a confidence interval depends on the population distribution from which the sample comes. For example, the variance of the median may be calculated from the relation

$$D(\tilde{x}_{0.5}) = \frac{1}{4nf^2(\text{med})}$$

where $f(\text{med})$ is the value of the probability density function at the position of median. For Laplace distribution, $f(\text{med}) = 1/(\sigma\sqrt{2})$ and therefore $D(\tilde{x}_{0.5}) = \sigma^2/2n$, and the asymptotic confidence interval of the median is given by

$$\tilde{x}_{0.5} - u_{1-\alpha/2}\frac{0.707\,s}{\sqrt{n}} \leq \mu \leq \tilde{x}_{0.5} + u_{1-\alpha/2}\frac{0.707\,s}{\sqrt{n}} \quad (30)$$

Eqn (*30*) is valid only if the sample size n is big enough for the median of the Laplace distribution to have approximately normal distribution.

Analysis of Small Samples

The analysis of small samples is not reliable and results are usually rather uncertain. Small samples are used in cases when experiment repetition is expensive or scarcely possible.

For $n = 2$, the statistical analysis is very difficult. If observations are close enough, the arithmetic mean is calculated. If observations do not agree, it is not possible to say which is the outlier. The 100 $(1-\alpha)$ % confidence interval of the mean μ may be calculated by an approximation

$$\frac{x_1 + x_2}{2} - T_\alpha \frac{|x_1 - x_2|}{2} \leq \mu \leq \frac{x_1 + x_2}{2} + T_\alpha \frac{|x_1 - x_2|}{2} \quad (31)$$

The critical value of T_α depends on the distribution of data population from which the two values come. For the normal distribution it is $T_\alpha = \operatorname{cotg}(\pi\alpha/2)$ and for $\alpha = 0.05$ T_α is 12.71. For the rectangular distribution $T_\alpha = 1/\alpha - 1$, *i.e.* $T_\alpha = 19$, *cf.* Ref. [1].

For $n = 3$ it is also difficult to use a statistical analysis. The calculation of the arithmetic mean $\bar{x}$ from two near observations is better than the use of the median from all three values. The 100 $(1-\alpha)$ % confidence interval of the mean μ is then calculated by an approximation

$$\bar{x} - T_\alpha^* \frac{s}{\sqrt{3}} \leq \mu \leq \bar{x} + T_\alpha^* \frac{s}{\sqrt{3}} \quad (32)$$

For the normal distribution, $T_\alpha^* \approx 1/\sqrt{\alpha} - 3\sqrt{\alpha}/4 + \ldots$, and when $\alpha = 0.05$, T_α^* is 4.30. For rectangular distribution $T_\alpha^* = 5.74$, *cf.* Ref. [1].

For $4 \leq n \leq 20$ a procedure based on order statistics was introduced by *Horn* [1]. This is based on the depth which corresponds to the sample quartiles (the letter F). The pivot depth is expressed by $H_L = \text{int}((n+1)/2)/2$ or $H_L = \text{int}((n+1)/2+1)/2$ according to which H_L is an integer. The lower pivot is $x_L = x_{(H)}$ and the upper one is $x_U = x_{(n+1-H)}$. The estimate of parameter of location is then expressed by the *pivot halfsum*

$$P_L = \frac{x_L + x_U}{2} \quad (33)$$

and the estimate of parameter of spread is expressed by the *pivot range*

$$R_L = x_U - x_L \quad (34)$$

The random variable

$$T_L = \frac{P_L}{R_L} = \frac{x_L + x_U}{2(x_U - x_L)} \quad (35)$$

has approximately a symmetric distribution and its quantiles are in Table 2.

The 95 % confidence interval of the mean is expressed by pivot statistics as

$$P_L - R_L\, t_{L,0.975}(n) \leq \mu \leq P_L + R_L\, t_{L,0.975}(n) \quad (36)$$

and analogously hypothesis testing may be also carried out. For small samples ($4 \leq n \leq 20$), the pivot statistics lead to more reliable results than the application of Student's *t*-test or robust *t*-tests.

Table 2. Quantile $t_{L,1-\alpha}(n)$ of the T_L-Distribution [1]

n	$1-\alpha = 0.90$	0.95	0.975	0.99	0.995
4	0.477	0.555	0.738	1.040	1.331
5	0.869	1.370	2.094	3.715	5.805
6	0.531	0.759	1.035	1.505	1.968
7	0.451	0.550	0.720	0.978	1.211
8	0.393	0.469	0.564	0.741	0.890
9	0.484	0.688	0.915	1.265	1.575
10	0.400	0.523	0.668	0.878	1.051
11	0.363	0.452	0.545	0.714	0.859
12	0.344	0.423	0.483	0.593	0.697
13	0.389	0.497	0.608	0.792	0.945
14	0.348	0.437	0.525	0.661	0.775
15	0.318	0.399	0.466	0.586	0.685
16	0.299	0.374	0.435	0.507	0.591
17	0.331	0.421	0.502	0.637	0.774
18	0.300	0.380	0.451	0.555	0.650
19	0.288	0.361	0.423	0.502	0.575
20	0.266	0.337	0.397	0.464	0.519

COMPUTATION

1. When no preliminary information about data is available, the full exploratory data analysis (see Part I of this series) is applied.

2. Analyzing any new data batch the basic assumptions about data are always examined using i) a test for sample homogeneity; ii) a test for sample normality; iii) a test for independence of sample elements; and iv) a test for minimal sample size (see Part II of this series).

3. Analyzing routine data a knowledge of sample distribution is supposed and moreover distribution is

supposed to be normal and data elements should be homogeneous and independent otherwise the data transformation is applied (see Part III of this series).

4. Classical point and interval estimates of parameters of location, spread, and shape, *i.e.* the sample arithmetic mean, the sample variance (and related sample standard deviation and coefficient of variance), the skewness and curtosis sufficiently describe corresponding population the sample of medium size comes from.

5. For analysis of small samples the measures of location and spread are calculated by the Horn procedure being based on the pivot halfsum and the pivot range.

SOFTWARE

Module Univariate Data Analysis from the package ADSTAT contains procedures for an estimation of sample location, scale, and distribution shape. These procedures represent a part of programs for confirmatory data analysis of large and small samples.

RESULTS

Study Case 1. *Calibration of a pipette, with large sample size*

The pipette of volume 10 cm^3 was calibrated by weighing the water delivered and 32 measurements were obtained. The point and interval estimates of the real volume of the pipette should be determined.

Data: the pipette volume V/cm^3: 9.9889, 9.9820, 9.9656, 9.9940, 9.9877, 9.9865, 9.9755, 9.9820, 9.9794, 9.9184, 9.9848, 9.9914, 9.9905, 9.9726, 9.9661, 9.9857, 9.9889, 9.9832, 9.9923, 9.9877, 9.9779, 9.9936, 9.9666, 9.9903, 9.9666, 9.9713, 9.9762, 9.9840, 9.9723, 9.9999, 9.9887, 9.9921.

Solution: The first step of the univariate data treatment is the exploratory data analysis. The dot diagrams and the box-and-whisker plots for original data (n = 32) in Fig. 1*a* exhibit one very significant lower value and one higher value which can be understood as the outliers in sample. The probability density function (Fig. 1*b*) and the quantile plot (Fig. 1*c*) shows that the distribution is rather skewed to higher values and that the assumption of normality is not fulfilled either when using classical $\bar{x}$ and s^2 or robust (median) characteristics.

The point estimates of location, spread, and shape, $\bar{x}$ = 9.9807 cm^3, s = 0.0147 cm^3, skewness $\hat{g}_1 = -2.45$, and curtosis $\hat{g}_2 = 11.17$ prove that a distribution shape is skewed to higher values ($\hat{g}_1 < 0$) and exhibits a sharp peak ($\hat{g}_2 > 3$). Therefore, the sample mean and standard deviation cannot be taken as the final values and examination of assumptions about the data is carried out:

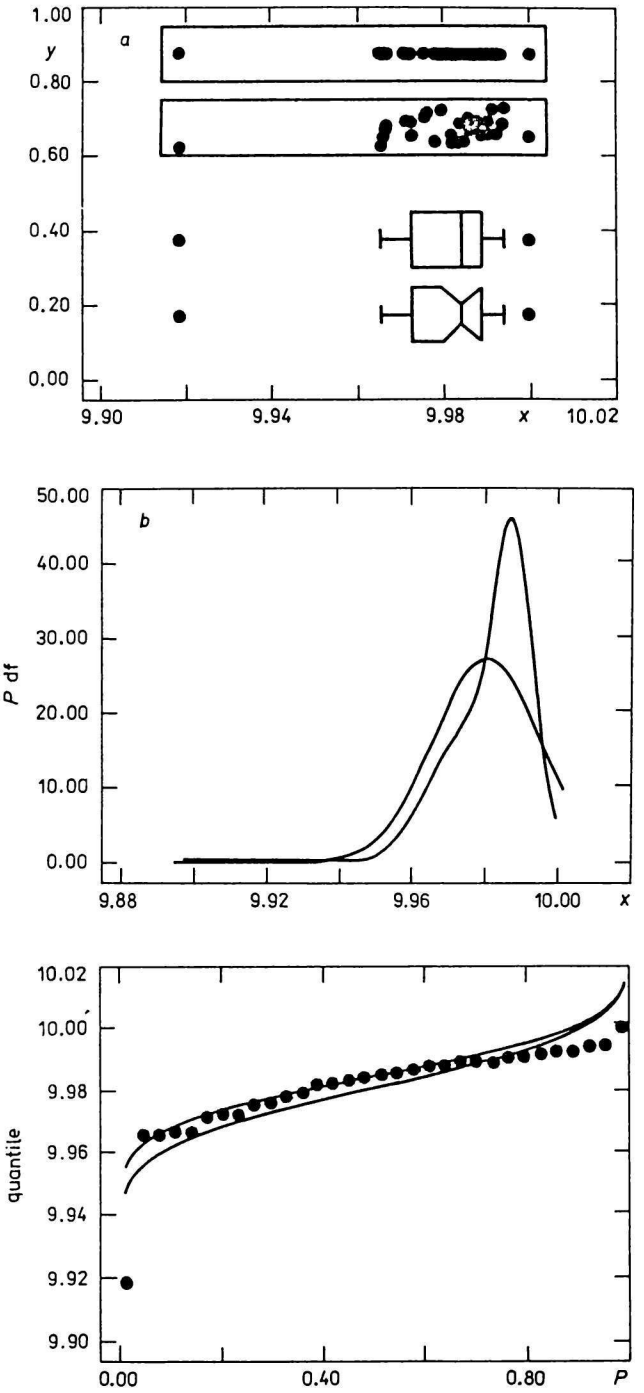


Fig. 1. Exploratory data analysis of the original sample: *a*) the dot diagrams and the box-and-whisker plots, *b*) the plot of the probability density function (sharper curve for robust and flatter one for classical estimates), and *c*) the quantile plot (upper curve for robust and lower one for classical estimates).

a) Data are independent of the time from the test criterion $t_n = 0.19 < t_{0.975}(32 + 1) = 2.035$;

b) Testing sample skewness and curtosis, C_1 = 185.7 reaches a higher value than the quantile $\chi^2_{0.95}(2)$ = 5.992 so the sample distribution is not Gaussian (normal);

c) Because the value $x_1 = 9.9184$ cm^3 lies outside modified external hinges $V_L^* = 9.9407$ cm^3 and $V_U^* = 10.0230$ cm^3, x_1 is a significant outlier and should be excluded from the sample. Estimates of location, spread, and shape of the resulting new sample are: $\bar{x} = 9.9827$ cm^3, $s = 0.0094$ cm^3, $\hat{g}_1 = -0.42$, and $\hat{g}_2 = 2.22$.

After exclusion of two outliers, $x_{(1)} = 9.9184$ cm^3 and $x_{(32)} = 9.9999$ cm^3, EDA plots (Fig. 2*a*—*c*) indicate no outliers in a sample and the distribution is closer to the normal one. The resulting sample is described by statistics: $\bar{x} = 9.9821$ cm^3, $s = 0.0090$ cm^3, $\hat{g}_1 = -0.54$, and $\hat{g}_2 = 2.02$ which can be taken as final.

Classical and robust statistics of the original sample ($n = 32$) are in Table 3. The classical arithmetic mean $\bar{x}$ is also calculated for the sample after elimination of one ($n = 31$) or two ($n = 30$) outliers.

Table 3. Point and Interval Estimates of Location

Parameter	Estimate $\hat{\mu}$/cm^3	Estimate $\hat{\sigma}$/cm^3	95 % Confidence interval	
			L_1/cm^3	L_2/cm^3
Mean $\bar{x}$, $n = 32$	9.9807	14.67×10^{-3}	9.9754	9.9860
Mean $\bar{x}$, $n = 31$	9.9827	9.43×10^{-3}	9.9793	9.9862
Mean $\bar{x}$, $n = 30$	9.9821	9.02×10^{-3}	9.9788	9.9855
Median $\tilde{x}_{0.5}$, $n = 32$	9.9844	12.54×10^{-3}	9.9791	9.9897

It may be concluded that assumption of normality is not fulfilled because two outliers are in the sample. Excluding outliers the relative error of pipette volume decreases from 0.026 % for the original data ($n = 32$) to 0.016 % for reduced data ($n = 30$). The use of robust estimate median is equivalent to excluding outliers from the sample. Calibration of pipette leads to the fact that the real volume is less than the declared one of 10 cm^3.

Study Case 2. *Calibration of a pipette, with small sample size*

The pipette of volume 25 cm^3 was calibrated by weighing the water delivered and 7 measurements were obtained. The point and interval estimates of the real pipette volume should be estimated.

Data: The pipette volume V/cm^3: 24.96439, 24.97758, 24.96809, 24.97409, 24.96880, 24.94759, 24.97119.

Solution: The point estimates of location, spread, and shape are $\bar{x} = 24.9670$ cm^3, $s = 0.0100$ cm^3, $\hat{g}_1 = -1.25$, and $\hat{g}_2 = 3.64$. Examination of assumptions about the sample data leads to the conclusions:

a) The test criterion $t_n = 0.61$ is smaller than the quantile $t_{0.975}(7 + 1) = 2.306$ and shows that the sample elements are independent.

b) The criterion $C_1 = 8.55$ is higher than the quantile $\chi^2_{0.95}(2) = 5.992$ so the sample has not a normal distribution.

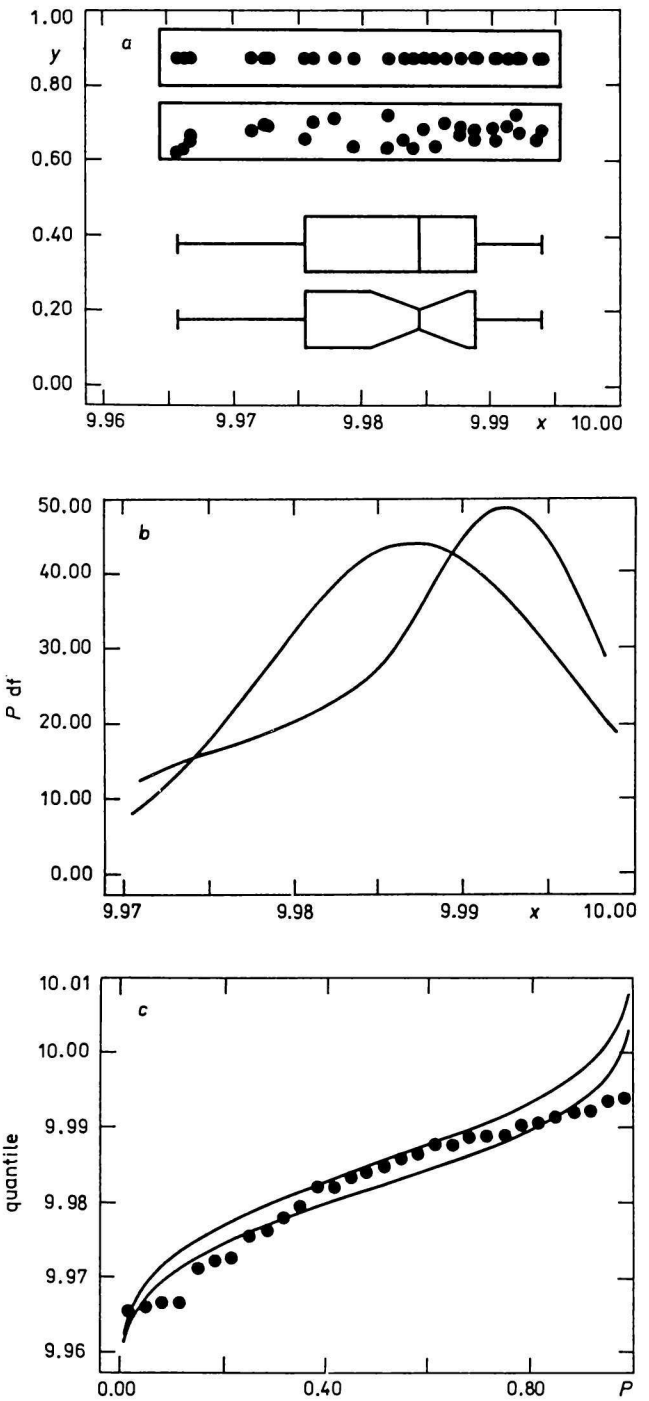
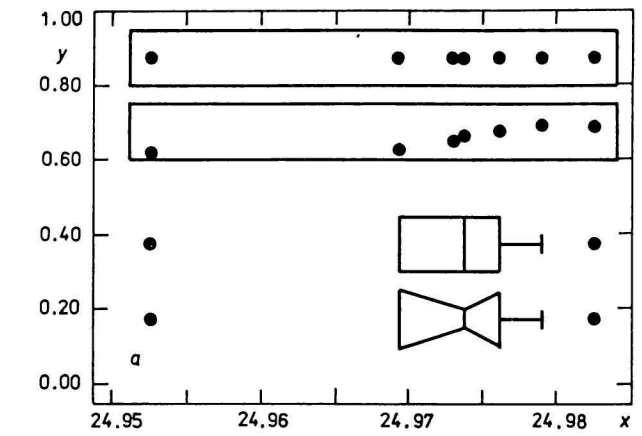


Fig. 2. Exploratory data analysis for the sample after elimination of two outliers: *a*) the dot diagrams and the box-and-whisker plots, *b*) the plot of the probability density function (sharper curve for robust and flatter one for classical estimates), and *c*) the quantile plot (upper curve for robust and lower one for classical estimates).

c) Outside the modified interval hinges $V_L^* = 24.955$ cm^3 and $V_U^* = 24.984$ cm^3 there is the value 24.94759 cm^3 (Fig. 3).

The 95 % confidence interval of the mean is

$$24.958 \text{ cm}^3 \leq \mu \leq 24.976 \text{ cm}^3$$



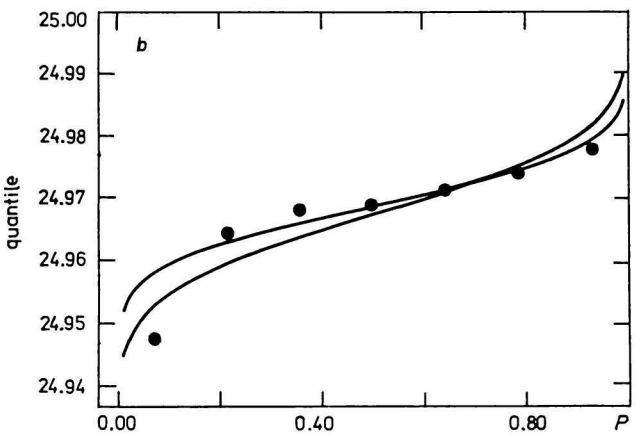


Fig. 3. Exploratory data analysis for the small sample: *a*) the dot diagrams and the box-and-whisker plots, and *b*) the quantile plot.

Horn procedure for the small samples leads to the following statistical characteristics: the pivot depth H_L = 2, the lower pivot is $x_{(2)}$ = 24.96439 cm^3 and the upper pivot is $x_{(6)}$ = 24.97409 cm^3. From the pivot halfsum P_L = 24.9692 cm^3 (eqn (*33*)), the pivot range R_L = 0.0097 cm^3 (eqn (*34*)) and from Table 2 the quantile $t_{L,0.975}(7)$ = 0.72, the 95 % confidence interval of the mean can be calculated

$$24.9622 \text{ cm}^3 \leq \mu \leq 24.9762 \text{ cm}^3$$

It may be concluded that for the small sample the pivot technique is more suitable. For a high precision of weighing, one outlier has a small influence on the estimate of the mean. Calibration of pipette leads to the fact that the real volume is less than the declared one of 25 cm^3

CONCLUSION

When in exploratory data analysis the tests of assumptions about sample data confirm that the sample comes from the population of the normal distribution, classical estimates of parameters of location, spread, and distribution shape sufficiently describe the sample. Interactive data investigation by modules of ADSTAT enables an easy determination of all classical estimates.

REFERENCES

1. Horn, J., *J. Am. Stat. Assoc. 78*, 930 (1983).

Translated by M. Meloun