Clin Chem Lab Med 2001; 39(1):53–61 © 2001 by Walter de Gruyter · Berlin · New York

Analysis of Large and Small Samples of Biochemical and Clinical Data

Milan Meloun[1], Martin Hill[2], Jiří Militký[3] and Karel Kupka[4]

[1] Department of Analytical Chemistry, Faculty of Chemical Technology, Pardubice University, Pardubice, Czech Republic
[2] Institute of Endocrinology, Prague, Czech Republic
[3] Department of Textile Materials, Technical University, Liberec, Czech Republic
[4] Trilobyte Statistical Software Ltd., Pardubice, Czech Republic

Statistical software often offers a list of various descriptive statistics of location and scale, but rarely selects an efficient estimate that is statistically adequate for an actual univariate sample. The sample interval estimate for a specified degree of uncertainty seems to be more meaningful if it covers an unknown value of the population parameter. The concept of an interval estimate in medicine is then used for medical decision-making. The proposed methodology, which uses the S-Plus algorithm for biochemical, biological and clinical data analysis contains the following steps: (i) Exploratory data analysis identifies basic statistical features and patterns of the data, the distributions of which are mostly non-normal, non-homogeneous and often corrupted by outliers. (ii) Sample assumptions about data, independence of sample elements, normality and homogeneity are examined. (iii) Power transformation and the Box-Cox transformation to improve sample symmetry and stabilize the spread. (iv) Classical and robust statistics for both large (n>30) and medium-sized samples (15<n<30), point and interval estimates for the parameters of location, scale and shape. For an analysis of small samples (4 < n < 20) the Horn procedure of pivot measures is recommended. The proposed methodology is demonstrated in two case studies, a large sample analysis of mean pregnenolone concentrations in the umbilical blood of newborns, and a small sample analysis of mean haptoglobin concentrations in human serum.

Key words: Statistics, M-estimate; Small samples; 3/3-hydroxy-5en-steroids; Pregnenolone; Sample; Statistical distribution.

Abbreviations: CDA, confirmatory data analysis; EDA, exploratory data analysis; RSD, relative standard deviation.

1. Introduction

The aim of the analysis of biochemical, biological and clinical data is the extraction of relevant information. With the advent of computers and sophisticated analytical instruments, the evaluation and interpretation of results seems to be the main problem. Due to the well-known fact that much experimental data in biochemistry exhibits a non-normal asymmetric distribution, classical analyses based on the assumption of normality cannot be employed; moreover, measurements are often corrupted by outliers. Tukey (1) has claimed that the techniques allowing the isolation of certain basic statistical features and patterns of data can be collectively named the exploratory data analysis (EDA). The EDA toolbox represents a collection of classical and computer-assisted parametric, non-parametric and function estimation methods for graphical visualization and data treatment (1–5). After an exploratory data analysis, a confirmatory data analysis (CDA) delivers measures of how adequate a model is.

Data generally come from populations with an unknown probability distribution. The corresponding univariate population is characterized by (i) measures of the *location* or *central tendency*, (ii) the degree of the *dispersion* (or *spread, scatter, scale, variability*), and (iii) the *shape* parameters of distribution. As a large population of all possible measured quantities is rarely available, a *representative random sample* (or just *sample*) of a few quantities (or measurements, elements) is analyzed. The sample is characterized by information about the *mean value* of the sample elements and their *variability* around this sample mean. The main purpose of analysis is to draw inferences about a population from the study of samples.

The population parameters are estimated by statistics computed from data termed the *point estimate*. The *interval estimate* gives a range of possible values of the population parameter with a pre-chosen probability. The interval estimate is more informative than the point estimate. Interval estimates can be used for testing of single hypotheses about population mean. If the assumption of normality is valid, the interval estimate for the mean, variance, and standard deviation may be computed very simply. The concept of an interval estimate, also called *a reference interval in medicine*, is based on determining a set of values within which some percentage, *e.g.* 95%, of the values of a particular analysed variable in a healthy population would fall. This interval is then used for medical decision-making. Recommendations on how to obtain such reference intervals have focused on the types of statistics best used to calculate such a reference interval (4). However, the standard error of each of these intervals depends on the sample variance, which is sensitive to outliers. In the case of non-normal distribution, simple variance is obviously inflated and classic inter-

val estimates are not applicable. When an exploratory data analysis (1–3) indicates that the sample distribution strongly differs from a normal one, the problem arises as to how to analyze such biochemical or medical data. Raw data may require re-expression to produce an informative display, an effective summary, or a straightforward analysis (6–12). Difficulties may arise because the raw data have (i) a strong asymmetry, or (ii) no constant variance. Altering the distribution by use of *e.g.* data transformation may alleviate these problems.

This paper focuses on classical and robust estimates of parameters of location, scale and distribution shape. When the sample distribution is systematically skewed or exhibits a variance heterogeneity, power transformation or Box-Cox transformation can improve sample symmetry and also stabilize variance. The purpose of this paper is to propose an efficient methodology for the treatment of biochemical and clinical data treatment, and also to show the usefulness of robust statistical analysis for obtaining a good interval estimate, even with a small number of sample elements. The proposed methodology of univariate data treatment is demonstrated by the two case studies, determinations of the mean pregnenolone concentration in the umbilical blood of newborns and the mean haptoglobin concentration in human serum.

2. Theoretical

2.1 Point estimates of location and spread parameters

Location

Data are classically characterized by the *sample arithmetic mean* $\bar{x}$ and the *sample variance* s^2. Assuming that data come from a symmetric distribution, characterized by the mean μ, the variance σ^2, the skewness g_1 being equal to zero, and kurtosis g_2, it can be proven that

$$E(\bar{x}) = \mu \text{ with } D(\bar{x}) = \frac{\sigma^2}{n}, \text{ and } E(s^2) = \sigma^2$$

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

where *E(.)* is an operator of mathematical expectation and *D(.)* is an operator of dispersion. In addition to the sample arithmetic mean and the sample variance, other parameters of location and scale can be used: the *sample mode* (or *modes*) $\hat{x}_M$ are the values around which the data are generally locally concentrated. Sample values $x_1, \ldots, x_n$ arranged in order of ascending magnitude, $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ are called the *order statistics*. The *p-th sample quantile* (or *percentile*) is defined as the value of x below which p% of the sample value lies. The p-th *sample quantile* separates the order statistics into two parts, each containing the required percentage of the sample elements, p% and (100–p)%. If the sample is not normally distributed, or if some outliers are present, the efficiency of both $\bar{x}$ and s^2 decreases. Some statistics still remain approximately correct for reasonable departures from normality; in this regard they are said to be *robust* to non-normality. Robustness can relate to the separate effects of deviations from normality, independence, equal variance, and randomness.

The *sample median* $\tilde{x}_{0.5}$ separates order statistics into two parts: 50% of the elements lie below $\tilde{x}_{0.5}$ and 50% of the elements lie above $\tilde{x}_{0,5}$. For odd sample sizes it takes the form $\tilde{x}_{0.5} = x_{(k)}$, where $k=(n+1)/2$ and for even sample sizes $\tilde{x}_{0.5}=(x_{(k)}+x_{(k+1)})/2$, where $k=n/2$. The 25th and 75th percentiles are called the *lower quartile* and *upper quartile* of the sample. The median represents the maximum likelihood estimate of location for the Laplace distribution (double exponential). For this distribution, the variance of the median is expressed by $D_L(\tilde{x}_{0.5})=\sigma^2/2n$. For a normal distribution, however, the sample median is not efficient (Table 1). For a rectangular (box type) distribution, an efficient estimate of location is the *midsum* x_P defined by $\hat{x}_P=(x_{(1)}+x_{(n)})/2$, where $x_{(1)}$ is the smallest and $x_{(n)}$ the largest element of an ordered sample. The variance of the midsum estimate for a rectangular distribution is defined by $D_R(\hat{x}_P)=6\sigma^2/[(n-1)(n-2)]$, where the index R denotes the rectangular distribution. The variance of $\hat{x}_P$ for the normal distribution is much higher.

Often the condition of constant variance in all sample elements is not maintained. If each x_i has a normal distribution with variance σ_i, the statistical weight is

Tab. 1 Estimates of location and dispersion for a sample of size n from a population with normal distribution $N(\mu, \sigma^2)$.

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	$N(\mu, \sigma^2)$
	$\hat{x}_P$	$\sigma^2\pi^2/(24 \ln n/(\pi^2 n))$	$24 \ln n/(\pi^2 n)$	$N(\mu, \sigma^2)$
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	
	R	$1.36\ \sigma^2/n$	0.368	$N(\sigma, D(\sigma))$
	d	$\sigma^2/((\pi-2)n)$	0.876	

*) Explanation of terms efficiency and bias and symbols is described on the elementary level in the book [4]

calculated as $w_i=1/\sigma_i^2$. Instead of sample mean $\bar{x}$, the *weighted sample mean* $\bar{x}_w$ is computed from

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} x_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2} \qquad [1]$$

The variance of the weighted mean is $D(\bar{x}_w)=1/(\sum_{i=1}^{n} 1/\sigma_i^2)$.

One of the simplest and most efficient robust estimates of location is the *trimmed mean* $\bar{x}(\kappa)$ which is defined with the use of the order statistics $x_{(i)}$ as

$$\bar{x}(\kappa) = \frac{1}{n-2M} \sum_{i=M+1}^{n-M} x_{(i)} \qquad [2]$$

where M=int $(\kappa n/100)$. The parameter κ determines the percentage of order statistics $x_{(i)}$ that are to be trimmed off at each, low and high, tail. The usual value of κ is 10%, and this results in the 10% trimmed mean $\bar{x}$ (10). When there are many outliers $\bar{x}$(25), or $\bar{x}$(40) may be preferred. Trimmed mean is not efficient for cases when data obey a strongly asymmetric distribution.

Robust *M estimates* represent the maximum likelihood estimates of parameters for a selected distribution. When the selected distribution has long tails, the corresponding M estimate of location is robust. The *M-estimate of location parameter* $\hat{\mu}_M$ is generally defined by

$$\hat{\mu}_M = \sum_{i=1}^{n} x_i w_i \Big/ \sum_{i=1}^{n} w_i \qquad [3]$$

where $w_i=W[(x_i-\mu_M)/\sigma]$ and $W(u)=d\rho(u)/dt$. For a robust estimate the function $W(u)$ must be bounded; a bi-quadratic function $W(u)$ of the following type is recommended

$$W(u) = \begin{cases} \left(\left[1-\left\{\frac{u}{4.69}\right\}^2\right]^2\right) & \text{for } |u| < 4.69 \\ 0 & \text{for } |u| \geq 4.69 \end{cases} \qquad [4]$$

where the numerical constant 4.69 means that for normally distributed data the asymptotic efficiency of estimate of location $\hat{\mu}_M$ is equal to 0.95. Since the standard deviation σ is usually unknown, it is replaced by a suitable robust estimate. For the *M-estimate of standard deviation* the recommended expression is:

$$s_w = \sqrt{\frac{\sum_{i=1}^{n} V_i (x_i-\bar{x}_w)^2}{\sum_{i=1}^{n} V_i}} \quad \text{where } V_i = W\left[\Delta\left[\frac{x_i-\bar{x}_w}{s_w}\right]\right] \qquad [5]$$

The weight function $W(u)$ is defined above and Δ(u) is a deviation function for which

$$\Delta(u) = \begin{cases} u^2-\ln u^2-1 & \text{for } u \neq 0 \\ \infty & \text{for } u = 0. \end{cases}$$

Spread

Dispersion parameters describe the degree of dispersion, (scale, spread, variability or scatter) of the population elements. *Range* is a measure of spread which represents the difference between the largest and the smallest values in the sample. The *interquantile range* R is the quantile estimate of population standard deviation σ defined by

$$R=0.7413\,(\tilde{x}_{0.75}-\tilde{x}_{0.25}) \qquad [6]$$

where $\tilde{x}_{0.75}$ is the upper F_U and $\tilde{x}_{0,25}$ the lower quartile F_L. The main resistant rule for identifying outliers sets up *Hoaglin's inner bounds* B^* beyond the quartiles, $B_L^*=F_L-1.5\ R$, $B_U^*=F_U+1.5\ R$. Any observations that fall below the lower inner bound B_L^* or above the upper inner bound B_U^* are then termed the *outliers.*

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 is drawn from a population with a normal distribution $N(\mu, \sigma^2)$. The widely-used *coefficient of variation* δ (or *CV*), also known as the *relative standard deviation* s_{rel} (or *RSD*), is given by 100 σ/μ and may be estimated by the relationship $\hat{\delta}=s/\bar{x}$. The variance of $\hat{\delta}$ is approximately equal to $D(\delta)=\sigma^2[n+\delta^2\ (2n+1)]/[2n(n-1)]$. In descriptive statistics the error δ, expressed as a percentage, is also called the *relative error*. Relative errors are frequently used in the comparison of the precision of results with different units or of different magnitudes, and are again important in calculations of error propagation.

Shape

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

$$\hat{g}_1 = \frac{\sqrt{n}\sum_{i=1}^{n} (x_i-\bar{x})^3}{\left[\sum_{i=1}^{n} (x_i-\bar{x})^2\right]^{3/2}} \qquad [7]$$

with its asymptotic variance $D(\hat{g}_1) \approx [6n(n-1)]/[(n-2)(n+1)(n+3)]$.

The *kurtosis* characterizes the peakedness of the distribution near a modal value, and provides a picture of the shape of the distribution peak. For values of kurtosis greater than 3 the distribution has a sharper peak or longer tails than a normal distribution, while a flat shape is indicated for kurtosis values of less than 3. The *moment estimate of kurtosis* is defined by

$$\hat{g}_2 = \frac{n\sum_{i=1}^{n} (x_i-\bar{x})^4}{\left[\sum_{i=1}^{n} (x_i-\bar{x})^2\right]^2} \qquad [8]$$

with its asymptotic variance $D(\hat{g}_2) \approx 24\ n(n-1)^2/[(n-2)(n-3)(n+3)(n+5)]$. When the 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 with normal distribution, for example, the calculation of median $\hat{x}_M$ needs 1.6 times more measurements than the application of arithmetic mean $\bar{x}$.

2.2 Interval estimates of location and spread parameters

A more meaningful statement than the point estimate is the *confidence interval* for population parameters. This 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 confidence interval is bounded by numerical values, the lower limit L_L and the upper limit L_U. It is expected that the confidence interval (L_L, L_U) will include the unknown population parameter θ with preselected probability (1–α). The degree of trust associated with the confidence statement is called the *confidence coefficient*; it expresses the degree of statistical certainty (1–α) regarding the unknown population parameter θ, $P(L_L < \theta < L_U)=1-\alpha$, 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 (i) the confidence interval is small if the variance of estimate $D(\theta)$ is small; (ii) a large sample size n gives a small confidence interval [L_L, L_U]; and (iii) higher degrees of certainty (1–α) give broader confidence intervals [L_L, L_U].

Confidence interval [L_L, L_U] is referred to as a two-tailed interval, but one-tailed intervals may also be used in the biochemical laboratory. One-tailed confidence intervals can be left-side (lower-tail) interval [L_L, ∞), or the right-side (upper-tail) intervals (–∞ , L_U].

Large samples, n>30

To 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. The 100(1–α)% confidence interval $L_{L,U}$ for μ may then be found from

$$\bar{x}-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad [9]$$

where $u_{1-\alpha/2}$ is the 100(1–α/2)% quantile of the standardized normal distribution, (*e.g.*, for α=0.05 $u_{0.975}$=1.96 and $L_{L,U}=\bar{x} \pm 1.96\,\sigma/\sqrt{n}$).

Medium and small samples, n<30

In cases where the sample size n is not large enough, $n<30$ and the variance σ^2 is not known, the confidence limit for μ may be found, but using quantiles for a Student t-distribution instead of a normal one. The 100(1–α)% confidence limits $L_{L,U}$ are then given by

$$\bar{x}-t_{1-\alpha/2}(\nu)\frac{s}{\sqrt{n}} \le \mu \le \bar{x}+t_{1-\alpha/2}(\nu)\frac{s}{\sqrt{n}} \qquad [10]$$

where ν=n–1 is the number of degrees of freedom, and $t_{1-\alpha/2}(\nu)$ is the 100(1–α/2)% quantile of the Student distribution.

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

$$\frac{vs^2}{\chi^2_{1-\alpha/2}(\nu)} \le \sigma^2 \le \frac{vs^2}{\chi^2_{\alpha/2}(\nu)} \qquad [11]$$

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

2.3 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, then the arithmetic mean is calculated. If observations do not agree, it is not possible to say which is the outlier. The 100(1–α)% confidence interval of the mean μ may be calculated by the approximation

$$\frac{x_1+x_2}{2} - T_\alpha\frac{|x_1-x_2|}{2} \le \mu \le \frac{x_1+x_2}{2} + T_\alpha\frac{|x_1-x_2|}{2} \qquad [12]$$

The critical value of T_α depends on the distribution of the data population from which the two values are drawn. For a normal distribution it is T_α=cotg(πα/2) and for α=0.05, T_α is 12.71. For the rectangular distribution T_α=1/α–1, *i.e.* for α=0.05 is T_α=19 (14).

For n=3 it is also difficult to use 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–α) % confidence interval of the mean μ is then calculated by the approximation

$$\bar{x}-T^*_\alpha\frac{s}{\sqrt{3}} \le \mu \le \bar{x}+T^*_\alpha\frac{s}{\sqrt{3}} \qquad [13]$$

For a normal distribution, $T^*_\alpha \approx 1/\sqrt{\alpha}-3\sqrt{\alpha}/4+\ldots$, and when α=0.05, T^*_α is 4.30. For a rectangular distribution T^*_α=5.74 (14).

For $4 \le n \le 20$ a procedure based on order statistics was introduced by Horn (13, 15). This is based on a depth which corresponds to the sample quartiles. The pivot depth is expressed by H_L=int((n+1)/2)/2 or H_L=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 is $x_U=x_{(n+1-H)}$. The estimate of the parameter of location is then expressed by the *pivot halfsum*

$$P_L=(x_L+x_U)/2 \qquad [14]$$

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

$$R_L=x_U-x_L \qquad [15].$$

The random variable $T_L=P_L/R_L=(x_L+x_U)/[2(x_U-x_L)]$ has an approximately symmetrical distribution and its quan-

Tab. 2 Quantile $t_{L,\,1-\alpha}(n)$ of the T_L-distribution (15).

n	1–α=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

tiles are given in Table 2. The *95% confidence interval of the mean* is expressed by pivot statistics as

$$P_L - R_L\, t_{L,0.975}(n) \le \mu \le P_L - P_L\, t_{L,0.975}(n) \qquad [16]$$

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

3. Experimental

3.1 Proposed procedure

Many statistical software packages offer a list of various point parameters of location and spread, but rarely help the user to choose the only parameter *i.e.* the best and efficient estimate, which is statistically adequate for an actual sample batch. Exploratory data analysis and an examination of sample assumptions will provide an answer to this question.

1st step: Exploratory data analysis

When no preliminary information about data is available, a full exploratory data analysis is applied (1–4): for a graphical visualization of data, the quantile plot, dot diagram and jitter – dot diagram, the box-and-whisker plot and the notched box-and-whisker plot are supported. Sample distribution (represented by the symmetry and tail lengths, skewness and kurtosis) is investigated by the midsum plot and the symmetry plot. Construction of an actual sample distribution, *i.e.* the estimation of the probability density function, is carried out by a kernel density estimator of the probability density function, and the quantile-quantile plot is used to compare the sample distribution with the theoretical ones.

2nd step: Sample assumptions

In the analysis of any new data batch, the basic assumptions about the sample are always examined using a check for sample homogeneity, a check for sample normality, a check for the independence of sample elements and a check for minimal sample size (16).

3rd step: Data transformations

In the analysis of the routine data, the sample distribution is taken to be known. Moreover, distribution is assumed to be normal and data elements to be homogeneous and independent – otherwise data transformation should be applied (4, 17). Two procedures, power transformation and Box-Cox transformation, of the ADSTAT statistical systems (18), search parameters of a simple power transformation and parameters of the normalized Box-Cox transformation of data. Using transformed data, the mean $\bar{y}$, the variance $s^2(y)$, the skewness $\hat{g}(y)$, and the kurtosis $\hat{g}_2(y)$ are calculated. From these estimates, the re-expressed estimates of original variables $\bar{x}_R$, $s^2(\bar{x}_R)$, and the 95% confidence interval of the re-expressed variable μ are then calculated (17, 18).

4th step: Classical and robust statistics

The estimates of the parameters of location, scale and shape are calculated:

(a) Large samples: parameters of location: When investigating the center of a variable, the main descriptors are the mean, median, mode, and trimmed mean. Other averages, such as the geometric mean and harmonic mean, have specialized uses. If the data come from a symmetrical, normal distribution, the mean, median, mode and trimmed mean are all equal. If the mean and median are very different, then most likely there are outliers in the data or the distribution is skewed. If this is the case, the median is probably a better measure of location. The mean is very sensitive to extreme values and can be seriously contaminated by just one outlying observation. The trimmed mean is more robust than the mean, but is more sensitive than the median. The re-expressed estimate $\bar{x}_R$ of the power or Box-Cox transformation is one the most reliable estimates of the center.

(a) Large samples: parameters of spread: The next question is how closely the data fall about the center. There are numerous measures of variability: range, variance, standard deviation, interquantile range, pivot range, *etc.* All of these measures of spread or dispersion are affected by outliers to some degree, but some do much better than others. The standard deviation is one of the most popular measures of spread; unfortunately, it is greatly influenced by outliers and by the overall shape of the distribution. Robust alternatives are deviations based on the interquantile range.

(a) Large samples: parameters of shape: Skewness measures the direction and lack of symmetry. The more skewed a distribution is, the greater the need to employ a data transformation technique. Positive

skewness indicates a longtailedness to the right, while negative skewness indicates longtailedness to the left. Kurtosis measures the heaviness of the tails: a kurtosis value less than 3 indicates lighter tails than in a normal distribution, while kurtosis values greater than 3 indicate heavier tails than in a normal distribution. For heavy tailed distribution the median or generally robust methods are suitable.

(b) *Small samples:* For the analysis of small samples the measures of location and spread are calculated by the Horn procedure (15), this being based on the pivot halfsum and the pivot range.

3.2 Software used

The algorithm in *S-Plus* (19) was written for the creation of EDA diagnostic graphs and the computation of the quantile based characteristics of a sample distribution, and English version of ADSTAT software (18) with procedures for the estimation of sample location, scale and distribution shape was also used. The algorithm in S-Plus is available on request (18).

4. Illustrative Examples

The first case study using the proposed methodology runs on typical biochemical sample data, and illustrates the procedure of the statistical treatment of univariate data.

Case study 1. Determination of the pregnenolone concentrations in the umbilical blood of newborns

Low levels of free 5-ene steroids in umbilical blood and elevated levels of 5-ene steroid sulfates indicate a congenital sex-specific placental sulfatase insufficiency (20). A delayed onset of labor, frequently linked to the necessity of intervention (21) together with relatively low birth weights, is the common symptom of the disease. A defect of recessive X-linked type, also called the "dry skin" disease, may have phenotypic consequences in later postnatal life (22). The incidence of this disorder appears to be approximately one per 2000 male births (23). An exact assessment of the mean value and of the variance of steroid levels in controls is necessary for the correct judgment of samples from patients. The levels of pregenolone were evaluated and this evaluation was chosen as an example of correct data analysis. Statistical assumptions should be tested on the group of umbilical blood samples from newborns using plots of an extended exploratory data analysis, and statistical tests of the basic assumptions. A reliable estimate of the mean pregnenolone value is to be enumerated.

Solution

1. Descriptive statistics: NCSS2000 (24) statistical software for an actual sample batch calculates a survey of the parameters of location and spread for n=100 (for an explanation of the statistics see ref. 4): the arithmetic mean $\bar{x}$= 80.6 nmol/l, the median $\hat{x}_{0.5}$=66.0 nmol/l, the geometric mean $\bar{x}_g$=61.8 nmol/l, the harmonic mean $\bar{x}_h$=46.4 nmol/l, the mode $\bar{x}_M$=107 nmol/l, and following trimmed means $\bar{x}(5\%)$=74.7 nmol/l with $s(5\%)$=44.4 nmol/l and $n(5\%)$=90, $\bar{x}(10\%)$=71.6 nmol/l with $s(10\%)$=36.0 nmol/l and $n(10\%)$=80, $\bar{x}(25\%)$=67.9 nmol/l with $s(25\%)$=22.4 nmol/l and $n(25\%)$=50, $\bar{x}(45\%)$=66.2 nmol/l with $s(45\%)$=5.6 nmol/l and $n(45\%)$=10. Survey of parameters of spread: the variance s^2=3658.7, the standard deviation s=60.5 nmol/l, the unbiased standard deviation s=60.6 nmol/l, the interquantile range R_F=70.0 nmol/l. Finally a survey of parameters of shape: the skewness $\hat{g}_1$=1.57, the kurtosis $\hat{g}_2$=6.06. As the numerical values of the arithmetic mean, median, mode and trimmed mean are very different, there are definitely outliers in the data or the distribution is strongly skewed. The median will probably be a better measure of location. On the basis of EDA, the investigator should select the most convenient parameter of location for the actual sample batch.

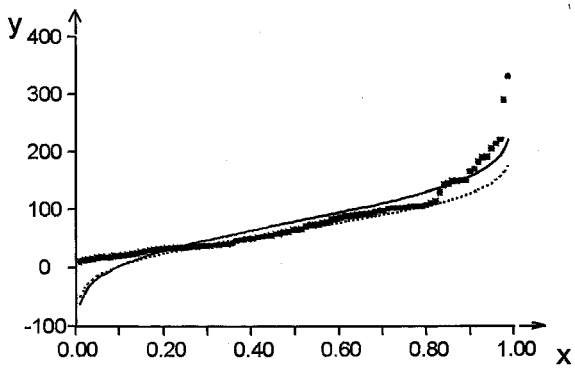

Fig. 1 Quantile plot of pregnenolone data showing the empirical curve (dot curve) and the normal distribution approximation (full curve).

Tab. 3 Input data: Univariate sample of concentration of pregnenolone (nmol/l), n=100.

33.6	35.4	32.8	150.0	149.0	98.8	107.0	91.9	37.7	67.1	107.0	183.0
332.0	77.5	104.0	50.0	93.1	206.0	93.4	215.0	107.0	292.0	36.4	152.0
71.4	222.0	47.6	66.1	75.3	48.8	89.2	78.1	87.1	170.0	89.2	54.6
97.0	20.6	26.6	31.4	90.6	18.0	12.2	25.2	39.8	61.6	21.5	17.6
106.0	39.4	23.8	191.0	33.5	41.0	62.7	56.5	59.8	58.7	29.8	53.3
13.8	192.0	42.8	38.4	32.2	104.0	17.7	102.0	109.0	84.5	41.2	145.0
152.0	167.0	18.4	50.9	143.0	14.8	99.4	10.8	92.1	34.4	65.8	34.5
131.0	34.1	74.2	103.0	20.3	40.5	83.9	74.2	51.1	27.8	116.0	105.0
23.0	37.0	55.8	112.0								

2. Exploratory data analysis is used for a graphical visualization of data: the quantile plot (Figure 1) shows a strong deviation from a normal distribution, as the sample points do not fit a classic curve, and two outliers at high values are indicated. Both dot diagrams (Figure 2) and the box-and-whisker plot (Figure 3) indicate five outliers at high values and an asymmetric, skewed distribution. In the midsum plot (Figure 4) and in the symmetry plot (Figure 5) most sample points are outside the confidence limits, and both diagnostic plots indicate that the sample distribution is strongly skewed.

3. Sample distribution represented by symmetry, skewness and kurtosis is examined by two plots: the kernel density estimator of the probability density function (Figure 6) indicates a skewed sample distribution with several outliers, while the *Q-Q* rankit plot (Figure 7) checking a normal distribution does not exhibit close agreement of the sample points with a straight line.

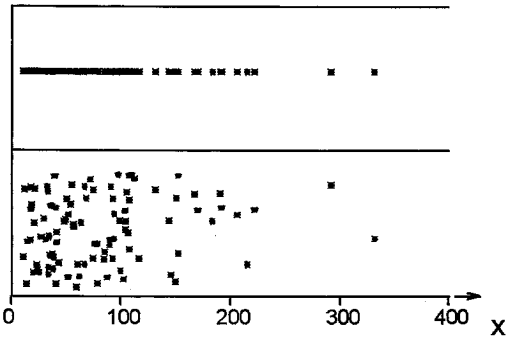

Fig. 2 Dot and jitter dot diagram of pregnenolone data.

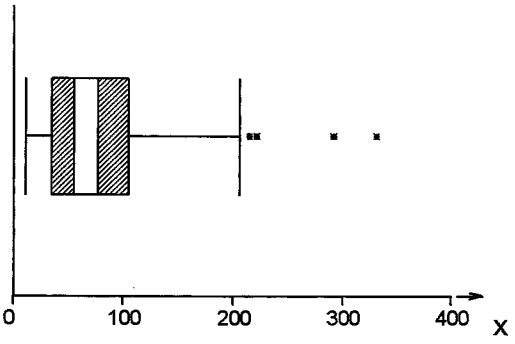

Fig. 3 Box-and-whisker plot of pregnenolone data.

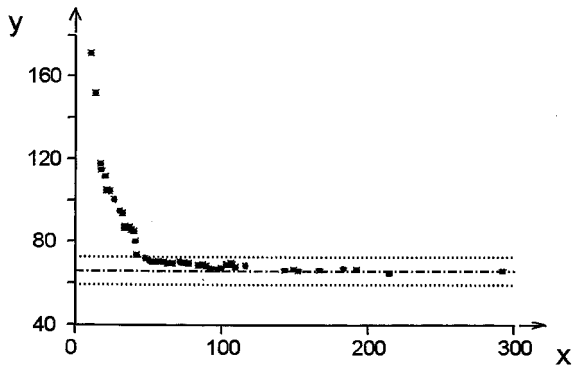

Fig. 4 Midsum plot of pregnenolone data.

The highest value of the correlation coefficient of the *Q-Q* rankit plot r=0.9951 is reached for exponential distribution.

The midsum Z_Q and positive skewness S_Q indicate a skewed distribution and the presence of outliers. The point estimate of skewness of 1.57 and kurtosis at 6.06 indicate that the sample distribution is strongly asymmetric with a slim and sharp peak and is definitely not normal.

4. Sample assumptions, (cf. pp. 78–82 in ref. 4 or 16): applying an analysis of basic assumptions about data the following conclusions were arrived at:

(a) *Examination for normality of sample distribution:* a combined sample skewness and kurtosis test leads to

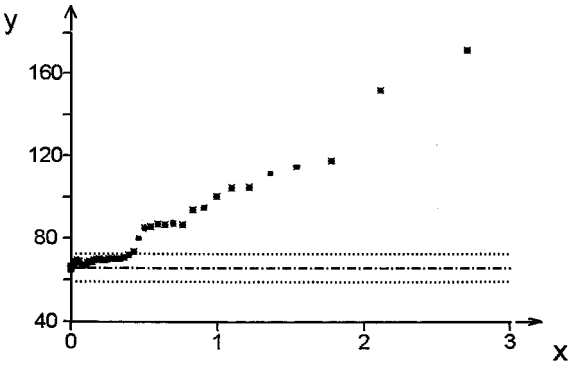

Fig. 5 Symmetry plot of pregnenolone data.

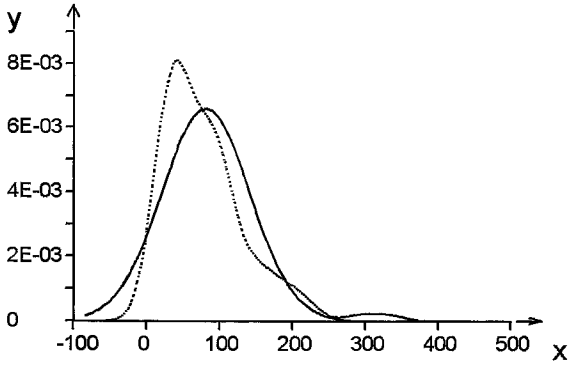

Fig. 6 The kernel estimator of the probability density plot of pregnenolone data, showing the empirical curve (dot curve) and the normal distribution approximation (full curve).

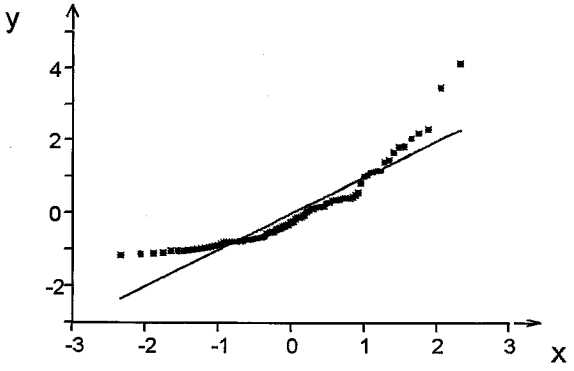

Fig. 7 Quantile-quantile plot (for normal distribution called a rankit plot) of pregnenolone data.

Tab. 4 The quantile measures of location, spread and shape for pregnenolone concentration (nmol/l).

Quantile	P	Lower quantile Q_U	Upper quantile Q_L	Range R_Q	Midsum Z_Q	Skewness S_Q	Tails length T_Q	Pseudo-Sigma G_Q
Median	0.5	66.0	66.0	–				
Quartile	0.25	35.2	104.3	69.1	69.7	0.21	0.000	51.2
Octile	0.125	24.3	149.6	125.3	87.0	0.09	0.595	54.5
Sedecile	0.0625	18.1	189.5	171.4	103.8	0.06	0.909	56.0

Tab. 5 Point and interval estimates of location pregnenolone concentration (nmol/l).

Parameter	Estimate	Estimate	95% Confidence interval	
	$\hat{\mu}$	$\hat{\sigma}$	L_D	L_H
Mean $\bar{x}$	80.6	60.5	68.6	92.6
Median $\tilde{x}_{0.5}$	66.0	61.3	48.8	83.1
Midsum $\hat{x}_P$	171.4	*	*	*
Mode $\hat{x}_M$	33.9	*	*	*
M-estimate of mean value μ_M	71.1	49.4	60.7	81.4
Re-transformed mean (Power Transf.) $\bar{x}_R$	64.2	48.2	55.2	74.4
Re-transformed mean (Box-Cox Transf.) $\bar{x}_R$	65.8	*	30.6	125.4

* indicates values that software does not estimate

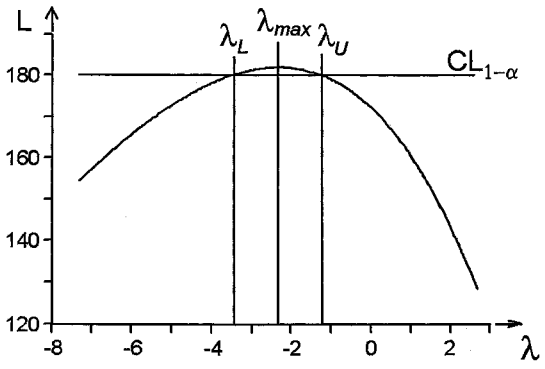

Fig. 8 Plot of the logarithm of maximum likelihood L in dependence on the power λ for pregnenolone data, and estimation of the optimal power λ_{max} with its lower λ_L and upper λ_U limits of the confidence interval, for the confidence level $(1-\alpha)$, $CL_{1-\alpha}$.

the test statistic C_1=90.73 > χ^2(0.95, 2)=5.992 and therefore normality in the data distribution is rejected.

(b) *Examination of sample homogeneity:* there are two observations outside the interval of both of Hoaglin's inner bounds [B_L^*=–119.0 nmol/l; B_U^*=258.4 nmol/l] and therefore two points $x_{(13)}$=332.0 nmol/l and $x_{(22)}$=292.0 nmol/l may be classed as outliers and sample homogeneity rejected. The classic measures of location, scale and distribution shape for data without two outliers are $\bar{x}$= 75.9 nmol/l, $s(\bar{x})$=51.0 nmol/l, $\hat{g}_1(x)$=0.99 and $\hat{g}_2(x)$=3.41.

5. Data transformations: Figure 6 shows the asymmetric distribution of the original sample data, and therefore the need for data transformation. In the case of the Box-Cox transformation the true mean value of the sample distribution and both asymmetric confidence limits L_L and L_U were calculated. From the plot of the logarithm of the likelihood function for the power transformation, the maximum on the curve was read from a graph at $\hat{\lambda}$ = 0.13, for the Box-Cox transformation the maximum of the curve is at $\hat{\lambda}$ = –2.32. For both transformations, the corresponding 95% confidence interval does not contain the exponent value λ=1, so both transformations are statistically significant.

The measures of location, spread and shape for the original data, *i.e.* mean $\bar{x}$=80.6 nmol/l, standard deviation $s(x)$=60.5 nmol/l, skewness $\hat{g}_1(x)$=1.57 and kurtosis $\hat{g}_2(x)$=6.06 are outside statistical significance and may be taken as false estimates. Power transformation estimated the corrected mean value $\bar{x}_R$=64.2 nmol/l and the Box-Cox transformation the corrected mean value $\bar{x}_R$=65.8 nmol/l.

6. Classical and robust statistics (n=100): a survey of various point and interval estimates of location is given in Table 5. It may be concluded that the assumption of normality is not fulfilled because of a skewed sample distribution shape, and that therefore the mean $\bar{x}$ and midsum $\hat{x}_P$ cannot be used. The use of the robust estimate median is equivalent to excluding outliers from the sample. Both transformations lead to the statistically same value $\bar{x}_R$=65 nmol/l. This value is quite close to the median $\tilde{x}_{0.5}$=66 nmol/l and the trimmed mean $\bar{x}$(45%)=66.2 nmol/l. The M-estimate μ_M=71 nmol/l is not far from the re-transformed means and median. The confidence interval of the power transformation estimate [L_L=55 nmol/l, L_U=74 nmol/l] is narrower than that of the Box-Cox transformation [L_L=31 nmol/l, L_U=125 nmol/l], and the power transformation estimate can thus be taken as more efficient.

Case study 2. The mean value of a haptoglobin concentration in the human serum

Concentration of haptoglobin [g l^{-1}] were measured in human serum taken from eight adults yielding values of 1.82, 3.32, 1.07, 1.27, 0.49, 6.79, 0.15, 1.98. Applying the Horn procedure the measure of location and spread for this small sample is calculated.

Calculations: the Horn procedure consists of six steps:

1. Order statistics for $n=8$, [g l^{-1}]: $x_{(i)}$=0.15, 0.49, 1.07, 1.27, 1.82, 1.98, 3.32, 6.79.
2. The pivot depth reachs the value int(2.75) ≈ 2.
3. The lower pivot is $x_D=x_{(H)}=x_{(2)}$=0.49 and the upper pivot $x_H=x_{(n+1-H)}=x_{(7)}$=3.32.
4. The pivot halfsum [15] is $P_L=(x_D+x_H)/2$=1.905 g l^{-1}.
5. The pivot range [16] is $R_L=x_H-x_D$=3.32–0.49=2.83 g l^{-1}.
6. The 95% confidence interval of the measure of location μ is calculated for the Horn quantile $t_{L,\ 1-\alpha/2}$=0.564 with the use of relation [16] and leads to the range 0.31 g l^{-1} ≤ μ ≤ 3.50 g l^{-1}. Thus, for the small sample the pivot technique is more suitable.

5. Conclusion

Biochemical data are often less than ideal, and do not fulfill all basic assumptions. Interactive data treatment by modules of the proposed algorithm in S-Plus enables the following steps:

(i) Exploratory data analysis identifies the basic statistical features and patterns of the data.

(ii) Sample distribution is examined by the symmetry and tail lengths, skewness and kurtosis.

(iii) Sample assumptions about the data concern the independence of sample elements, normality and sample homogeneity.

(iv) Data transformations enable power transformation and Box-Cox transformation, improving sample symmetry and also stabilizing the spread.

(v) Classical and robust statistics are calculated for large (n>30) and medium samples (15<n<30), while for small samples ($4 \le n \le 20$) the Horn procedure of pivot measures is recommended.

6. Acknowledgements

The financial support of the Grant Agency of the Czech Republic (Grant No 303/00/1559) is gratefully acknowledged.

References

1. Tukey JW. Exploratory data analysis. Reading, Massachusetts: Addison Wesley, 1977.
2. Chambers J, Cleveland W, Kleiner W, Tukey P. Graphical methods for data analysis. Boston: Duxbury Press, 1983.
3. Hoaglin DC, Mosteler F, Tukey JW. Exploring data tables, trends and shapes. New York: Wiley, 1985.
4. Meloun M, Militký J, Forina M. Chemometrics for analytical chemistry, Vol 1, PC-Aided statistical data analysis. Chichester: Ellis Horwood, 1992.
5. Shapiro SS, Gross AJ. Statistical modeling techniques. New York: Marcel Dekker Inc, 1981.
6. Silverman BW. Density estimation. London: Chapman and Hall, 1986.
7. Lejenne M, Dodge Y, Koelin E. Proceedings of the Conference COMSTAT'82. Toulouse 1982: 173p, Vol III.
8. Hoaglin DC, Mosteler F, Tukey JW, editors. Understanding robust and exploratory data analysis. New York: Wiley, 1983.
9. Kafander K, Spiegelman CH. An alternative to ordinary QQ plot. Comput Stat Data Anal 1986; 4:167.
10. Hines WGS, Hines RJH. Quick graphical power law transformation selection. Am Statist 1987; 41:21.
11. Hoaglin DC. Performance of some resistant rules for outlier labeling. J Am Statist Assoc 1986; 81:991.
12. Stoodley K. Applied and computational statistics, Chichester: Ellis Horwood, 1984.
13. Horn PS, Pesce AJ, Copeland BE. A robust approach to reference interval estimation and evaluation. Clin Chem 1998; 44:622–31.
14. Blackman NM, Machol RE. IEEE Trans on inform theory 1987; IT–33:373.
15. Horn J. Some easy T-statistics. J Am Statist Assoc 1983; 78:930.
16. Patil GP, Ord K, (Patil GP, editor). Statistical distributions in scientific work. Vol. II, Holland: D Reidel, 1975.
17. Meloun M, Hill M, Militký J, Kupka K. Transformation in the PC-aided biochemical data analysis. Clin Chem Lab Med 2000; 38(6):553–59.
18. Statistical package ADSTAT, Pardubice: TriloByte Statistical Software, 1999. http://www.trilobyte.cz, E-mail: kupka@trilobyte.cz.
19. S-Plus. Seattle, Washington: MathSoft, Inc, 1997.
20. Hirato K, Yanaihara T. Serum steroid hormone levels in neonates born from the mother with placental sulfatase deficiency. Endocrinol Jpn 1990; 37:731–9.
21. Rabe T, Hosch R, Runnebaum B. Sulfatase deficiency in the human placenta: clinical findings. Biol Res Pregnancy Perinatol 1983; 4:95–102.
22. Shapiro LJ, Cousins L, Fluharty AL, Stevens RL, Kihara H. Steroid sulfatase deficiency. Pediatr Res 1977; 11:894–7.
23. Lykkesfeldt G, Nielsen MD, Lykkesfeldt AE. Placental steroid sulfatase deficiency: biochemical diagnosis and clinical review. Obstet Gynecol 1984; 64:49–54.
24. NCSS Statistical Software, 329 North 1000 East, Kaysville, Utah 84037. Email: sales@ncss.com.

Received 17 July 2000, revised 16 November 2000, accepted 20 November 2000

Corresponding author: Prof. Dr. Milan Meloun, Department of Analytical Chemistry, University of Pardubice,
532 10 Pardubice, Czech Republic
Tel: +4240-603 7026, Fax: +4240-603 7068
E-mail: milan.meloun@upce.cz, http://meloun.upce.cz