Clin Chem Lab Med 2001; 39(3):244–255 © 2001 by Walter de Gruyter · Berlin · New York

Exploratory Biochemical Data Analysis: a Comparison of Two Sample Means and Diagnostic Displays

Milan Meloun[1], Martin Hill[2] and David Cibula[3]

[1]Department of Analytical Chemistry, Faculty of Chemical Technology, Pardubice University, Pardubice, Czech Republic
[2]Institute of Endocrinology, Prague, Czech Republic
[3]Department of Obstetrics and Gynecology, General Teaching Hospital in Prague, Charles University Institute of Endocrinology, Prague, Czech Republic

The occurrence of acne in women with hyperandrogenemia is well known; a question remains, however, as to whether a further positive relationship can be detected between the intensity of acne and the levels of testosterone, androgen precursors and sex hormone binding globulin (SHBG). A procedure of interactive data analysis extracting relevant information from original data was applied. Exploratory data analysis (EDA) identifies basic statistical features and patterns of data using a variety of diagnostic displays. The need for this step is particularly acute in biochemical and clinical data, the distribution of which is mostly non-Gaussian and often corrupted by the outliers. The omission of EDA can lead to incorrect results and false conclusions. In the EDA (i) several graphical tools for summarizing data are applied, (ii) the peculiarities of a sample distribution are investigated, (iii) a construction of distribution is carried out, (iv) a graphical comparison of the sample distribution with selected theoretical distributions is employed. The proposed procedure is illustrated by typical case study in the evaluation of differences between mean values of serum levels of testosterone, androgen precursors and SHBG in a group of patients with mild and severe forms of acne. A knowledge of the interval estimate of the mean value in both groups enables their comparison at the chosen probability level. As will be apparent from the evaluation of inter-group SHBG differences, an incorrect approach to the determination of group mean values could result in a complete misinterpretation of the data. The results indicate that androgens are not significantly related to the intensity of acne, and that SHBG is higher in patients with more severe forms of acne.

Key words: Two sample means; Statistics; Exploratory data analysis; Chemometrics; Biometrics.

Abbreviations: ADION, androstenedione; CDA, confirmatory data analysis; DHEA, dehydroepiandrosterone; DHEAS, dehydroepiandrosterone sulfate; EDA, exploratory data analysis; SHGB, sex hormone binding globulin.

1. Introduction

Statistics, when correctly used, can be a useful and constructive tool in the analysis of biochemical and clinical data; in careless or unscrupulous hands, however, it can be a dangerous weapon. Used properly, statistics will allow an investigator to quantify concepts and conclusions, and help both to take into account sources of systematic variation and to minimize the effect of random error. It will draw attention to the accuracy of the data, and to the type and quality of the inferences. This paper will introduce the most important of the basic techniques used in interactive computerized statistics; at the same time the limitations of these diagnostic methods and the conditions under which they are valid will also be discussed.

The classical approach to statistical data analysis, based on the use of the arithmetic mean and sample variance, is possible when some stringent assumptions are valid. This approach is based on assumptions about the statistical nature of a sample, such as independence, normality and homogeneity. However, experience and further research have forced the recognition that classical techniques can perform badly when the biochemical and clinical data depart from the ideal described by such assumptions. More recently developed robust exploratory methods are broadening the effectiveness of statistical analyses. Techniques of exploratory data analysis help researchers cope with sets of data in a fairly informal manner, guiding them towards structure relatively quickly and easily. Good statistics practitioners have always looked in detail at biochemical or clinical data before producing summary statistics and testing statistical hypotheses. One description of the general steps and operations that make up practical data analysis identifies two broad phases: the exploratory and the confirmatory.

The first phase is an exploratory data analysis (EDA). According to Tukey (1), EDA represents "detective work", in which data are treated to uncover typical relationships and patterns. EDA uses various descriptive and graphic techniques which are typically free of strict statistical assumptions about data (2, 3); often called a "distribution-free technique", EDA provides the first contact with the data, and isolates certain basic statistical features and patterns within it. An important element of the exploratory approach is the diagnostic flexibility, both in tailoring the analysis to the structure of the data and in responding to the patterns that the successive steps of the analysis uncover.

The second phase is confirmatory data analysis (CDA), which assesses the reproducibility of the observed patterns or effects; its role is closer to that of traditional statistical inference in providing statements of significance and confidence (4).

The role of androgens in the pathogenesis of acne vulgaris is well known; elevated androgen levels in women with acne have been confirmed in many studies (5–13). Androgens elevate *sebum* production and *follicular ceratosis*, which plays a major role in the etiology of acne (14). However, the question remains as to whether the severity of acne is directly related to androgenicity, or whether acne severity is influenced more by other factors. Discrepant results have been indicated in the literature regarding the relationship between androgenicity and acne severity. For this reason, a prospective study was conducted to evaluate the relation of acne severity to clinical and laboratory markers of androgenicity in a group of women with acne. Evaluation of the differences between the levels of steroids in two groups with different acne severity was one of the constitutive steps in the examination of the relationship. Differences were evaluated in dehydroepiandrosterone sulfate (DHEAS), dehydroepiandrosterone (DHEA), androstenedione (ADION), testosterone and sex hormone binding globulin (SHBG).

This paper employs efficient diagnostic displays and plots of an EDA, the new words frequently appearing almost side by side, to make connections between the EDA and an existing background of statistical knowledge in the analysis of biochemical and clinical data. A typical study case illustrates an application of the proposed procedure for the evaluation of differences in mean serum levels of testosterone, androgen precursors and SHBG in a group of patients with mild and severe forms of acne. The best estimates of the mean values of the steroid levels, with their confidence intervals, were calculated in groups of patients with mild (denoted as data set Acne 0) and severe (denoted as data set Acne 1) forms of acne. Knowledge of the interval estimate of the mean value in both groups enables their mutual comparison at a chosen probability level; such comparisons contribute to answering the question as to whether testosterone, androgen precursors and SHBG can influence the severity of acne. As will be apparent from the evaluation of the inter-group differences of SHBG, an incorrect approach to the determination of group mean values could result in a complete misinterpretation of the data.

2. Theory

The most common biochemical data structure is a *batch of numbers*. This simple data structure may have characteristics not easily discerned by scanning or studying the numbers. It is necessary to display the batch as a whole and to notice such features as:

- how close to symmetrical it is;
- how spread out the numbers are;
- whether there are a few values that are far removed from the rest;
- whether there are concentrations of data.

In exposing these features of data to the analyst, some EDA terminology and diagnostics are offered here at the outset. One of the most frequent acts in statistical data analysis is the *one-sample problem* based on a batch of numbers statistically denoted as *sample* $x_1, ..., x_n$, this representing behavior of a *univariate* (random) *variable* x. Observations are the *random quantities*. The complete collection of all possible outcomes from the experiment in question is called the *population*, and observations represent points in this population (4). The sample values can be sorted in order of ascending magnitude, $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ and the sorted values $x_{(1)}, x_{(2)}, ..., x_{(n)}$ are known as the *order statistics*. The order statistic $x_{(i)}$ is a rough estimate of sample quantile $\tilde{x}_{Pi}$. Under quantile $\tilde{x}_{Pi}$ the $100\,P_i$ % of the sample values lie, and the parameter P_i is the *cumulative* or *rank probability* given by $P_i=i/(n+1)$. For a normal distribution the expression $P_i=(i-3/8)/(n+1/4)$ is often used, but the EDA algorithms also use the empirical expression $P_i=(i-1/3)/(n+1/3)$. The plot of order statistic $x_{(i)}$ against cumulative probability P_i, for $i = 1, ..., n$ is an estimate of the *quantile function* $Q(P)$. This is, in fact, an inverse function of the sample distribution function. For any value α from the interval [0, 1] the 100α-th *quantile* $\tilde{x}_\alpha$ may be calculated by linear interpolation

$$\tilde{x}_\alpha=(n+1)\left(\alpha-\frac{i}{n+1}\right)(x_{(i+1)}-x_{(i)})+x_{(i)},$$

$$\text{where } \frac{i}{n+1}\le\alpha\le\frac{i+1}{n+1}$$

Some methods of EDA are based on the selected *quantiles* Q being calculated for selected cumulative probabilities $P_i = 2^{-i}$, $i = 1, 2, ...$. These quantiles are also termed the *letter values* (16), as in Table 1, where the symbol u_{Pi} denotes the quantile of the standard normal distribution $N(0, 1)$. Excepting a *median* ($i = 1$), for each $i > 1$ there is a pair of extreme quantiles, the lower Q_L and upper Q_U letter value. The lower letter value is calculated for a cumulative probability $P_i = 2^{-i}$ while for the upper $P_i = 1 - 2^{-i}$ is used (Figure 1).

2.1 Basic diagnostic EDA displays

The features and statistical properties of a data sample are described by the symmetry, peakedness and tail length of the sample distribution, the local concentration of data and a presence of outliers. The various exploratory plots display such information.

A *quantile plot* (2) enables the identification of the peculiarities of shape in a sample distribution, which may be symmetric, or skewed to higher or lower values. It is a plot of the quantiles of an actual distribution against the corresponding probability P_i. The scale of the vertical axis of the quantile is the scale of the ordered variable $x_{(i)}$. To compare an actual sample distribution with a normal one, the quantile function of the normal distribution $Q(u_{Pi}) = \mu + \sigma\, u_{Pi}$ for $0 \le P_i \le 1$ is plotted:

1. The classic estimates of μ and σ^2, *i. e.* $\hat{\mu}=\bar{x}$ and $\hat{\sigma}^2=s^2$ are used where $\bar{x}$ is the sample arithmetic mean and s^2 is the estimate of sample variance σ^2.

Tab. 1 A survey of some letter values.

i	i-th quantile	Probability for lower quantile	Symbol for letter value	Normal quantile u_{Pi}
1	Median	$2^{-1} = 0.500$	M	0
2	Quartiles	$2^{-2} = 0.250$	F	–0.674
3	Octiles	$2^{-3} = 0.125$	E	–1.15
4	Sedeciles	$2^{-4} = 0.0625$	D	–1.53

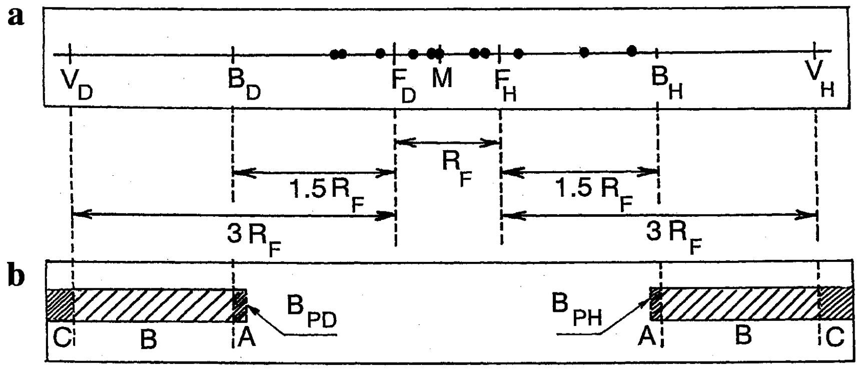

Fig. 1 Construction of the dot diagram with letter values indicating outliers: (a, upper part) the dot diagram with median M, F_L lower and F_U upper quartiles, inner B_L lower and B_U upper limits, outer V_L lower and V_U upper limits; (b, lower part) the area of outliers: *A* close outliers, *B* near far outliers, *C* far outliers.

2. Robust estimates of μ and σ^2, $\hat{\mu}=\tilde{x}_{0.5}$ and $\hat{\sigma}^2=(R_F/1.349)^2$ are used where $\tilde{x}_{0.5}$ is the median and R_F is the interquartile range, $R_F = F_U - F_L$.

A *dot diagram* (1), being a one-dimensional scatter plot of data, represents an univariate projection of the quantile plot onto the *x*-axis. The dot diagram simply shows the local concentration of data, outliers, and extremes in the data. The data points are plotted along a straight line parallel to, and above, the horizontal axis. Stacking or jitter can be applied to better display overlapping points.

A *stacking dot diagram* has data points of equal value plotted above each other on a vertical line orthogonal to the main line of the plot.

A *jitter dot diagram* (2), being similar to a dot diagram, also represents an univariate projection of the quantile plot, but sample points are randomly displayed on a strip above the axis. The width of the strip is kept small compared to the range of the horizontal axis. The vertical position of a point within the strip is random. Although a scatter plot with jitter loses accuracy, its information content is enhanced.

A *box-and-whisker plot* (1) provides a graphical display of a five-letter values summary in the form of the median, two quartiles and two extremes. The bulk of data is represented as a rectangle with the lower and upper quartiles being the bottom and the top of the rectangle respectively, and the median is portrayed by a vertical line within the rectangle. The box-and-whisker plot has a length R_F from the lower F_L to upper F_U quartile, $R_F = F_U - F_L$, and its width is proportional to the value $\sqrt{n}$ even where this has no meaning. The plot is useful in illustrating the skewness of a sample. If the distribution has a long heavy tail to the right (*positive skewness*) then the right-hand section of the box will be longer than the left, and the upper extreme point will be further from the median than the lower extreme. The converse will be true if the distribution has *negative skewness* with its longer tail to the left. Two lines, called whiskers, extend from the ends of the box to the *adjacent values* B_U and B_L. Adjacent values lie within the *inner bounds* nearest to their boundary values, B_U and B_L, expressed by $B_U = F_U+1.5\ R_F$ and $B_L = F_L-1.5\ R_F$. Values outside the inner bounds but within the *outer bounds* V_U and V_L, expressed by $V_U = F_U+3\ R_F$ and $V_L = F_L - 3\ R_F$, are called the *near far outliers* (Figure 1). Data points outside the inner bounds, smaller than V_L or larger than V_U, are called the *far outliers*, and are marked on this plot by a circle. This plot enables (i) the determination of a robust estimate of the median, (ii) the illustration of the spread and skewness of the sample, (iii) examination of the symmetry and length of distribution tails, and (iv) identification of the outliers. Displaying several box-plots side by side gives a graphical comparison of the corresponding distributions. To emphasize the relative locations, box plots can be drawn with notches in their sides, and are then called *notched box-and-whisker plots* (2). Such plots enable an examination of the variability of the median, which is expressed by notches given by the robust confidence interval $I_L \le M \le I_U$, where the lower and upper limits are

$$I_L = M - 1.57\, R_F/\sqrt{n} \text{ and } I_U = M + 1.57\, R_F/\sqrt{n} \quad [1a,1b].$$

The notches I_L and I_U are located symmetrically around the median and form the beginning and the end of the white strip in the box.

2.2 EDA Diagnostics for distribution shape examination

The main statistical features of sample distribution are represented by the asymmetry and tails length in comparison with a normal (Gaussian) distribution. Asymmetry (skewness) and peakedness (kurtosis) can be characterized at different distances from the median by the following statistical diagnostics based on quantiles: the *halfsum* (or *midsum*) $Z_Q = (Q_L + Q_U)/2$, the *interquantile range* $R_Q = Q_U - Q_L$, the *skewness* $S_Q = (M - Z_Q)/R_Q$, the *pseudosigma* $G_Q = R_Q/(-2u_{Pi})$, the *tails length* $T_Q = \ln(R_Q/R_F)$, where Q stands for the letter value and u_{Pi} is the quantile of standardized normal distribution for $P_i = 2^{-i}$. The resulting diagnostics are summarized in Table 2.

For a selected symmetric distribution the theoretical values of the length of tails, T_E and T_D, were computed: for a normal distribution $T_E = 0.534$ and $T_D = 0.822$, for a rectangular distribution $T_E = 0.405$ and $T_D = 0.559$, and for a Laplace (both sides exponential) distribution $T_E = 0.693$ and $T_D = 1.098$. The skewness S_Q has negative values for distributions skewed to higher values and positive values for distributions skewed to lower values. For distributions with longer tails than the normal, the values of pseudosigma G_Q increase with the distance from the median. When the values of pseudosigma G_Q decrease with the distance from the median, the sample distribution has shorter tails than the normal. To examine all statistical features of the sample various plots of characteristics from Table 2 are used. For large samples letter values are examined, while for small samples the quantile $\tilde{x}_{Pi} = x_{(i)}$ usually for $P_i = (i - 1/3)/(n + 1/3)$ is used.

The *halfsum (or midsum) plot* (2) indicates the symmetry of distribution. It has on the *x*-axis the order statistic $x_{(i)}$ and on the *y*-axis the halfsum (or midsum) $Z_i = (x_{(n+1-i)} + x_{(i)})/2$. For symmetric distribution the halfsum (or midsum) plot forms a horizontal line $y = \tilde{x}_{0.5}$. If all points are situated in between the two dashed lines of the confidence bounds of the horizontal line, then the distribution is symmetric.

A *symmetry plot* (2) indicates the symmetry of distribution. It has on the *x*-axis the quantile u_{Pi}^2 for $P_i = i/(n+1)$ and on the *y*-axis the halfsum (or midsum) $Z_i = (x_{(n+1-i)} + x_{(i)})/2$. For a symmetric distribution the symmetry plot forms a horizontal line $y = \tilde{x}_{0.5}$. When this line has a non-zero slope, the slope gives an estimate proportional to skewness. When the data distribution is asymmetric, the plot shows a clear trend (increasing for a negative skewness and decreasing for a positive skewness), going far beyond the dashed lines.

The *quantile-box plot* (17): examining the statistical features of data is based on an estimate of the sample quantile function, formed connecting points $\{x_{(i)}, P_i\}$ with straight lines. On the *x*-axis it has probability P_i and on the *y*-axis the order statistic $x_{(i)}$, where P_i is calculated by $P_i = (i - 1/3)/(n + 1/3)$. For symmetric distributions, the sample quantile function exhibits a sigmoid shape, whereas for asymmetric the quantile function is convex or concave increasing. The following quantile boxes are on the graph:

a) The *quartile box F* has on the *y*-axis two vertices given by quartiles F_L and F_U with corresponding values on the *x*-axis equal to the cumulative probability values $P_2 = 2^{-2} = 0.25$ and $1 - 2^{-2} = 0.75$.

b) The *octiles box E* has on the *y*-axis octiles E_L and E_U and on the *x*-axis the cumulative probabilities $P_3 = 2^{-3} = 0.125$ and $1 - 2^{-3} = 0.875$.

c) The *sedeciles box D* has on the *y*-axis sedeciles D_L and D_U and on the *x*-axis the cumulative probabilities $P_4 = 2^{-4} = 0.0625$ and $1 - 2^{-4} = 0.9375$.

The position of the *median M* is marked by a horizontal line inside the quartile box. A robust estimate of the median confidence interval, $M \pm 1.57\, R_F/\sqrt{n}$, is drawn as a vertical line at $P_1 = 0.5$. On the basis of this plot, the following statistical features of the sample distribution may be stated (17): i) *a symmetric unimodal sample distribution* contains individual boxes arranged symmetricly inside themselves, and the value of relative skewness is close to zero, $S_Q \approx 0$. ii) *An asymmetric sample distribution*: in the case of a distribution skewed to higher values, there are significantly shorter distances between the lower parts of the boxes when compared with those between the upper ones. The skewness S_Q then has a negative value. For a distribution skewed to lower values the skewness S_Q is positive. iii) *Outliers* are indicated by a sudden increase of the quantile function outside the *F* box, and the slope may approach infinity. iv) *A multimodal sample distribution* is indicated by several parts of the quantile function inside box *F* reaching a zero slope.

Tab. 2 Diagnostics describing the distribution shape.

Diagnostic	Used for exploration of	Valid for *L*
Halfsum (or midsum) Z_Q	symmetry (at $Z_Q = M$ where M is median)	*F, E, D, ...*
Interquantile range R_Q	spread	*F, E, D, ...*
Skewness S_Q	symmetry (at $S_Q = 0$)	*F, E, D, ...*
Pseudosigma G_Q	peakedness (for Gaussian distribution $G_Q = 0$)	*F, E, D, ...*
Tails length T_Q	peakedness	*E, D, ...*

2.3 Determination of sample distribution

Some graphical displays can show overall patterns or trends; they can also reveal surprising, unexpected, or amusing features of the data that might otherwise go unnoticed. When a large number of observations is available, the estimation of probability density function or other function characterizing the data distribution can help to elucidate the statistical behaviour of the sample.

A *histogram* is one of the oldest classic presentations of grouped frequency distributions. The range of the continuous variable is partitioned into several intervals, usually of equal length, and the counts (frequencies) of the observations in each interval are plotted as a bar length. Histograms give a visual information about asymmetry, kurtosis and outliers. When the histogram is plotted on a square root vertical scale, which is an approximate variance stabilizing transformation, it is called a *rootogram.*

A *kernel density estimator for the sample probability density function* $\hat{f}(x)$ for small and medium-sized samples may be calculated by the relationship $\hat{f}(x)=[1/(nh)]\sum_{i=1}^{n} K[(x-x_i)/h]$, where h is bandwidth, which controls a smoothness of $\hat{f}(x)$, and $K(x)$ is the kernel density function. The kernel density estimator $K(x)$ is symmetric around zero, is a nonnegative density function, and has properties of a probability density function. To take a bi-quadratic kernel estimate

$$\mathrm{K(x)}=\begin{Bmatrix} 0.9375(1-x^2)^2 & \text{for } -1 \le x \le 1 \\ 0 & \text{for } x \text{ outside}[-1;1] \end{Bmatrix} \qquad [2]$$

The quality of the kernel estimate $\hat{f}(x)$ is controlled by the choice of parameter h. If h is too small, the estimate is rough; if it is too large, the shape $\hat{f}(x)$ is flattened too much. Selection of optimal h for EDA purposes is described by Lejenne *et al.* (16). The plot brings a comparison of the normal probability density curve (solid curve) with a kernel density estimate, computed from the data (dashed curve): when the data are not homogeneous and show a clustering tendency, several local maxima of the density estimate can occur. For normal data, both curves should be close to each other. On the other hand, it must be realized that with a small enough smoothing parameter h, local maxima can occur for any data.

A *quantile-quantile Q-Q plot* (3, 16) allows a comparison of sample distribution described by the empirical $Q_E(P_i)$ quantile function and the given theoretical quantile function $Q_T(P_i)$. The values of the empirical $Q_E(P_i)$ function are approximated by the sample order statistic $x_{(i)}$. If data points lie along a straight line, then there is close agreement between the sample and theoretical distributions $x_{(i)} \approx Q_T(P_i)$ where P_i is the cumulative probability $P_i = (i - 1/3)/(n + 1/3)$. Use of this equation requires a knowledge of all of the parameters for theoretical quantile functions $Q_T(P_i)$. Theoretical distributions can be standardized to the form $Q_T(P_i) = \mu + \sigma Q_{TS}(P_i)$. Here μ is usually the location parameter, s is the spread parameter, and $Q_{TS}(P_i)$ is the standardized quantile function. For most bi-parametric distributions $Q_{TS}(P_i)$ is free of adjustable parameters. For some tri-parametric distributions the shape factor is usually a parameter of the plot. The standardized quantile function $Q_T(P_i)$ is therefore used for practical construction of the quantile-quantile plot. The x and y co-ordinates of the Q–Q plot for selected theoretical distributions are given in Table 3.

Due to a strong dependence among order statistics x_i and their non-constant variance, the quantile-quantile plot for small samples has a very patterned appearance. When the normal distribution is used, this plot is called the *rankit plot* (or *normal-probability plot*). This plot is very effective for testing an assumption of normality on a variable, and enables classification of the sample distribution according to its skewness, kurtosis and tails' length. A convex or concave shape indicates a skewed sample distribution. It is the best graphical tool for checking normality and outliers presence; for normal data without outliers the points should fit closely to a line; for normal data with outliers, points in the central parts should fit closely to a line and the endpoints fur-

Tab. 3 Standardized probability density $f_T(s)$ and distribution $F_T(s)$ functions, and corresponding coordinates (x, y) of the Q–Q plot.

Distribution	$F_T(s)$	$f_T(s)$	y	x
Rectangular	s	1	$x_{(i)}$	P_i
Exponential	$1 - \exp(-s)$	$\exp(-s)$	$x_{(i)}$	$-\ln(1 - P_i)$
Normal	$\Phi(s)$	$(2\pi)^{-1/2} \exp(-0.5\, s^2)$	$x_{(i)}$	$\Phi^{-1}(P_i)$
Laplace, $x \le 0$	$0.5 \exp(s)$	$0.5 \exp(s)$	$x_{(i)}$	$\ln(2P_i)$ for $P_i \le 0.5$
Laplace, $x > 0$	$0.5(2 - \exp(-s))$	$0.5 \exp(-s)$	$x_{(i)}$	$-\ln(2(1 - P_i))$ for $P_i > 0.5$
Log-normal	$\Phi[\ln(s)]$	$(2\pi)^{-1/2} \exp(-0.5 \ln s^2)$	$x_{(i)}$	$\exp[\Phi^{-1}(P_i)]$

In Table 3 the normal distribution function $\Phi(s)$ is defined as $\Phi(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} \exp(-0.5\, u^2)\mathrm{d}u$.

For calculation of an inverse function $\Phi^{-1}(P_i)$ the simple approximate relation may be used

$$\Phi^{-1}(\mathrm{P_i}) = \frac{-9.4 \ln\left(\frac{1}{P_i} - 1\right)}{\mathrm{abs}\left(\ln\left(\frac{1}{\mathrm{P_i}} - 1\right)\right) + 14}.$$

ther away from the line; for data coming from a positively skewed distribution (*e. g.* log-normal, exponential) the shape should be nonlinear, convex; for data coming from a negatively skewed distribution the shape should be nonlinear, concave; for data coming from a distribution with kurtosis higher than normal, *i. e.* those showing a high concentration around the mean (*e. g.* Laplace), the shape should be concave-convex; for data coming from a distribution with kurtosis smaller than normal, *i. e.* those with low concentration around the mean (*e. g.* uniform), the shape should be convex-concave. One advantage of the rankit plot, as compared to statistics describing skewness, kurtosis, etc., is that one can visually check whether the lack of normal appearance (non-linearity) is caused by just a few points, or whether it is a general tendency shared by all data.

2.4 Comparison of two sample means

To consider a random sample of size n_1, with mean $\bar{x}$ and variance s_x^2, and sample of size n_2 with mean $\bar{y}$ and variance s_y^2: the hypothesis H_0: $\bar{x} = \bar{y}$ is tested against the alternative H_A: $\bar{x} \neq \bar{y}$. When both samples are not from a normal distribution, the modified test criterion T_3 is used

$$T_3 = \frac{\left|\bar{x}-\bar{y}\right| + C + D\,(\bar{x}-\bar{y})^2}{\sqrt{\frac{s_x^2}{n_1}+\frac{s_y^2}{n_2}}}, \qquad [3].$$

where

$$C = \frac{1}{6}\,\frac{\frac{\bar{g}_x}{n_1^2}\,\frac{s_x^3}{\sqrt{n_1}} - \frac{\bar{g}_y}{n_2^2}\,\frac{s_y^3}{\sqrt{n_2}}}{\frac{s_x^2}{n_1}+\frac{s_y^2}{n_2}} \text{ and } D = \frac{1}{3}\,\frac{\frac{\bar{g}_x}{n_1^2}\,\frac{s_x^3}{\sqrt{n_1}} - \frac{\bar{g}_y}{n_2^2}\,\frac{s_y^3}{\sqrt{n_2}}}{\left[\frac{s_x^2}{n_1}+\frac{s_y^2}{n_2}\right]^2}.$$

The test criterion T_3 for H_0 has a Student *t*-distribution with $v = n_1 + n_2 - 2$ degrees of freedom. This statistic is robust for skewed sample distributions, also for heteroscedasticity in data and different sample variances $\sigma^2_x \neq \sigma^2_y$.

3. Procedure

3.1 Procedure of univariate data analysis

The main task of statistical analysis is to collect information about a population, so sample estimates are used to find confidence intervals of parameter of location. With a given probability, the confidence interval of a population parameter will include the true value of this „unknown" parameter. Statistical hypotheses testing "unknown" parameters of the population may also be carried out.

Step 1: Exploratory data analysis

When no preliminary information about data is available, a full exploratory data analysis is applied: for a graphical visualization of data, diagrams and simple plots, *i. e.* (1) the *quantile plot*, (2) the *dot diagram* and the *jitter dot diagram*, (3) the *box-and-whisker plot* and the *notched box-and-whisker plot* are used. Sample distribution represented by the symmetry and tail lengths, skewness and kurtosis is investigated by plots, (4) the *halfsum (or midsum) plot* and (5) the *symmetry plot* (6) the *quantile-box plot*. Construction of an actual sample distribution, *i. e.* the estimate of probability density function, is carried out by (7) the *histogram* or *rootogram* and (8) the *kernel density estimate of the probability density function* while (9) the *quantile-quantile plot* and (10) the *circle plot* are used for comparison of the sample and theoretical distributions.

Step 2: Confirmatory data analysis

When analyzing any data, the sample assumptions are always examined using a check for sample homogeneity, a check for sample normality and a check for independence of sample elements.

Step 3: Data transformations

Power transformation and Box-Cox transformation are used to calculate the re-expressed mean value $\bar{x}_R$, the variance $s^2(\bar{x}_R)$ and the 95% confidence interval of the re-expressed mean value.

Step 4: Determination of point and interval estimates of parameters

The classic and robust point and interval estimates of the parameters of location, scale and shape are calculated. The choice of the type of statistics depends on the asymmetry of the distribution and the results of a check of sample assumptions. If outliers are present in data, robust characteristics are preferred.

Step 5: Statistical hypothesis testing

A simple test of the parameters of population on the basis of one sample uses the 100(1-α)% confidence interval of parameter μ. For testing hypotheses about two populations on the basis of two samples, the first step is the test of normality, followed by a test of the homogeneity of variance in the two samples by the Fisher-Snedecor *F*-test. However, this test is rather sensitive to any deviation of the sample distribution from normality. The classic Student *t*-test T_1 (equal variance in both samples) and T_2 (unequal variances in both samples) are known from most statistical software and textbooks. When both samples deviate in skewness from the normal distribution, the test criterion T_3 (see eq. [3]) is the only convenient method and can be applied only with some advanced statistical software, as the T_3 criterion is rarely available in more general statistical software.

3.2 Software used

For EDA, the creation of diagnostic graphs and the computation of quantile-based characteristics of a sample distribution, the algorithm in *S-Plus* was constructed. In addition, English version of ADSTAT and NCSS2000 were used (18, 19).

4. Results

4.1 Illustrative example

To describe the correct application of statistical techniques, the value of SHBG was chosen as the optimum example. For the best understanding of a procedure resulting in the correct evaluation of differences between the mean values of SHBG in groups with mild and severe acne, each step is described in detail. The original data are shown in Table 4.

Solution

Step 1: Basic diagnostic plots of EDA are used for graphical visualization of both data sets.

Data set SHBG 0: the quantile plot (Figure 2a) exhibits 2 or 3 outliers and an asymmetric distribution as the two curves, classic and empirical robust, are not identical. indicate that the distribution is asymmetric in tails. The halfsum plot (Figure 5a) and the symmetry plots (Figure 6a) exhibit an asymmetric distribution as many points are outside the confidence limits. The quantile-box plot (Figure 7a) shows an asymmetry in the sample distribution, as all three boxes are located asymmetrically inside themselves along the median. The position of the median M is marked by a horizontal line inside the quartile box; a robust estimate of the median confidence interval, $M \pm 1.57\ R_F/\sqrt{n}$, is drawn as a short vertical line at $P_1 = 0.5$. The kernel density estimate of the probability density function (Figure 8a) proves a non-normal distribution, as both curves, the theoretical for a normal distribution and the sample, differ. In the rankit plot (Figure 9a) some points do not fit the line, and therefore a normal distribution is rejected. Analysis of the quantile-quantile plot compares various distributions with the sample and shows that the highest value of the correlation coefficient $r = 0.9843$ detects an exponential distribution. A circle plot (Figure 10a) also proves asymmetric distribution as both, theoretical for normal distribution and empirical, differ.

Because the halfsums Z_Q are not constant and skewnesses S_Q are positive, the right skew distribution is identified. The point estimate of skewness is 1.08 and of kurtosis 4.38, indicating that the sample distribution is asymmetric and not Gaussian.

Data set SHBG 1: the quantile plot (Figure 2b) contains some sample points which are far from fitting the curve, and this proves a strong deviation from the normal distribution. The two curves, the classic one for normal distribution and the empirical one, are not quite identical. Both the dot diagrams (Figure 3b) and the box-and-whisker plot (Figure 4b) indicate an asymmetric distribution with about seven outliers. The halfsum plot (Figure 5b) and the symmetry plot (Figure 6b) show many points outside the lower and upper confidence limits, also indicating an asymmetric distribution. The quantile-box plot (Figure 7b) indicates an asymmetry of distribution and about seven outliers. The circle plot (Figure 10b) shows an asymmetric distribution. The kernel density estimate of the probability density function (Figure 8b) does not prove a normal distribution, as the two curves, the theoretical for a normal distribution and the sample, are not close to each other. In the rankit (quantile-quantile) plot for a normal distribution (Figure 9b), the normal distribution is not proven, as many points do not lie on the straight line. Analysis of the quantile-quantile plot compares vari-

Tab. 4 Pairs of SHBG 0 (mild acne), $n = 42$ and SHBG 1 (severe acne), $n = 45$. The first value in each pair is SHBG 0 and the second is SHBG 1.

55.1	46.4	32.8	131.1	95.6	91.8	69.4	95.2	39.5	11.5
25.4	219.0	0.5	128.5	42.7	165.2	68.5	71.0	42.9	59.0
55.6	103.9	14.6	55.6	53.1	173.5	35.0	83.1	96.3	110.7
46.7	105.7	58.6	200.0	107.3	71.4	146.1	65.6	57.3	40.0
37.7	25.5	62.2	42.1	25.3	179.2	167.0	62.6	43.4	67.4
90.2	70.4	36.2	97.5	47.6	48.7	84.5	47.9	56.8	78.2
65.9	24.0	22.2	189.0	30.5	90.3	39.1	70.3	56.6	194.1
77.1	49.1	37.4	28.0	114.0	83.7	96.7	29.6	14.4	49.6
64.2	52.6	81.8	59.6	31.1	0.0	80.7	0.0	14.1	0.0

Tab. 5 The quantile measures of location, spread and shape for SHBG 0 data.

Quantile	P	Lower quantile Q_L	Upper quantile Q_U	Range R_Q	Halfsum Z_Q	Skewness S_Q	Tails Length T_Q	Pseudo-Sigma G_Q
Median	0.5	55.35	55.35	–	–	–	–	–
Quartile	0.25	37.48	75.18	37.7	56.33	0.73	0	27.97
Octile	0.125	26.04	96.21	70.18	61.13	0.35	0.62	30.51
Sedecile	0.0625	18.88	110.23	91.36	64.55	0.28	0.89	29.86

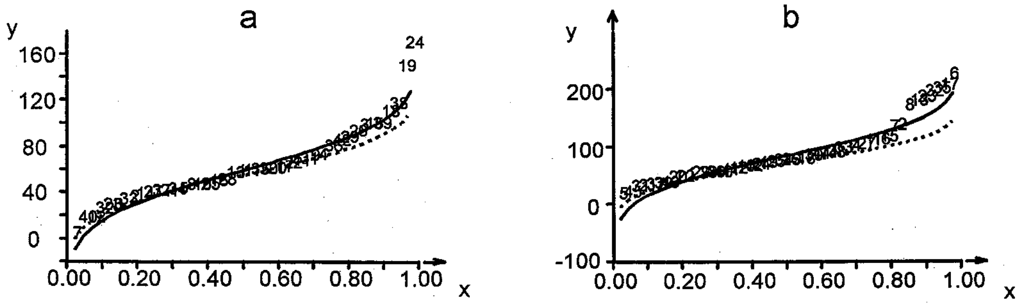

Fig. 2 The quantile plot (x axis: the order statistic $x_{(i)}$, y axis: the rank probability P_i) for (a) SHBG 0 data, (b) SHBG 1 data.

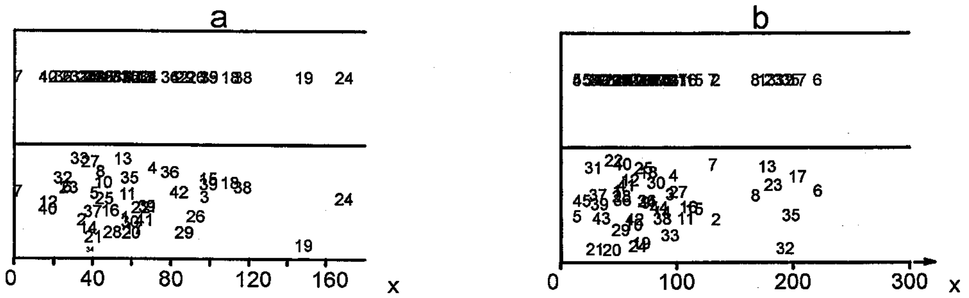

Fig. 3 The dot and jitter dot diagrams (x axis: the order statistic $x_{(i)}$, y axis: random variable) for (a) SHBG 0 data, (b) SHBG 1 data.

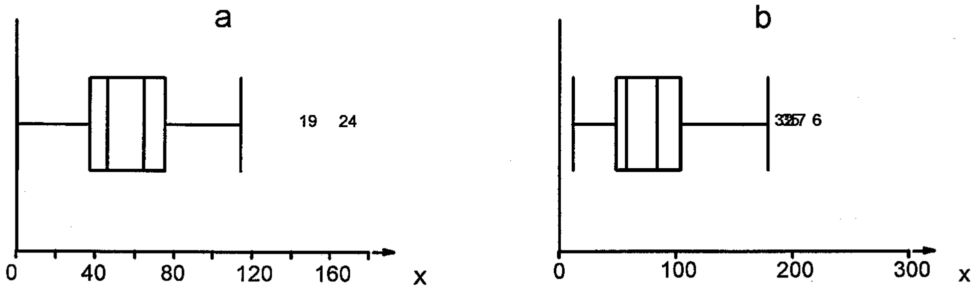

Fig. 4 The box-and-whisker plot (x axis: the order statistic $x_{(i)}$, y axis: no variable, diagram) for (a) SHBG 0 data, (b) SHBG 1 data.

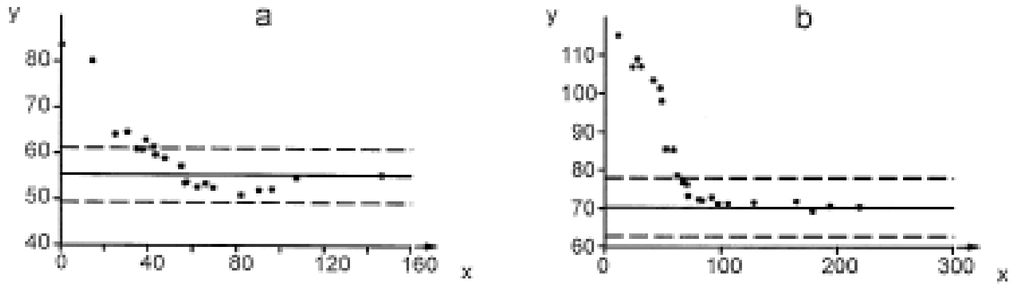

Fig. 5 The halfsum plot (x axis: the order statistic $x_{(i)}$, y axis: the halfsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2)$ for (a) SHBG 0 data, (b) SHBG 1 data.

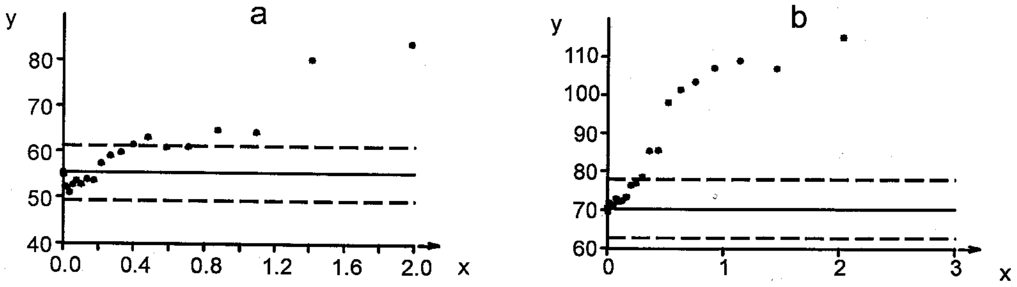

Fig. 6 The symmetry plot (x axis: the quantile of normalized Gaussian distribution $u_{P_i}{}^2$ for $P_i = i/(n+1)$, y axis: the halfsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2)$ for (a) SHBG 0 data, (b) SHBG 1 data.

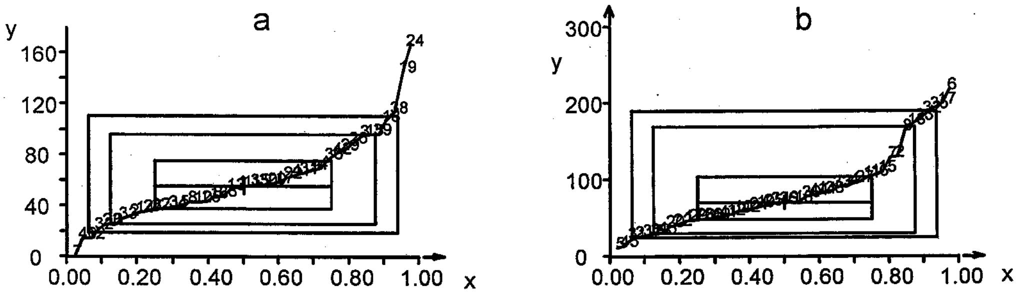

Fig. 7 The quantile-box plot (x axis: $P_i = (i - 1/3)/(n + 1/3)$, y axis: $x_{(i)}$) for (a) SHBG 0 data, (b) SHBG 1 data.

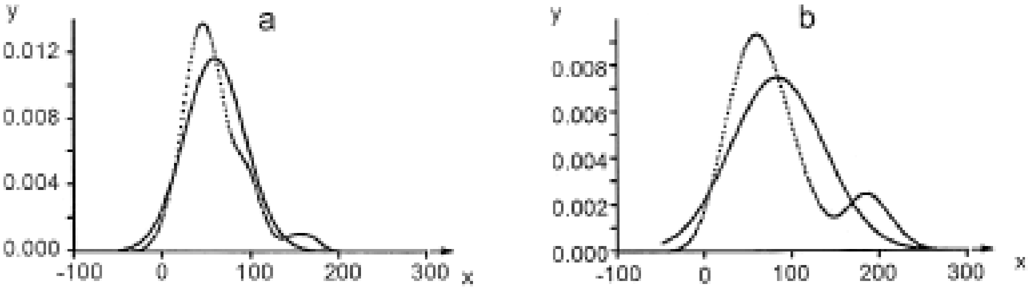

Fig. 8 The kernel estimate of the probability density plot (x axis: x_i, y axis: the kernel estimate $\hat{f}(x)$) for (a) SHBG 0 data, (b) SHBG 1 data.

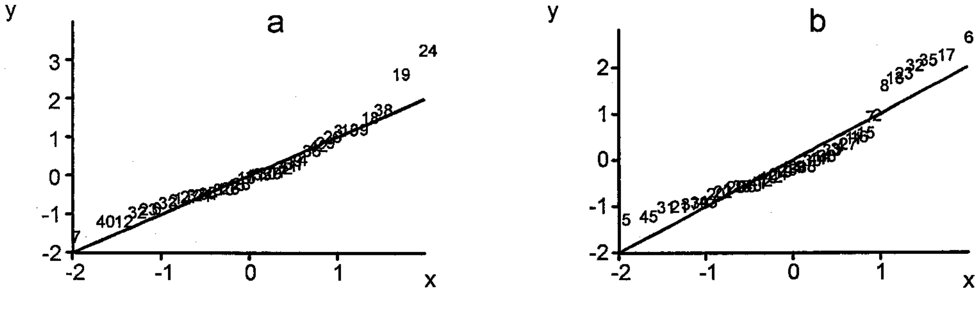

Fig. 9 The quantile-quantile plot (x axis: the theoretical quantile function $Q_T(P_i)$, y axis: the order statistic $x_{(i)}$,) for (a) SHBG 0 data, (b) SHBG 1 data.

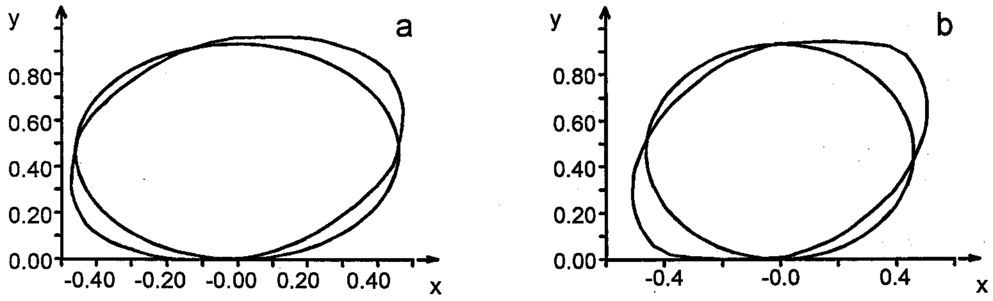

Fig. 10 The circle plot (x axis: the function of x_i, y axis: the relation of x_i) for (a) SHBG 0 data, (b) SHBG 1 data.

Tab. 6 The quantile measures of location, spread and shape for SHBG 1 data.

Quantile	P	Lower quantile Q_L	Upper quantile Q_U	Range R_Q	Halfsum Z_Q	Skewness S_Q	Tails Length T_Q	Pseudo-Sigma G_Q
Median	0.5	70.4	70.4	–	–	–	–	–
Quartile	0.25	48.7	103.9	55.2	76.3	0.53	0	40.95
Octile	0.125	30.35	169.35	139	99.85	0.72	0.92	60.43
Sedecile	0.0625	25.13	19.03	165.2	107.7	0.1	1.09	53.97

ous distributions with the sample one and shows the highest value of the correlation coefficient $r = 0.9744$, detecting the exponential distribution. The point estimate of skewness is 1.00 and of kurtosis 3.12, and this indicates that the sample distribution is asymmetric.

Tail lengths $T_E = 0.62$ (SHBG 0) and 0.92 (SHBG 1) are not close to the tabular value for a normal distribution $T_E = 0.534$, while $T_D = 0.89$ (SHBG 0) and 1.09 (SHBG 1) are not close to the tabular value for a normal distribution $T_D = 0.822$ (Tables 5 and 6). The non-constant halfsums Z_Q and the positive skewness S_Q clearly indicate a skew distribution. The point estimate of skewness and kurtosis indicate that both sample distributions are strongly asymmetric with a slim, sharp peak, and are definitely not normal.

Step 2: Assumptions about the sample

Applying an analysis of basic assumptions about data, the following conclusions were drawn:

(a) *Examination for independence of sample elements:* a test of sample elements independence leads to the test statistic $t_{17} = 0.075$, this being lower than the quantile $t_{0.975}(43) = 2.017$ (SHBG 0) and $t_{17} = 0.212$ also than the quantile $t_{0.975}(46) = 2.013$ (SHBG 1) and therefore indicating that the independence of both sample elements can be accepted.

(b) *Examination for normality of sample distribution:* a combined sample skewness and kurtosis test leads to the test statistic $C_1 = 15.15$ (SHBG 0) and $C_1 = 8.69$ (SHBG 1), this being higher than the quantile $\chi^2(0.95,2) = 5.992$, and therefore indicating that normality in both sample distributions can be rejected.

Step 3: Data transformation

From the plot of the logarithm of the likelihood function for the power transformation the maximum λ on the curve can be read from a graph. For both transformations the corresponding 95% confidence interval does not contain the exponent $\lambda = 1$, so all transformations are statistically significant (Figures 11a, b). The re-transformed mean $\bar{x}_R$ after power ($\lambda = 0.53$ for SHBG 0 and $\lambda = 0.27$ for SHBG 1) and Box-Cox ($\lambda = 0.53$ for SHBG 0 and $\lambda = 0.27$ for SHBG 1) transformations, is $\bar{x}_R = 54.69$ with $L_L = 44.73$, $L_U = 65.57$ for SHBG 0 and $\bar{x}_R = 72.85$ with $L_L = 59.47$, $L_U = 88.33$ for SHBG 1.

Step 4: Determination of point and interval estimates of parameters

A survey of descriptive statistics for parameters of location and the spread of SHBG in both groups of women with acne - computed on the base of EDA with use of ADSTAT and NCSS2000 software - is given below; on the basis of EDA the user should select the most convenient parameter for an actual sample batch:

Data set SHBG 0: $n = 42$, arithmetic mean $\bar{x} = 59.38$, median $\tilde{x}_{0.5} = 55.35$, mode $\hat{x}_M = 56.70$, and thus the trimmed means $\bar{x}(5\%) = 56.96$ with $s(5\%) = 31.46$, $\bar{x}(10\%) = 56.05$ with $s(10\%) = 30.72$, $\bar{x}(40\%) = 53.47$ with $s(40\%) = 29.32$. Parameters of spread: variance $s^2 = 1168.9$, standard deviation $s = 34.19$. Parameters of shape: skewness $\hat{g}_1 = 1.08$, and kurtosis $\hat{g}_2 = 4.38$.

Data set SHBG 1: $n = 45$, arithmetic mean $\bar{x} = 84.30$, median $\tilde{x}_{0.5} = 70.40$, mode $\hat{x}_M = 70.35$, and thus the trimmed means $\bar{x}(5\%) = 81.34$ with $s(5\%) = 56.70$, $\bar{x}(10\%) = 78.13$ with $s(10\%) = 59.54$, $\bar{x}(40\%) = 70.84$ with $s(40\%) = 44.81$. Parameters of spread: variance $s^2 = 2855.0$, standard deviation $s = 53.43$. Parameters of shape: skewness $\hat{g}_1 = 1.00$, and kurtosis $\hat{g}_2 = 3.12$.

For the best point estimate of the parameter of loca-

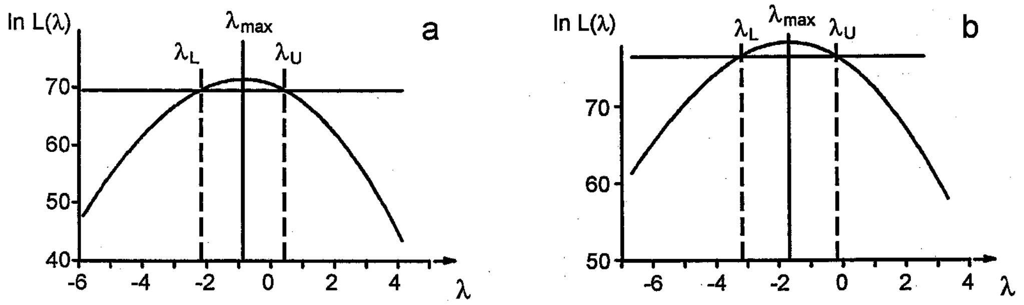

Fig. 11 The plot of the logarithms of the likelihood function ln $L(\lambda)$ in dependence on the power λ and estimation of the optimal power λ_{max} with its lower λ_L and upper λ_U limits of the confidence interval for the statistical probability 95% (*x* axis: ln $L(\lambda)$, *y* axis: λ) for (a) SHBG 0 data, (b) SHBG 1 data.

Tab. 7 A comparison of two sample means for selected steroids with the use of Student *t*-test of eq. [4].

Steroid	n	Mean (lower; upper limits)	SD	Median	Re-transf. mean (lower; upper limits)	Re-transf. stan. dev.	Skewness	Kurtosis	Normality	Test of H_0: equal variances	Test of H_0: equal means
TESTO-0	43	2.10 (1.78; 2.42)	1.05	2.00 (1.66; 2.34)	1.98 (1.67; 2.31)	1.04	0.57	2.85	Accepted	H_0 is accepted	H_0 is accepted
TESTO-1	46	1.91 (1.61; 2.21)	1.01	1.65 (1.37; 1.93)	1.71 (1.46; 1.99)	0.89	**1.12**	4.02	Rejected		
SHBG-0	42	59.4 (48.7; 70.0)	34.19	55.4 (44.1; 66.6)	54.7 (44.7; 65.6)	33.4	**1.08**	**4.38**	Rejected	H_0 is rejected	H_0 is rejected
SHBG-1	45	84.3 (68.2; 100.4)	53.4	70.4 (55.8; 85.0)	72.9 (59.5; 88.3)	47.9	**1.00**	3.12	Accepted		
ADION-0	42	9.22 (7.66; 10.79)	5.03	8.23 (6.98; 9.47)	8.04 (6.91; 9.38)	1.71	**1.71**	**6.52**	Rejected	H_0 is accepted	H_0 is accepted
ADION-1	46	9.66 (8.45; 10.86)	4.06	9.71 (8.25; 11.16)	9.25 (8.08; 10.49)	4.06	0.37	2.48	Accepted		
DHEAS-0	43	6.30 (5.45; 7.15)	2.76	5.70 (4.35; 7.05)	5.88 (5.11; 6.73)	2.62	0.67	2.69	Accepted	H_0 is accepted	H_0 is accepted
DHEAS-1	47	6.55 (5.64; 7.46)	3.09	5.75 (4.97; 6.53)	5.83 (5.10; 6.68)	1.67	**0.93**	3.04	Rejected		
DHEA-0	43	7.99 (5.89; 10.10)	6.83	6.70 (5.84; 7.56)	6.45 (5.58; 7.51)	0.71	**4.59**	**26.87**	Rejected	H_0 is accepted	H_0 is accepted
DHEA-1	46	7.34 (6.12; 8.56)	4.09	6.63 (5.67; 7.58)	6.43 (5.42; 7.60)	3.66	**0.93**	3.01	Rejected		

Bold: Gaussian distribution is rejected

tion, robust estimate median M = 55.35±11.27 (SHBG 0) and median M = 70.40±14.62 (SHBG 1); for parameter of spread also the robust estimate of standard deviation of the median s = 34.47 (SHBG 0), and the standard deviation of the median s = 46.08 (SHBG 1) may be used. The two parameters describing shape *i. e.* the estimates of skewness $\hat{g}_1$ = 1.08 (SHBG 0), 1.00 (SHBG 1) and kurtosis $\hat{g}_2$ = 4.38 (SHBG 0), 3.12 (SHBG 1), indicate that the distributions of both the SHBG 0 and the SHBG 1 samples are asymmetric.

The interval estimate of the parameter of location may be described by the confidence interval for medians M = 55.35 as being L_L = 44.08 and L_U = 66.62 (SHBG 0) and M = 70.40 as being L_L = 55.78 and L_U = 85.02 (SHBG 1), in which an unknown concentration exists with a 95% statistical probability.

Step 5: Statistical hypothesis testing: comparisons of the means of two samples

All of the EDA display techniques prove that the distribution of sample SHBG 0 does not come from a population with a symmetric and normal distribution, and two outliers, while the distribution of sample SHBG 1 is also asymmetric with seven outliers. For biochemical and clinical data no outliers can be excluded from data batch because of a danger of losing valuable information. Therefore, all the trimmed means are unusable (17).

Despite the overlap of the confidence intervals for a best estimate of the mean values indicating that no difference is found at a 95% probalility level between the mean values for SHBG 0 and SHBG 1, a more reliable statistical test of the two sample means T_3 can be applied: comparing two sample means, $\bar{x}_A = 59.38$, $s^2_A = 1168.9$ (for SHBG 0), $\bar{x}_B = 84.30$, $s^2_B = 28550.0$ (for SHBG 1); both samples have a non-normal distribution, and a modified Fisher-Snedecor test proves the heterogeneity of the variances of both samples, $F = 2.442 > F(0.95, 28, 30) = 2.092$. Therefore, for the comparison testing of two means a modified Student *t*-test for the non-normal distribution T_3 should be used, and here it was found that $T_3 = 2.467 > t(0.95, 77) = 1.991$, which indicates that the sample SHBG 0 means $\bar{x}_A$ and the sample SHBG 1 means $\bar{x}_B$ are not equal.

4.2 Comparisons of the means of two samples for selected steroids

The procedure demonstrated using an example of SHBG was applied to five selected steroids. As it is apparent from Table 7, no significant differences in the serum levels of the steroids were found between the groups of women with different acne severity, which means that no association was found between acne severity and the serum levels of steroids. On the other hand, higher levels of SHBG were found in women with more severe acne.

5. Conclusions

To obtain undistorted and accurate results in an effective statistical analysis of univariate biochemical data, EDA should be employed to uncover typical features and patterns. The second step is CDA, where probability models are created and tested. EDA is very effective in the investigation of the statistical behavior of experimental data coming from new or non-standard analytical techniques. From a list of various parameters of location and spread it enables the selection of an estimation of the best one. Regarding the evaluation of differences in the value of SHBG between two groups of women with different acne severity, it is evident that improper use of standard statistical technique could result in a misinterpretation of the data, and it is necessary to analyze data distribution prior to the selection of the appropriate statistical method.

As concerns the biological consequences of the results, no relationship were found between acne severity and the levels of serum testosterone and androgen precursors, while higher levels of SHBG were detected in patients with more severe form of acne. The later finding is in accordance with the study of Palatsi *et al.* (20).

Acknowledgements

The authors gratefully acknowledge the financial support of the Grant Agency of the Czech Republic for this project (Grant No 303/00/1559).

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. Stoodley K. Applied and computational statistics. Chichester: Ellis Horwood, 1984.
5. Cunliffe WJ, Shuster S. Pathogenesis of acne. Lancet 1969; 1:685–7.
6. Darley CR, Moore JW, Besser GM, Munro DD, Edwards CR, Rees LH, Kirby JD. Androgen status in women with late onset or persistent acne vulgaris. Clin Exp Dermatol 1984; 9:28–35.
7. Henze C, Hinney B, Wuttke W. Incidence of increased androgen levels in patients suffering from acne. Dermatology 1998; 196:53–4.
8. Lucky AW. Endocrine aspects of acne. Pediatr Clin North Am 1983; 30:495–9.
9. Lucky AW, McGuire J, Rosenfield RL, Lucky PA, Rich BH. Plasma androgens in women with acne vulgaris. J Invest Dermatol 1983; 81:70–4.
10. Schiavone FE, Rietschel RL, Sgoutas D, Harris R. Elevated free testosterone levels in women with acne. Arch Dermatol 1983; 119:799–802.
11. Scholl GM, Wu CH, Leyden J. Androgen excess in women with acne. Obstet Gynecol 1984; 64:683–8.
12. Timpatanapong P, Rojanasakul A. Hormonal profiles and prevalence of polycystic ovary syndrome in women with acne. J Dermatol 1997; 24:223–9.
13. Vexiau P, Husson C, Chivot M, Brerault JL, Fiet J, Julien R, *et al.* Androgen excess in women with acne alone compared with women with acne and/or hirsutism (see comments). J Invest Dermatol 1990; 94:279–83.
14. Cunliffe WJ. Acne, hormones, and treatment (editorial). Br Med J (Clin Res Ed) 1982; 285:912–3.
15. Hoaglin DC, Mosteler F, Tukey JW, editors.Understanding robust and exploratory data analysis. New York: Wiley, 1983.
16. Lejenne M, Dodge Y, Koelin E. Proceedings of the conference COMSTAT'82, Toulouse: P. 173 (Vol. III).
17. Meloun M, Militký J, Forina M. Chemometrics for analytical chemistry, Volume 1. PC-aided statistical data analysis. Chichester: Ellis Horwood, 1991. Volume 2. Regression model building and testing. Chichester: Ellis Horwood, 1994.
18. ADSTAT. Trilobyte Statistical Software Ltd, Pardubice, Czech Republic, 1998, http://www.trilobyte.cz.
19. NCSS. NCSS Statistical Software, Kayswille, Utah, USA, 1998, http://www.ncss.com/.
20. Palatsi R, Hirvensalo E, Liukko P, Malmiharju T, Mattila L, Riihiluoma P, Ylostalo P. Serum total and unbound testosterone and sex hormone binding globulin (SHBG) in female acne patients treated with two different oral contraceptives. Acta Derm Venereol 1984; 64:517–23.

Received 16 November 2000, revised 26 January 2001, accepted 30 January 2001

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