

Rigorous approach to the univariate data analysis

Rigorozní přístup k analýze jednorozměrných dat

Prof. RNDr. Milan Meloun, DrSc.

Katedra analytické chemie,
Universita Pardubice, 53210 Pardubice

Email: milan.meloun@upce.cz

Telefon: 466037026

Rigorous procedure
in 2 steps

UNIVARIATE DATA ANALYSIS

1st stage:

EXPLORATORY DATA ANALYSIS (EDA)

- 1. EDA DIAGNOSTIC PLOTS AND DISPLAYS**
- 2. EXAMINING A SAMPLE DISTRIBUTION**
- 3. DATA TRANSFORMATION**

2nd stage:

CONFIRMATORY DATA ANALYSIS (CDA)

- 1. TEST OF BASIC ASSUMPTIONS ABOUT DATA**
- 2. CONSTRUCTION OF PROBABILITY DENSITY FUNCTION**
- 3. POINT ESTIMATES FOR PARAMETERS OF LOCATION, SPREAD AND SHAPE**
- 4. INTERVAL ESTIMATES FOR PARAMETERS OF LOCATION AND SPREAD**
- 5. STATISTICAL HYPOTHESIS TESTING**

Vyšetření předpokladů o výběru

Examination of the sample assumptions

Outliers

Odlehlé body

Symmetry

Symetrie

Heteroscedasticity

Konstantní rozptyl

Sample size

Velikost výběru

Problem 2.1 Analysis of data with normal and log.-normal distribution

The exploratory data analysis of simulated sample data, from (a) normal distribution $N(10, 0.1)$ denoted by **norm** and (b) log.-normal distribution $In(5, 2)$ denoted by **log**.

Data: (a) Sample **norm**: 10.0005 10.185 10.05 10.042 10.197 10.021 10.033 9.99985 9.826 10.076
10.053 10.079 9.9998 10.026 9.9969 9.98995 10.035 10.064 9.9985 10.093 10.132 10.047 9.877 9.931
10.002 9.929 9.959 9.846 10.029 10.029 9.994 10.113 10.158 9.999 10.1414 10.004 10.067 9.995 10.091
10.088 10.06 9.9998 10.017 9.865 9.907 10.037 10.081 10.018 9.987 10.115 10.037 10.063 9.928 9.975
9.937 9.933 9.942 10.106 10.039 9.989 9.906 9.894 9.946 9.955 9.98 10.108 10.05 9.948 9.974 9.986
9.986 10.105 10.037 9.955 10.025 9.949 9.879 10.042 10.052 9.92 10.064 10.075 10.028 9.955 9.987
9.957 9.969 9.9999 9.9995 10.021 10.069 9.975 10.109 10.024 9.984 10.122 9.885 10.011 10.013 10.011

(b) Sample **log**: 2.408 5.389 2.259 2.439 2.173 1.157 0.892 0.498 0.351 1.229 1.356 4.719
1.445 1.023 1.723 0.572 2.012 0.212 0.305 0.993 11.993 2.247 0.973 0.418 2.27 12.03 1.321 3.076 1.355
4.54 0.216 10.159 0.346 1.078 0.206 0.116 1.733 0.55 0.762 2.689 1.798 1.522 2.763 0.536 0.21 2.462
0.516 0.421 1.588 2.54 7.48 0.881 0.841 1.039 0.966 0.49 1.476 1.185 0.875 0.557 1.464 0.308 0.097
1.137 2.247 0.084 0.217 1.885 0.204 2.786 2.341 0.466 0.712 0.401 0.404 1.027 0.623 0.139 2.905 0.111
0.958 0.188 0.611 0.243 5.331 0.745 0.367 0.919 1.236 1.912 2.816 0.666 4.972 0.451 1.316 3.241 0.316
2.2 8.291 0.815

Median or the mean ?

Medián nebo průměr ?

1st step:

EDA

Exploratory Data Analysis:
symmetry, outliers

Outliers and Extremes in Representative Random Sample

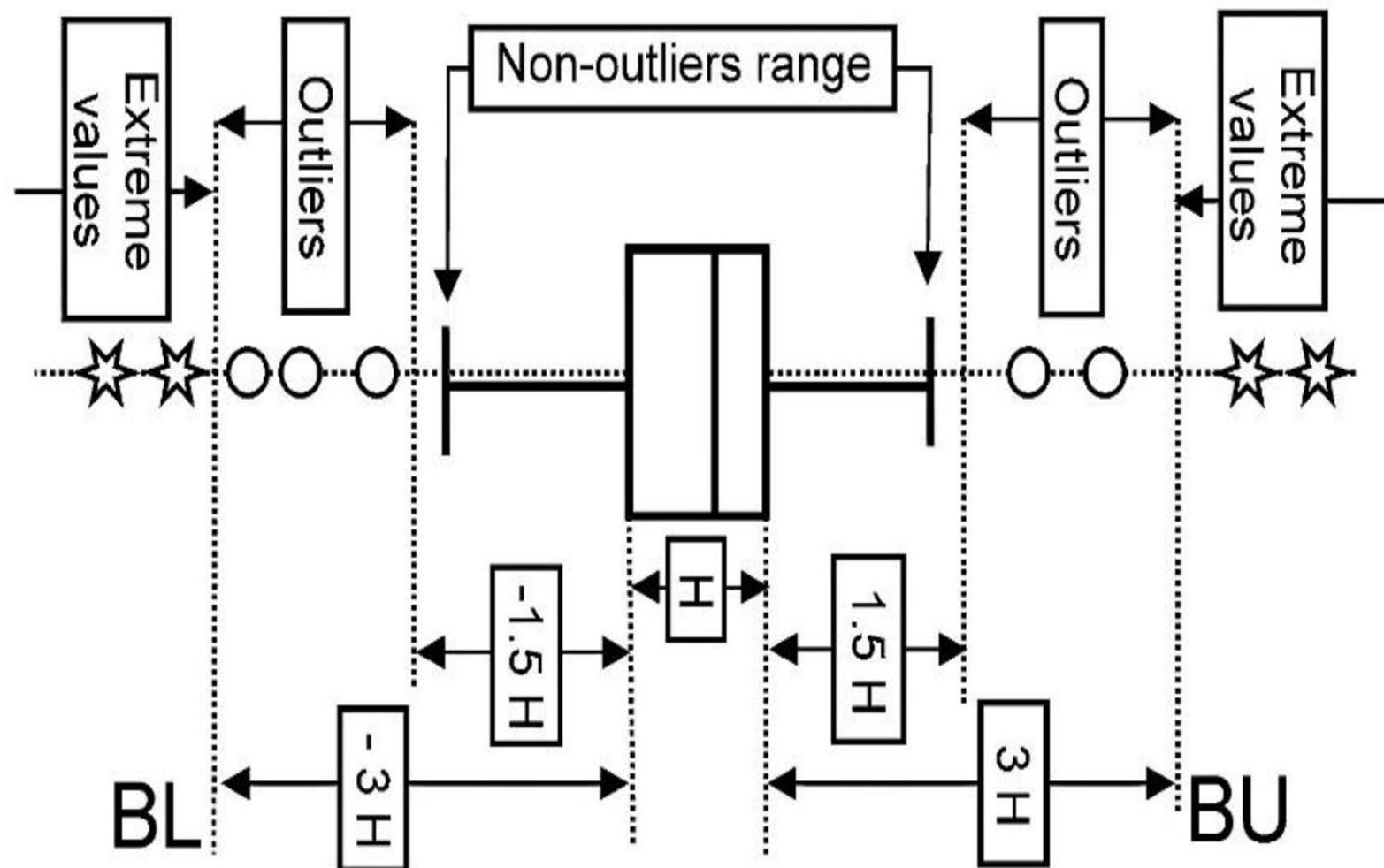
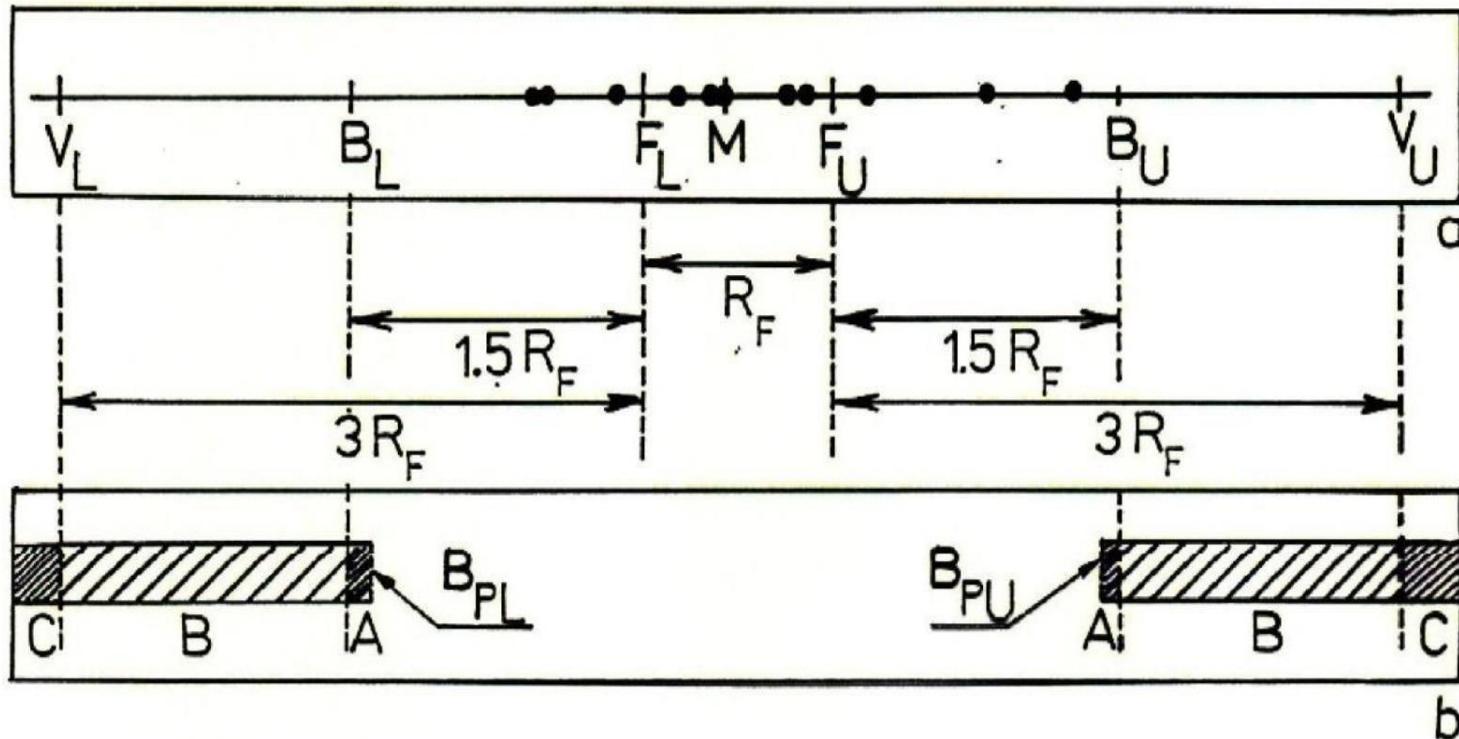


Figure 2.30 Outliers are points beyond the inner bounds BL and BU.

Letter values and Bounds

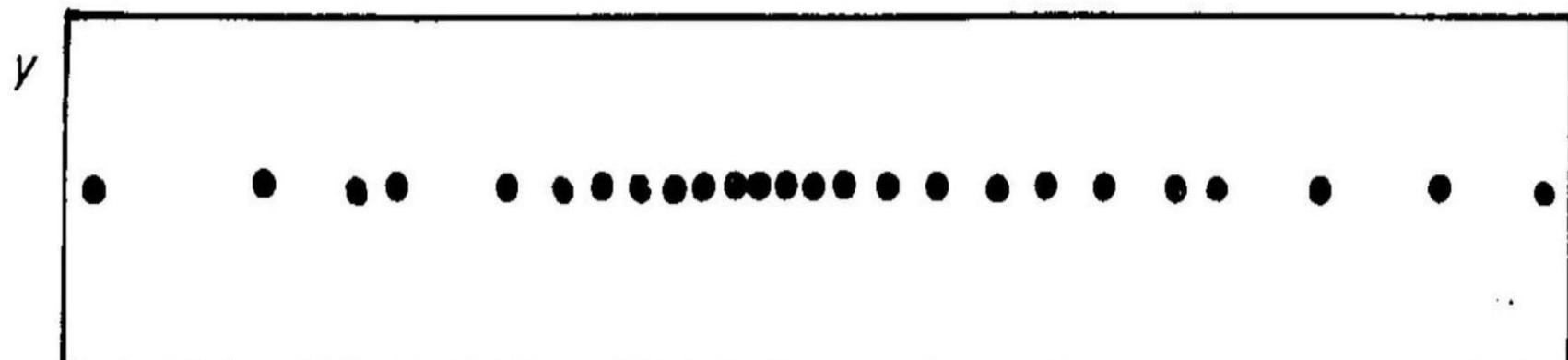


Dot diagram (x -axis x values, y -axis is a level $y = 0$) shows:

- (a) the **dot diagram with median M ,**
 F_L (lower) and F_U (upper) quartiles,
inner B_L (lower) and B_U (upper) bounds,
and **outer V_L (lower) and V_U (upper) bounds,**
- (b) the area of outliers: **A** close outliers, **B** near far outliers, **C** far outliers.

One-dimensional scatter plot

a



b

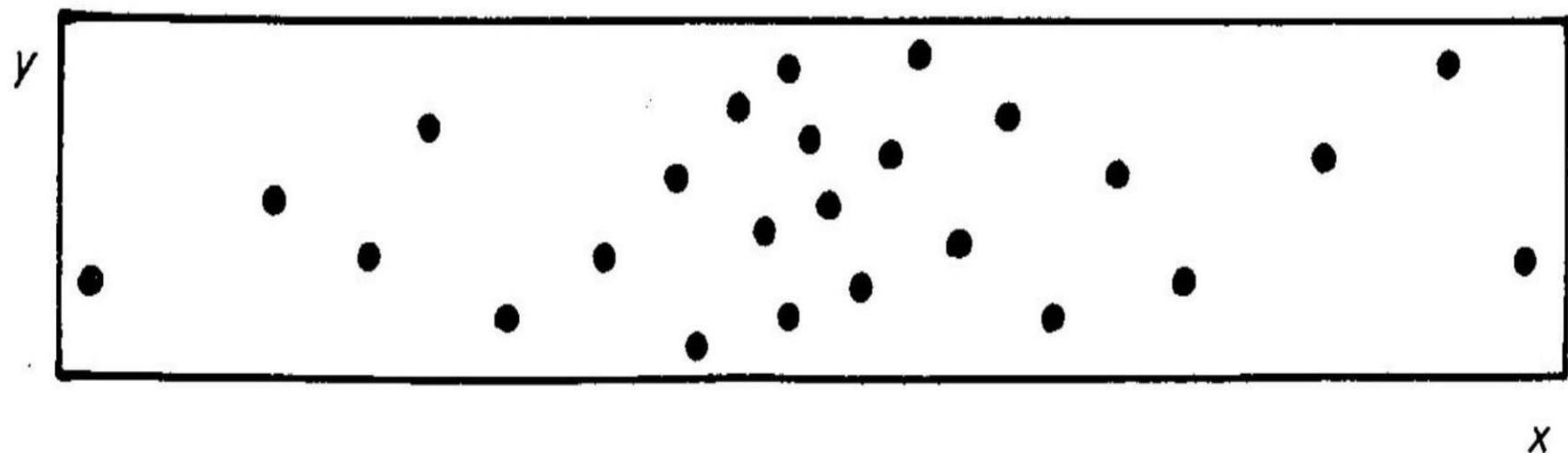


Fig. 2.6—Examples of (a) a dot diagram (G2) and (b) a jittered-dot diagram (G3).

Various graphical plots and displays in EDA

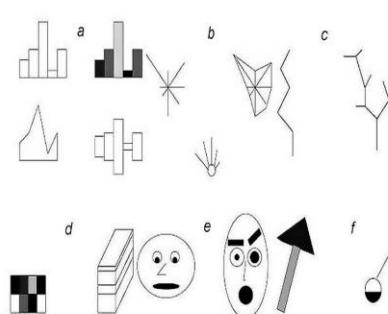
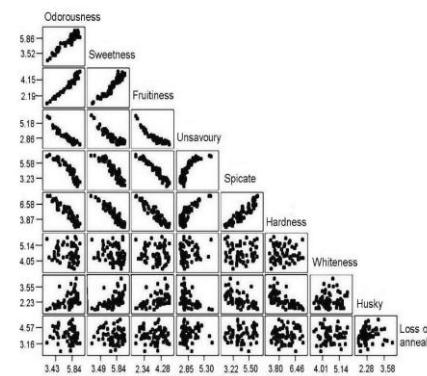
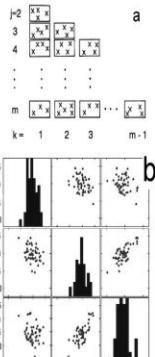
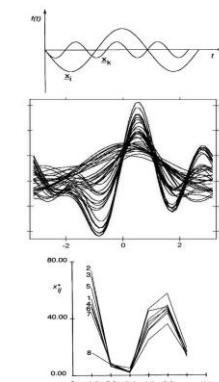
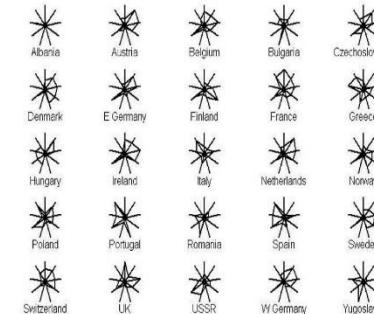
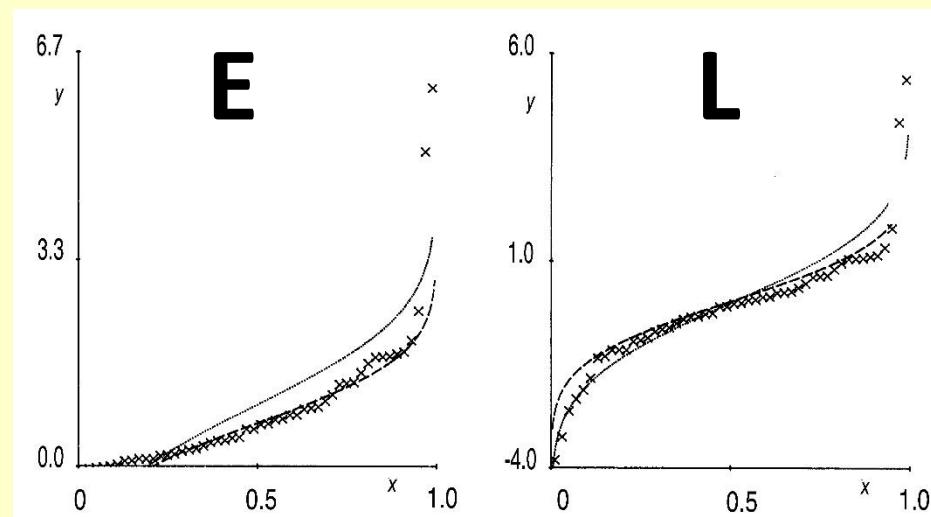
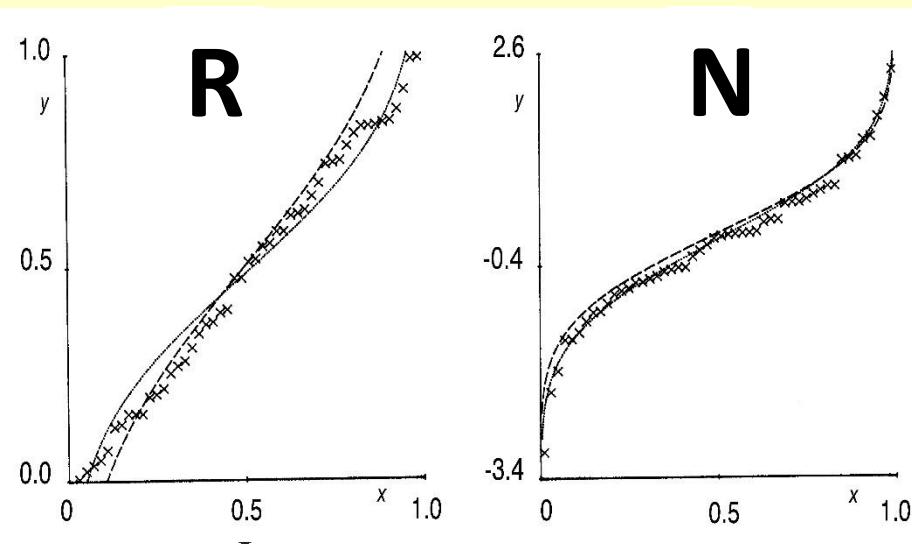


Figure 4.12 Examples of various glyphs: (a) Variations on profiles; (b) Stars/Metroglyphs; (c) Stick figure icons and Trees; (d) Autoglyphs and Boxes; (e) Chernoff faces; (f) Arrows and Weather vanes.

From left: Red meat, White meat, Eggs, Milk, Fish, Cereals, Starch, Nuts, Fruits & Vegetables

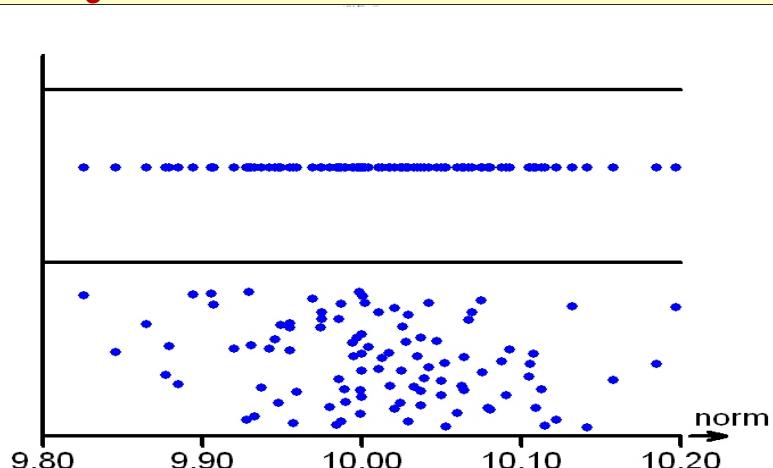


Quantile plot: symmetry and outliers

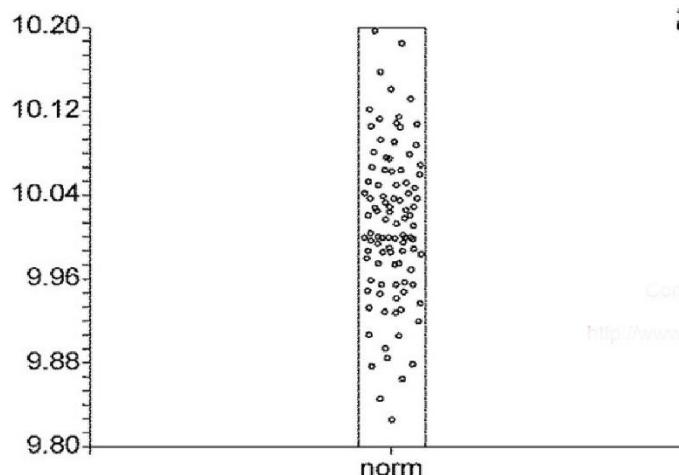
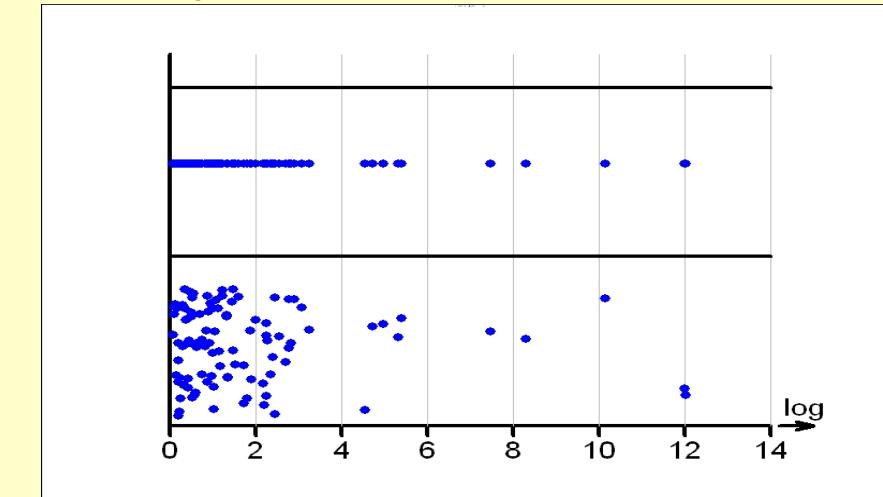


Jittered dot diagram

Symmetric distribution

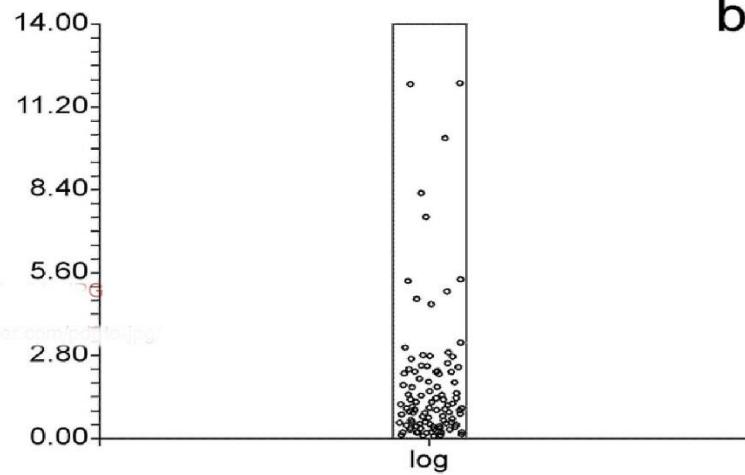


Asymmetric distribution



a

Converted by PDF
<http://www.PDF-Help.com/convert-to-Jpg/>



b

Figure 2.7 Another form of jittered dot diagram G3 (x-axis: x values; y-axis: a small interval of random numbers) for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log-normal) distributions, NCSS2000.

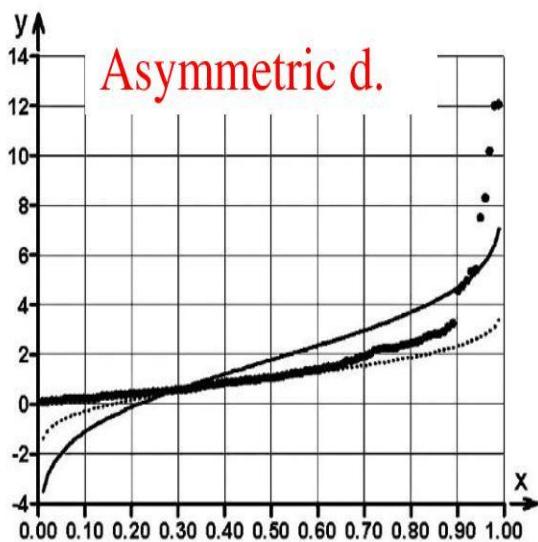
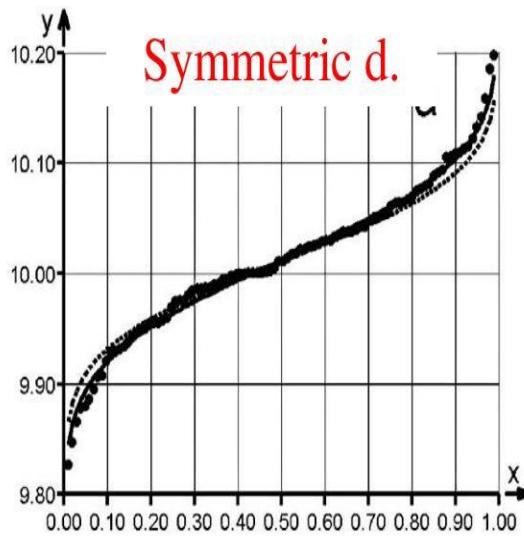


Figure 2.5 Quantile plot G1 (x-axis: the cumulative order probability P_i ; y-axis: the order statistic $x_{(i)}$) for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, (quantile functions for the normal distribution : classical – solid line, robust dotted line) QC-EXPERT.

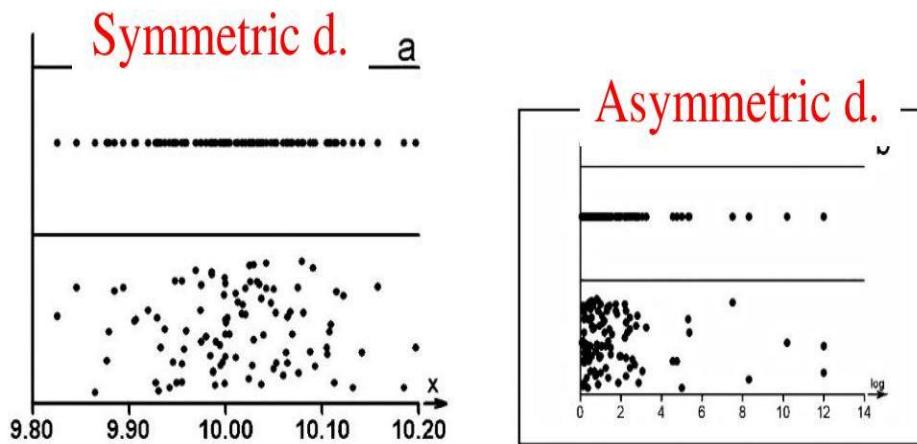


Figure 2.6 Dot diagram G2 above (x-axis: x values; y-axis: selected level, usually $y = 0$) and jittered dot diagram G3 below (x-axis: x values; y-axis: a small interval of random numbers), for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, QC-EXPERT.

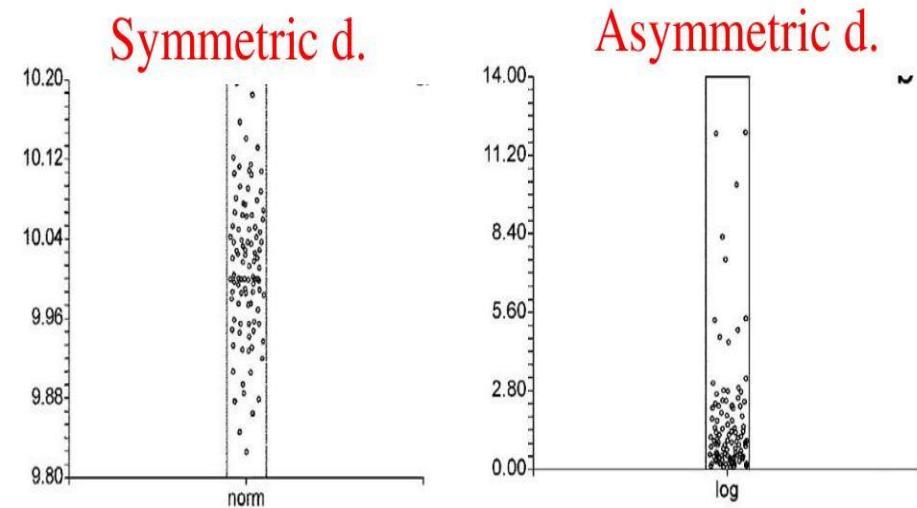


Figure 2.7 Another form of jittered dot diagram G3 (x-axis: x values; y-axis: a small interval of random numbers) for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, NCSS2000.

Symmetric d.

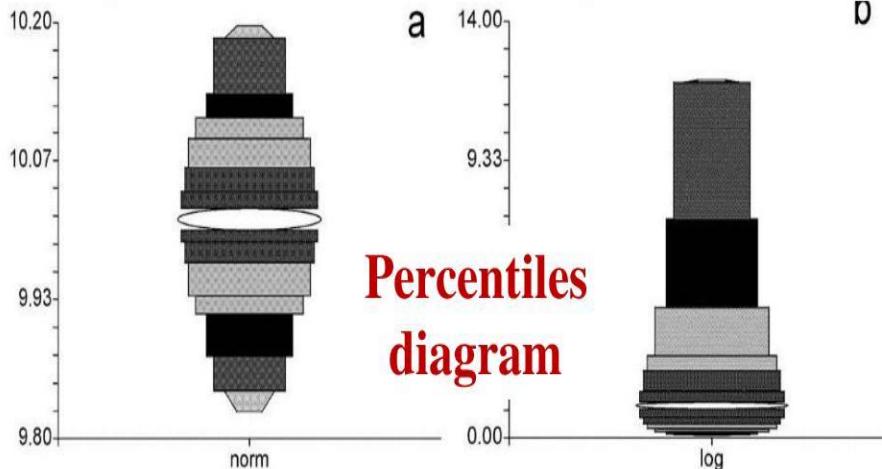


Figure 2.8 A percentiles diagram for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, NCSS2000.

Asymmetric d.

Symmetric d.

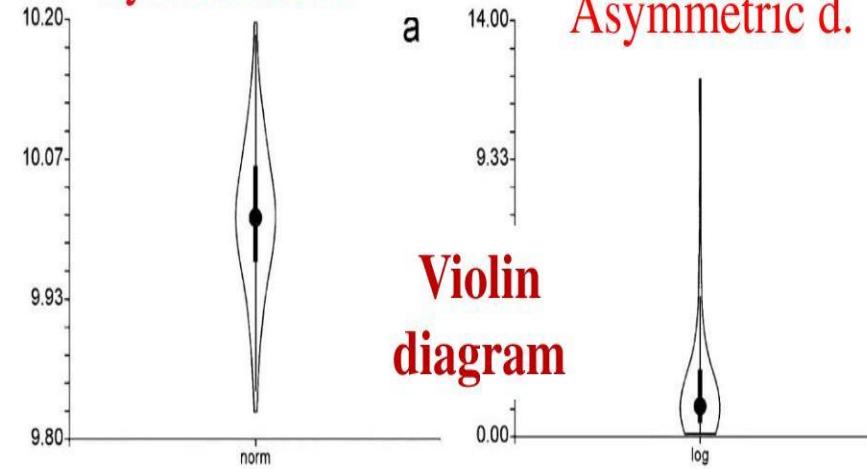


Figure 2.9 A violin diagram for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, NCSS2000.

Box-and-whisker plot

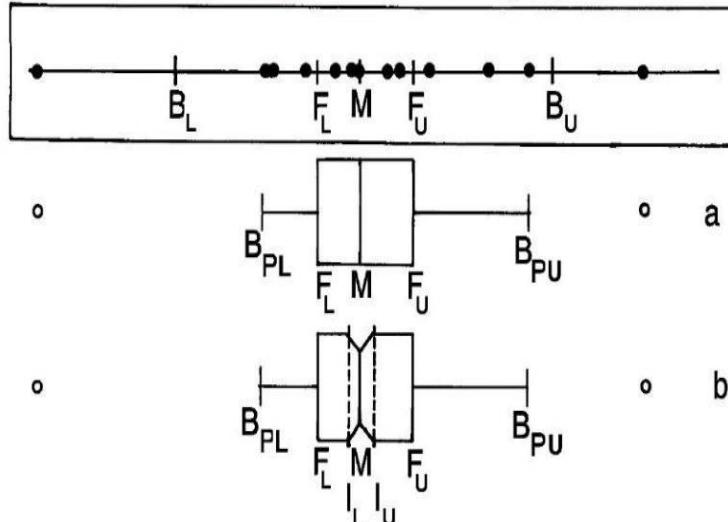


Figure 2.10 Scheme of the box-and-whisker plot (x-axis: x values; y-axis: any suitable interval).

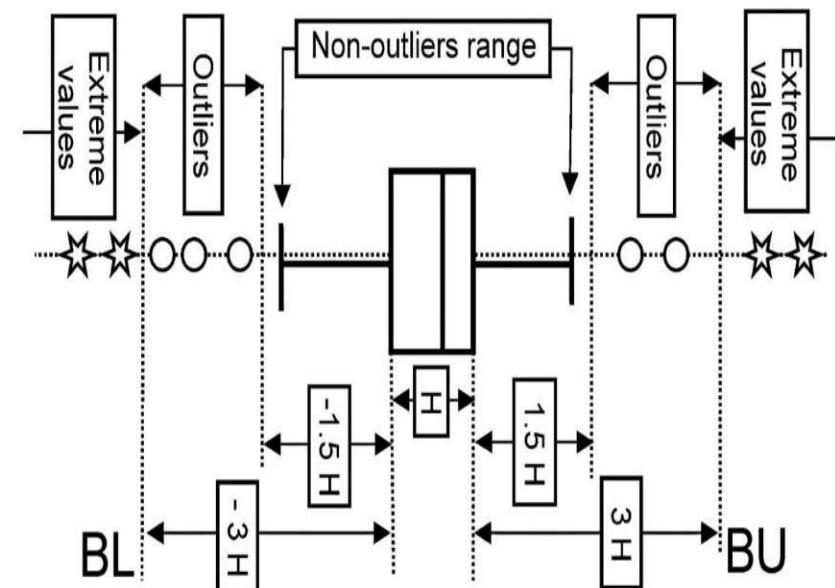
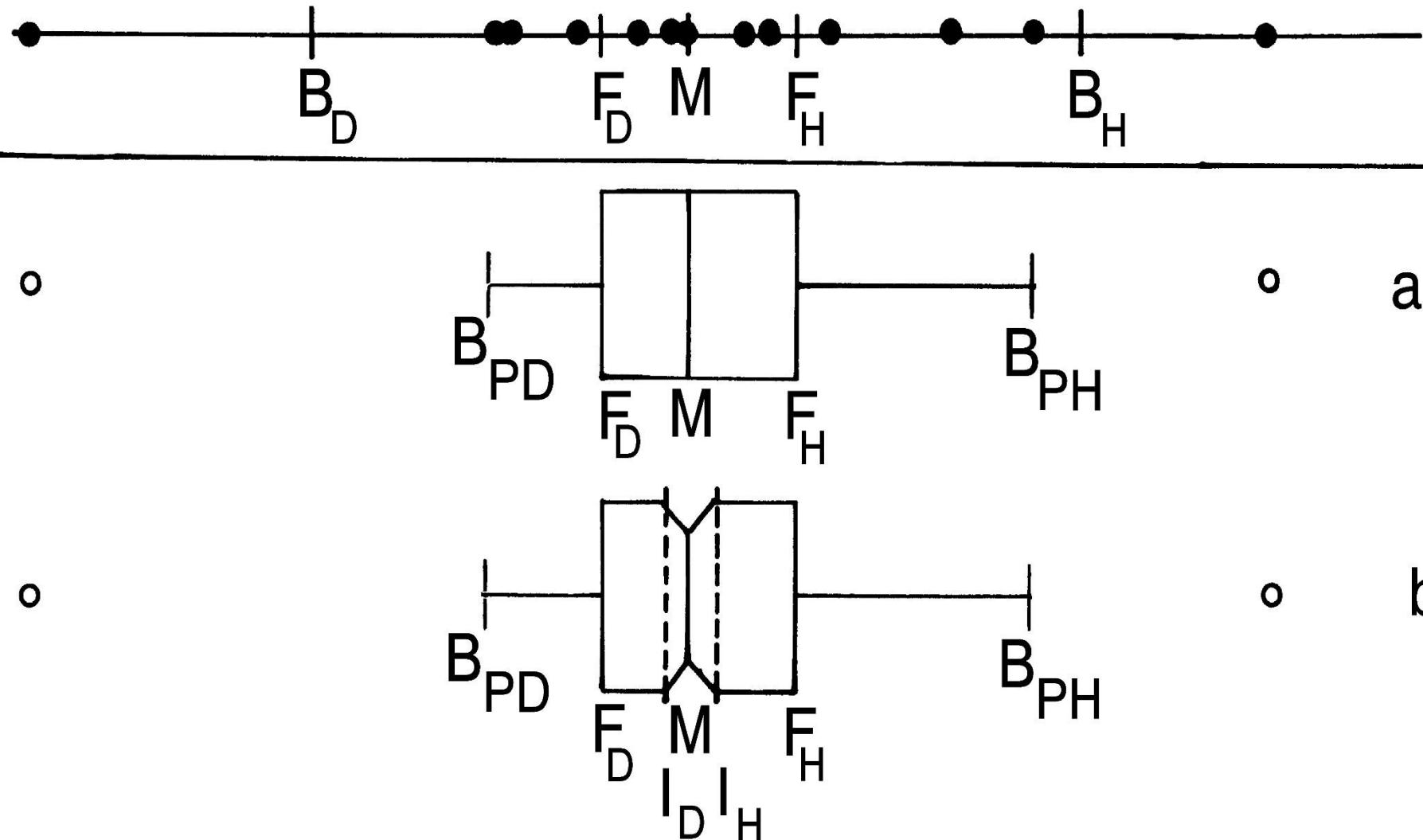


Figure 2.30 Outliers are points beyond the inner bounds BL and BU.

Box-and-Whisker Plot



Diagnostics for Examination of the distribution symmetry

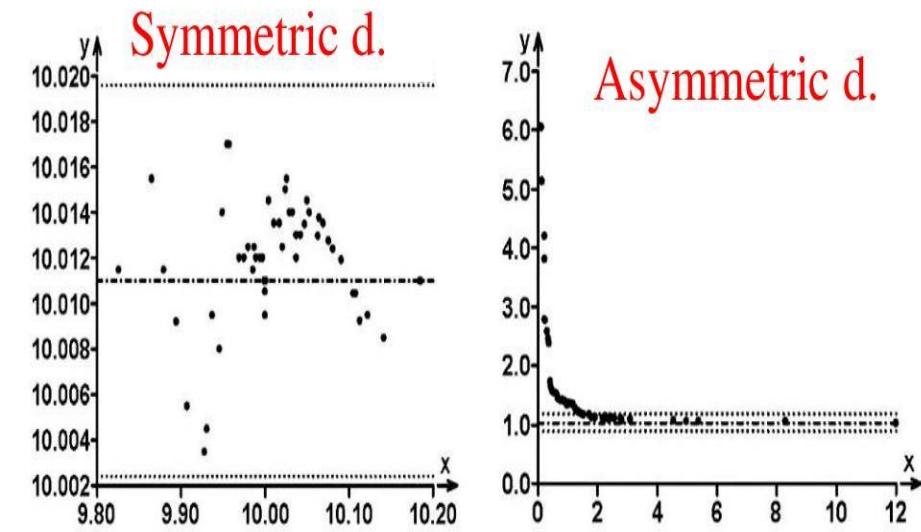


Figure 2.12 The midsum plot G6 (x-axis: the order statistic $x_{(i)}$; y-axis: the midsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2$) for samples with (a) norm, symmetric (Gaussian, normal), and (b) log, asymmetric (log.-normal) distributions, QC-EXPERT.

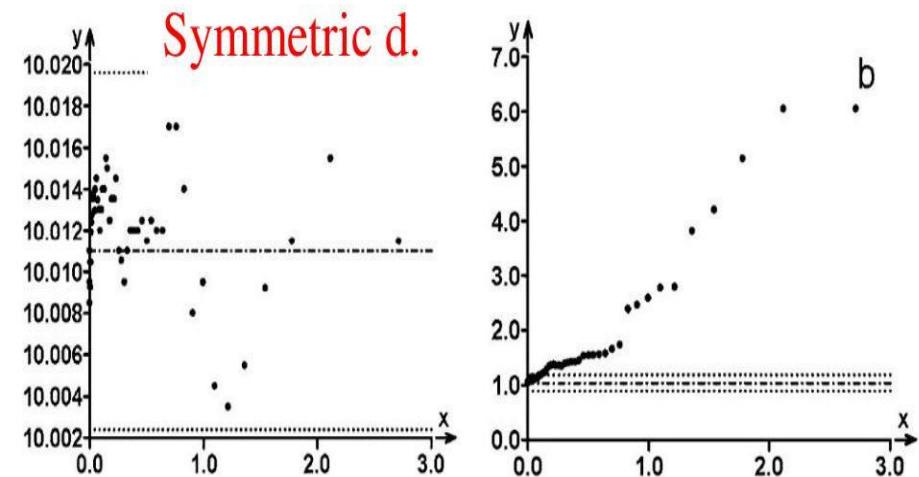


Figure 2.13 The symmetry plot G7 (x-axis: the quantile $u_{P_i}^2 / 2$ for $P_i = i/(n+1)$; y-axis: the midsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2$) for samples with (a) norm, symmetric (Gaussian, normal), and (b) log, asymmetric (log.-normal) distributions, QC-EXPERT.

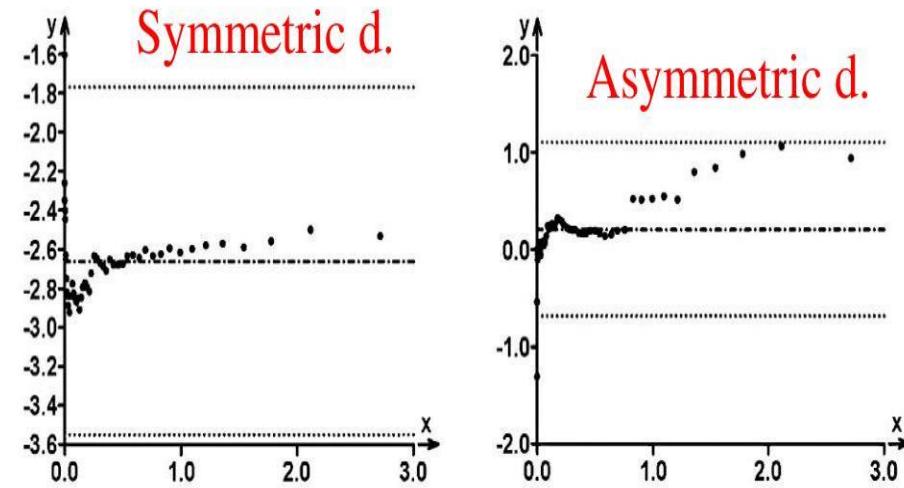
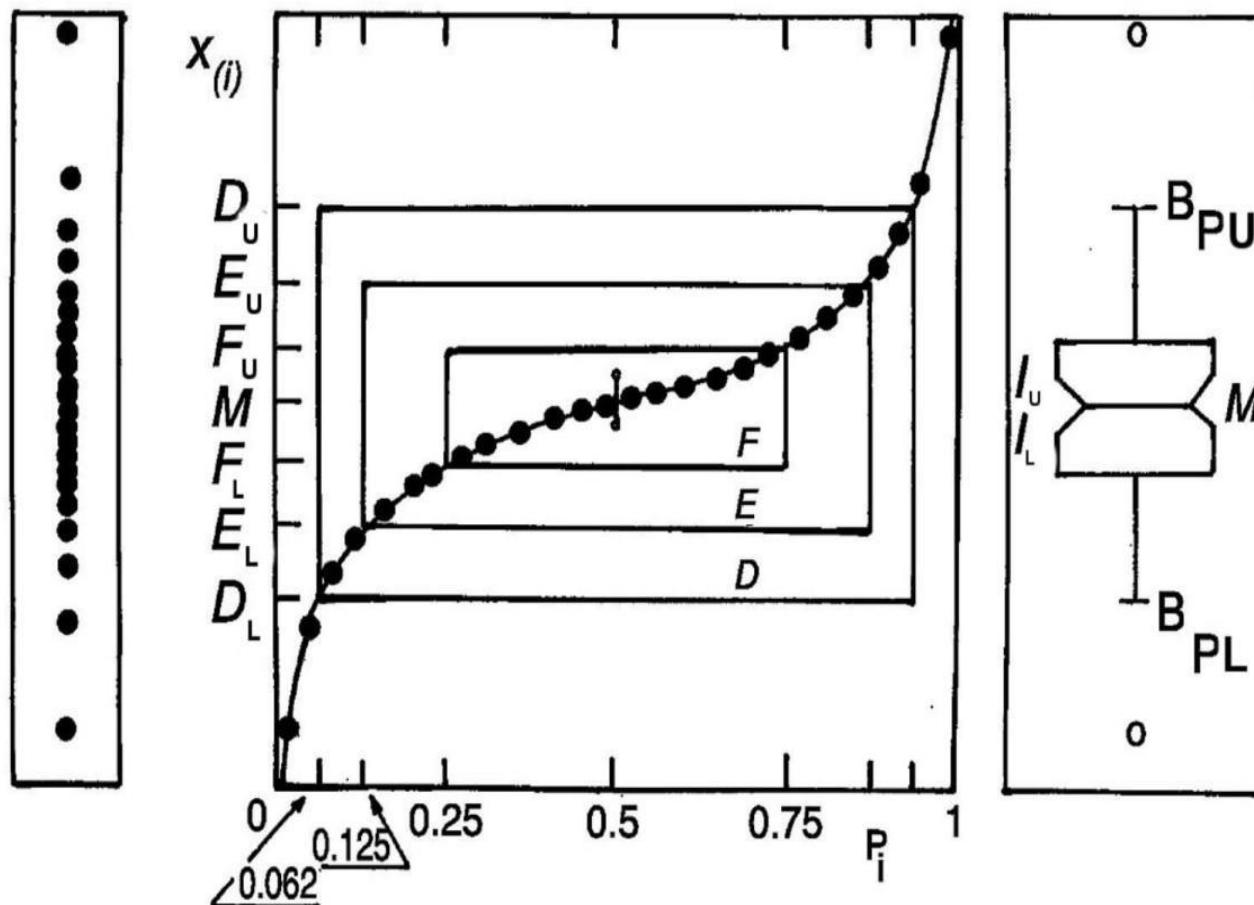


Figure 2.14 The kurtosis plot G8 (x-axis: the quantile $u_{P_i}^2 / 2$ for $P_i = i/(n+1)$; y-axis: the quantity $\ln[(x_{(n+1-i)} - x_{(i)}) / -2u_{P_i}]$) for samples with (a) norm, symmetric (Gaussian, normal), and (b) log, asymmetric (log.-normal) distributions, QC-EXPERT.

Quantile-box plot



Construction of the quantile-box plot

(*x*-axis: the order probability P_i , *y*-axis: the order statistic $x_{(i)}$).

The dot diagram (left) and the notched box-and-whisker plot (right) are for comparison.

Diagnostics for an Examination of

- 1) the distribution symmetry and
- 2) outliers

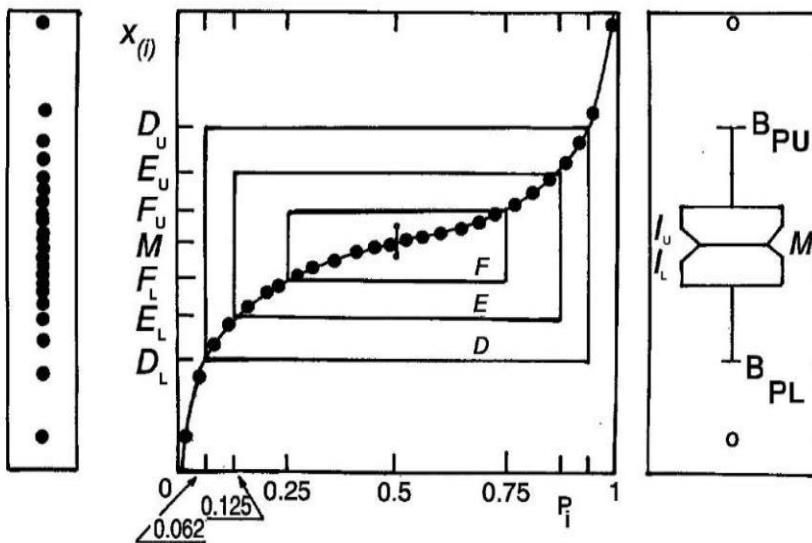


Figure 2.16 Construction of the quantile-box plot G10 (x-axis: the order probabil P_i , y-axis: the order statistic $x_{(i)}$). The dot diagram (left) and the notch box-and-whisker plot (right) are given for comparison of an acti distribution.

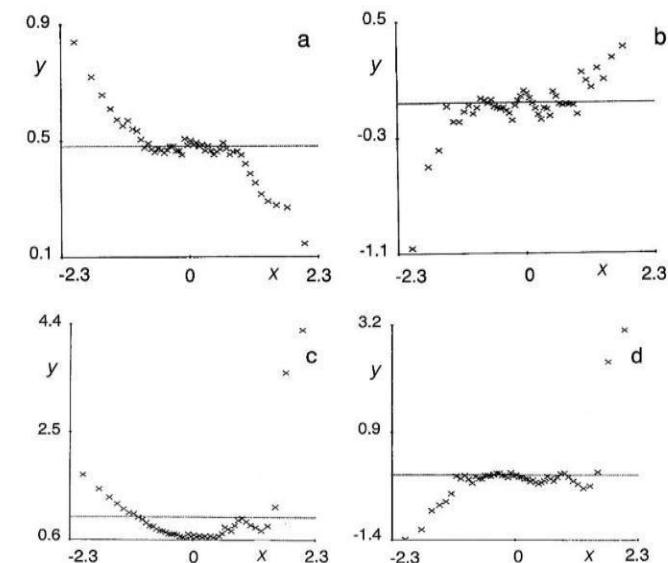


Figure 2.15 The differential quantile plot G9 (x-axis: the quantile u_{P_i} ; y-axis: the deviation of order statistics $d_{(i)} = x_{(i)} - \tilde{s}u_{P_i}$) for samples with (a) rectangular, (b) normal, (c) exponential, and (d) Laplace distributions, QC-EXPERT.

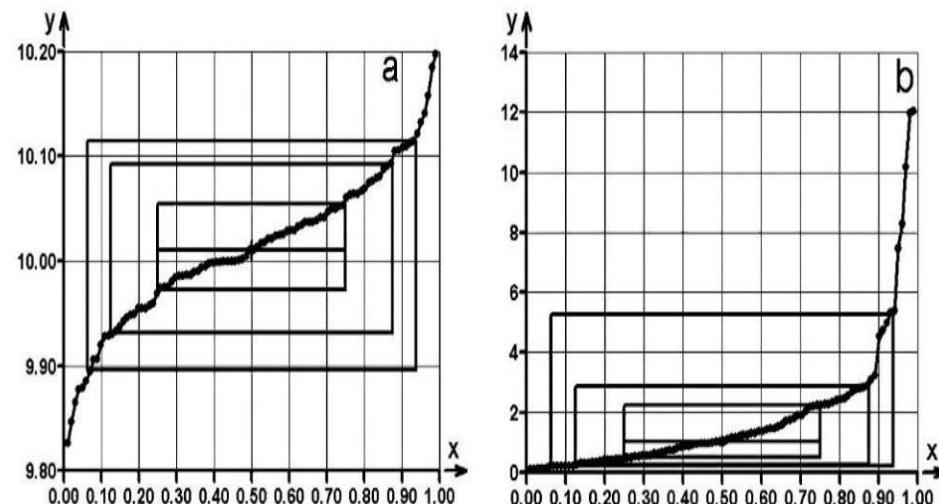
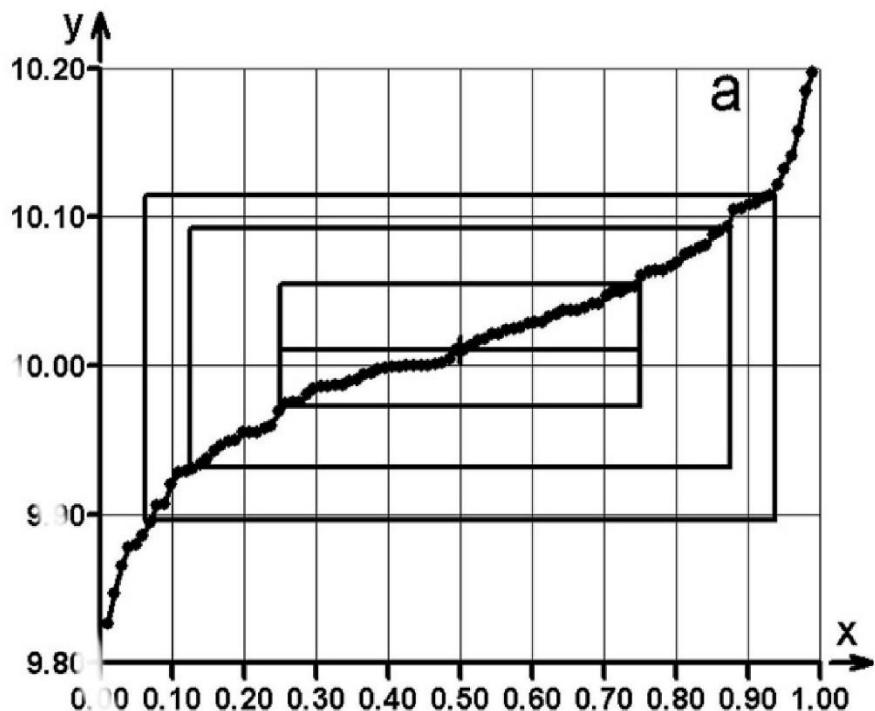


Figure 2.17 The quantile-box plot G10 (x-axis: the order probability P_i , y-axis: the order statistic $x_{(i)}$) for samples with (a) norm, symmetric (Gaussian, normal), and (b) log, asymmetric (log.-normal) distributions, QC-EXPERT.

Quantile-box Plot

Symmetric distribution



Asymmetric distribution

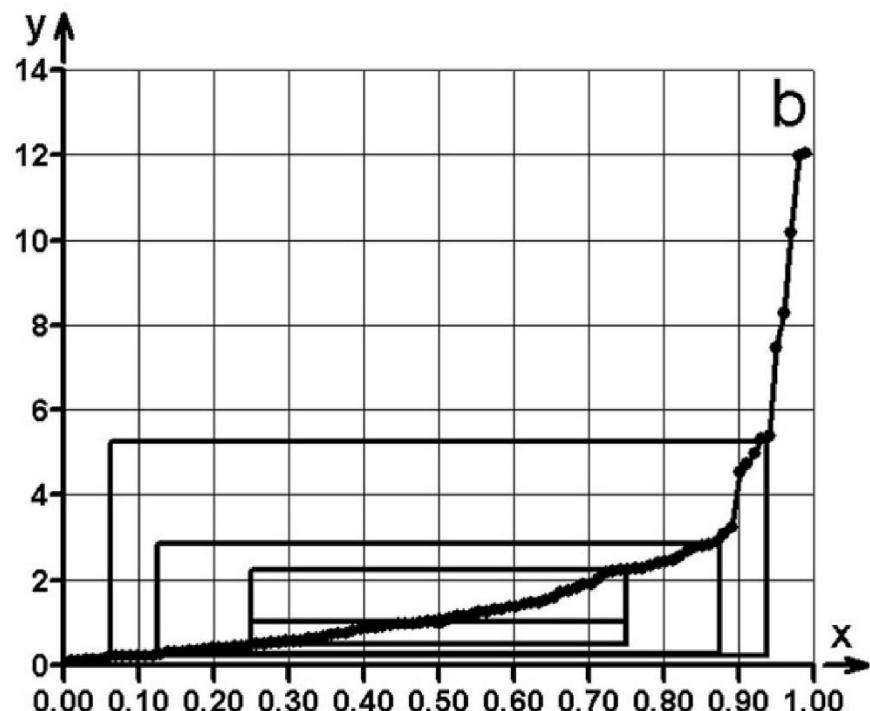
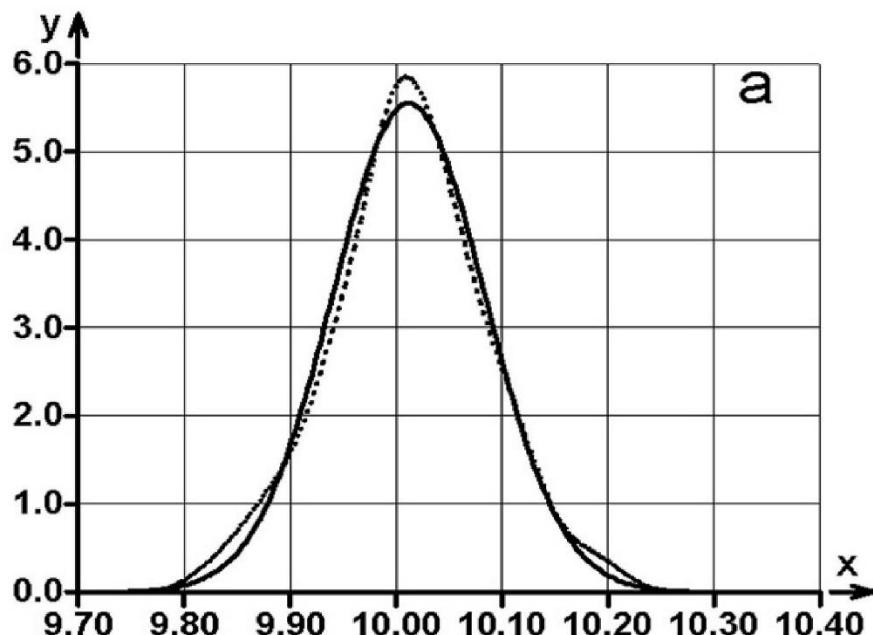


Figure 2.17 The quantile-box plot G10 (x-axis: the order probability P_i , y-axis: the order statistic $x_{(i)}$) for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, QC-EXPERT.

Kernel Estimation of Probability Density Plot

Symmetric distribution



Asymmetric distribution

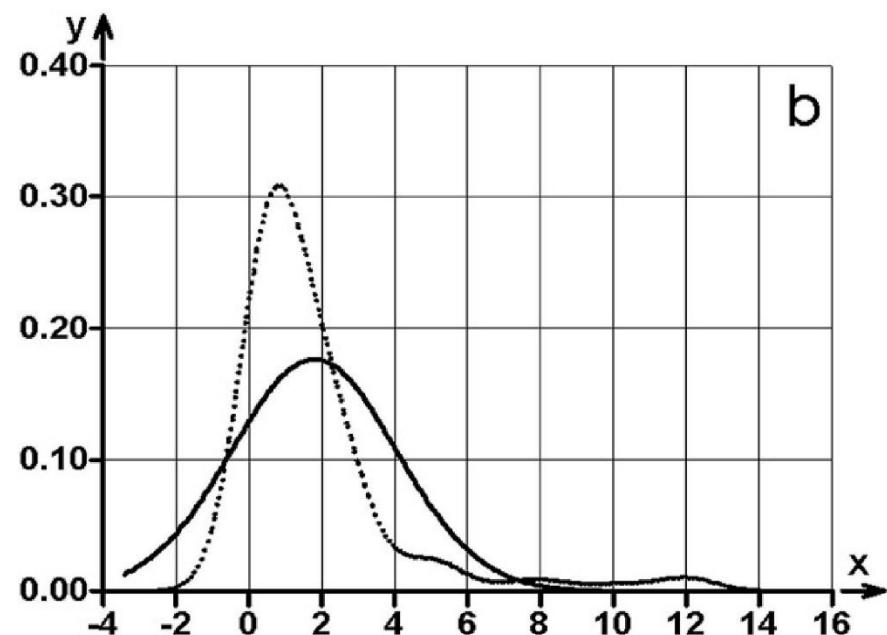
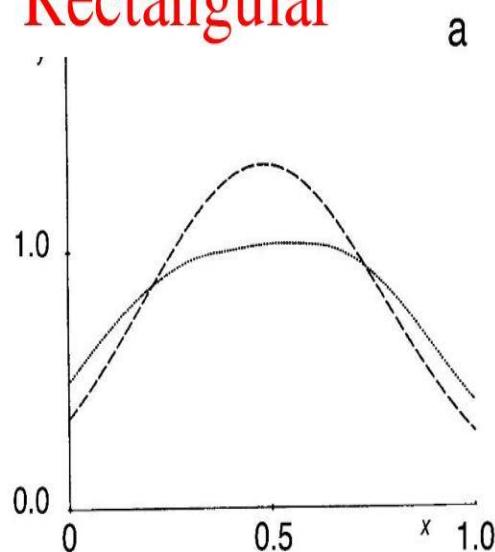


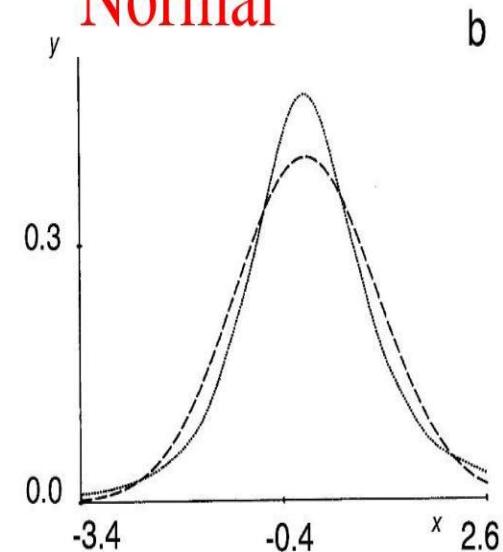
Figure 2.18 A Kernel estimation of probability density G12 (x -axis: the variable x ; y -axis: the probability density $f(x)$ (...) and the Gaussian (---) function) for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, QC-EXPERT.

Estimation of the Actual Distribution of laboratory data

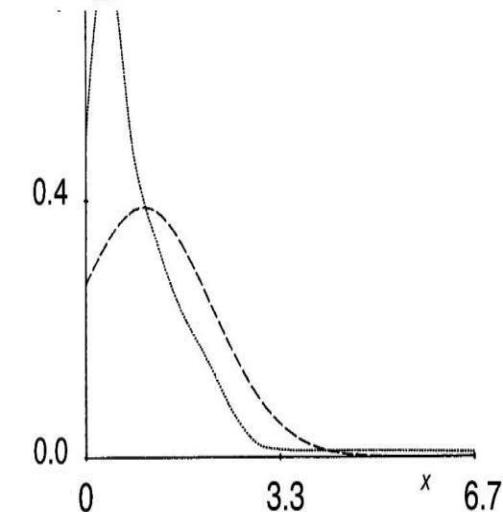
Rectangular



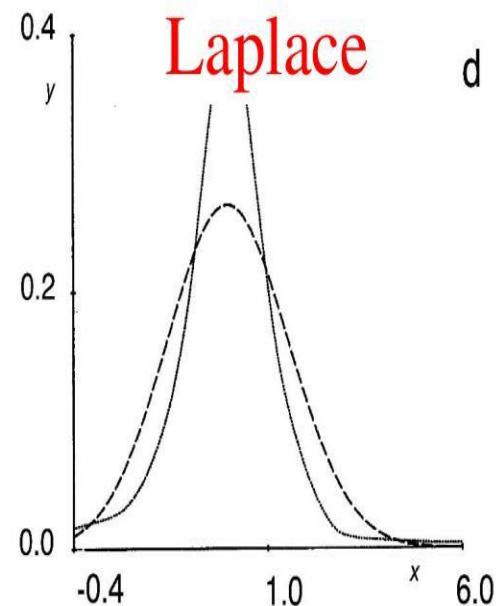
Normal



Exponential

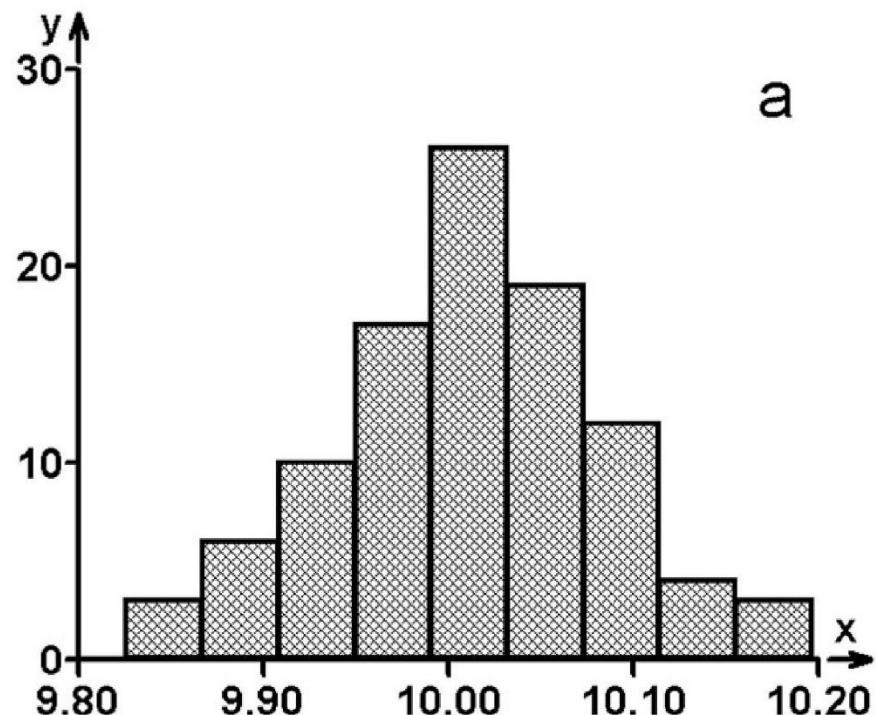


Laplace



Histogram

Symmetric distribution



Asymmetric distribution

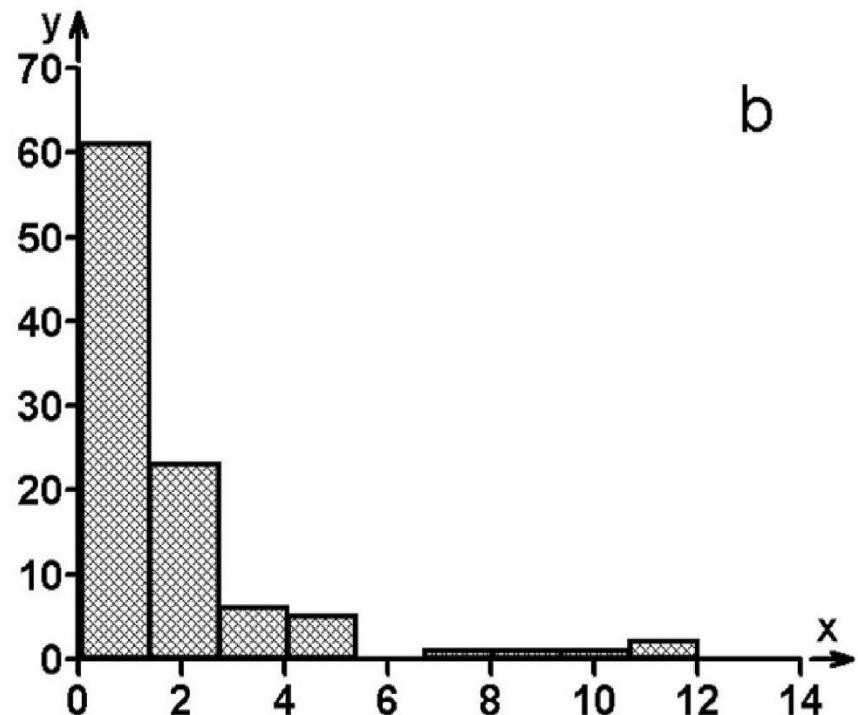


Figure 2.19 Histogram G13 (x-axis: the variable x ; y-axis: the probability density function $f(x)$) for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, QC-EXPERT.

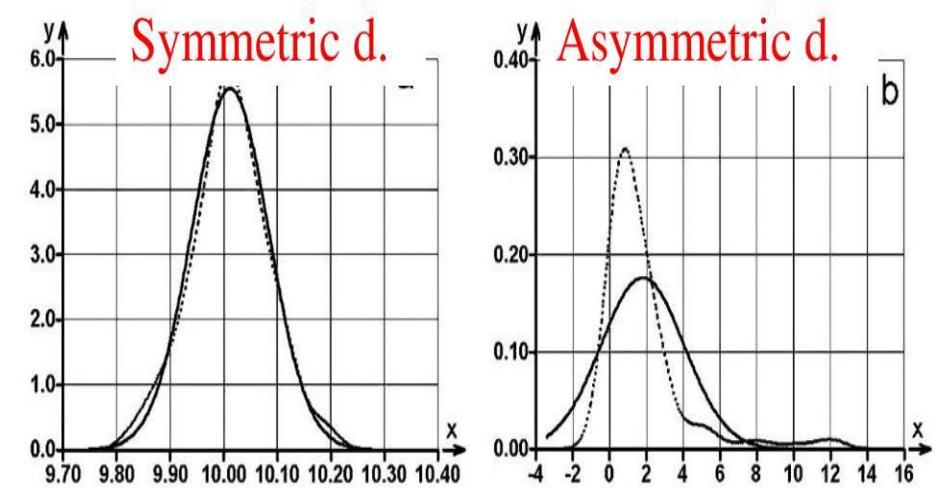
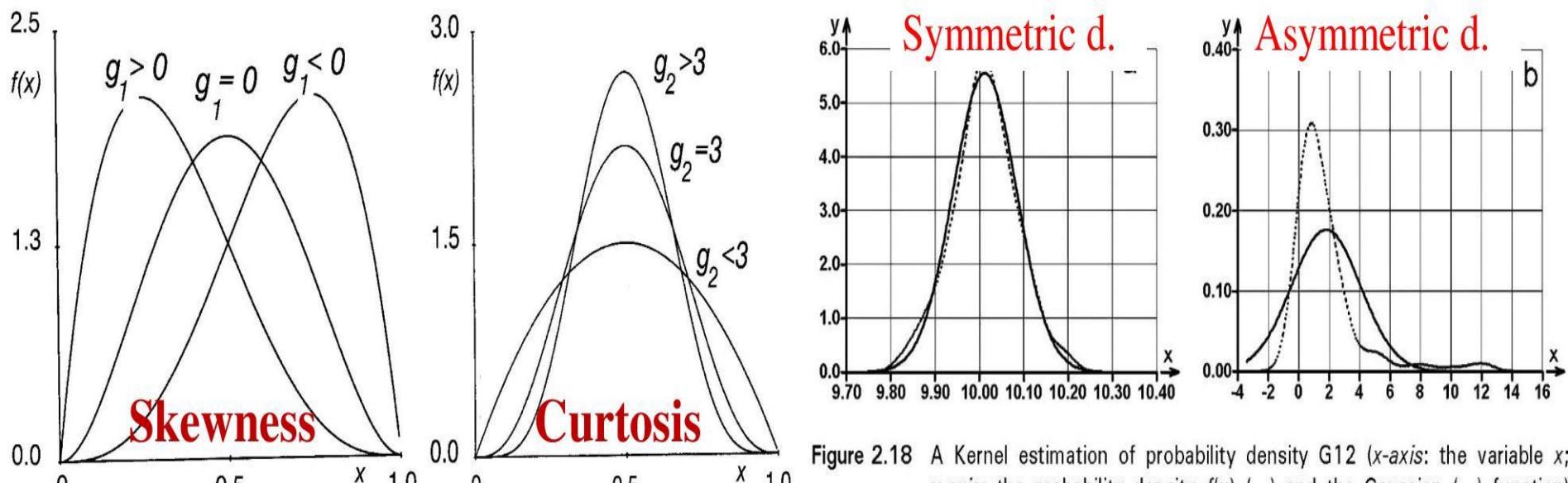


Figure 2.18 A Kernel estimation of probability density G12 (x-axis: the variable x ; y-axis: the probability density $f(x)$ (...) and the Gaussian (---) function) for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, QC-EXPERT.

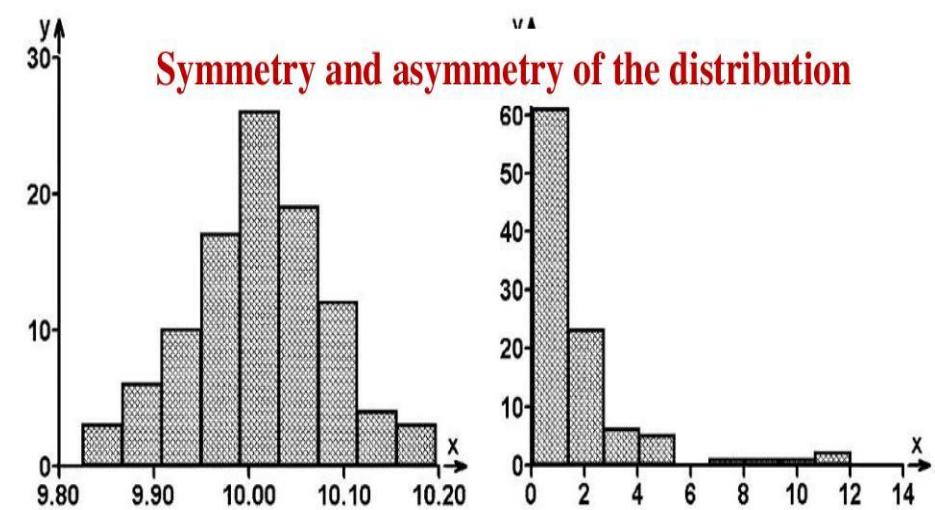
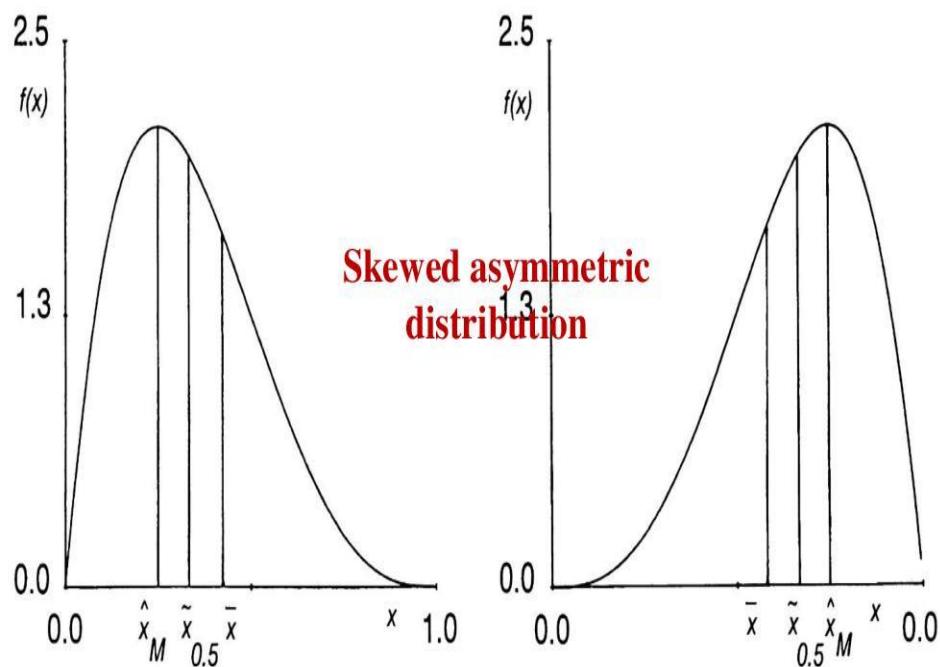
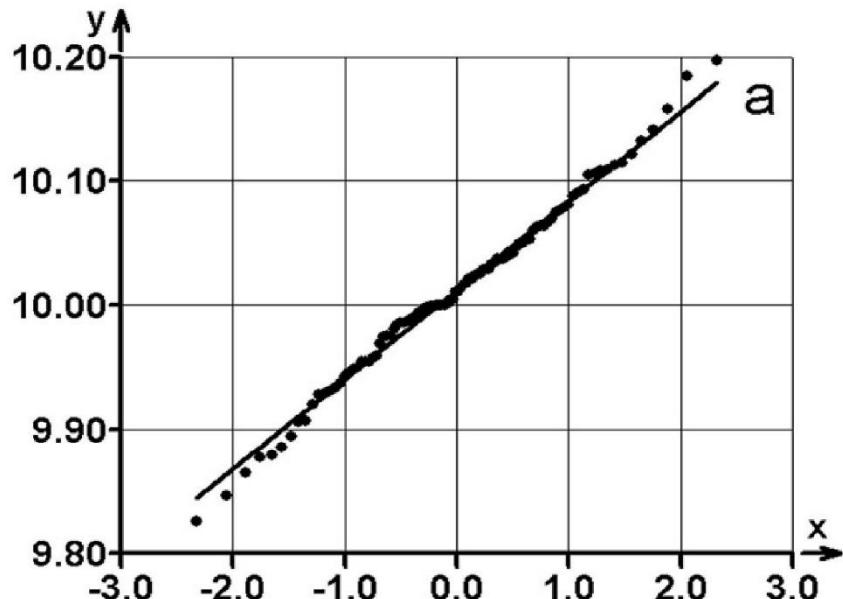


Figure 2.19 Histogram G13 (x-axis: the variable x ; y-axis: the probability density function $f(x)$) for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, QC-EXPERT.

Quantile-Quantile Plot (Rankit plot)

Symmetric distribution



Asymmetric distribution

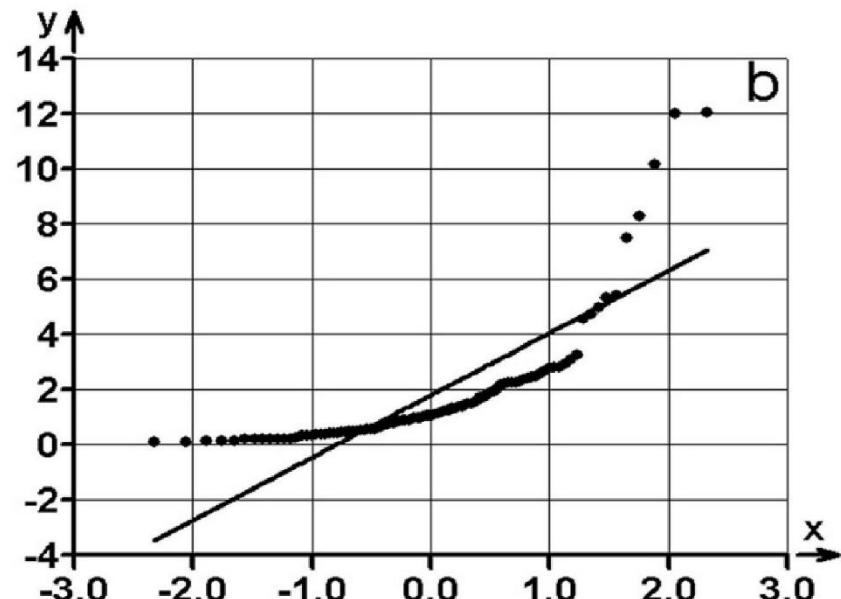
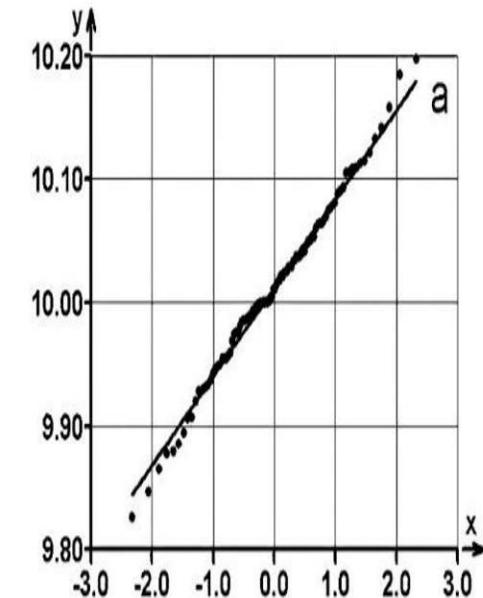


Figure 2.20 The rankit plot G15 (the normal probability plot, x -axis: the standardised normal quantile u_{P_i} , y -axis: the order statistic $x_{(i)}$) for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, QC-EXPERT.

Elucidation of Various Modifications of the Normal-Probability Plot (Q-Q plot)

Symmetric distribution
(Gaussian, Normal)



Asymmetric distribution
(Log-Normal)

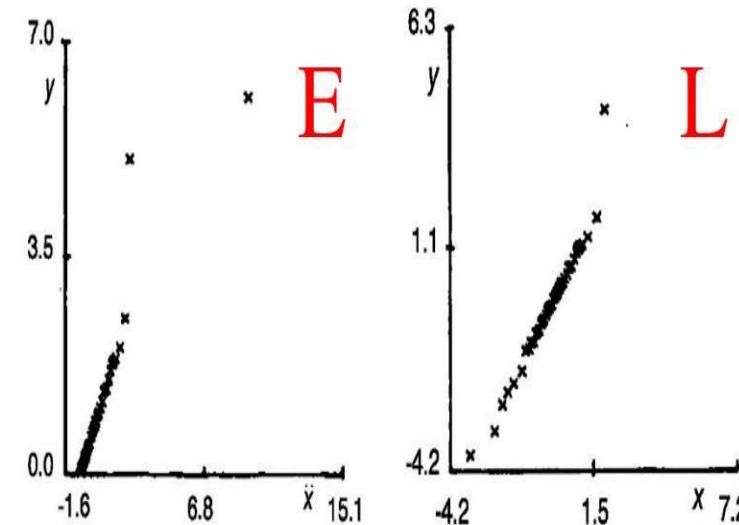
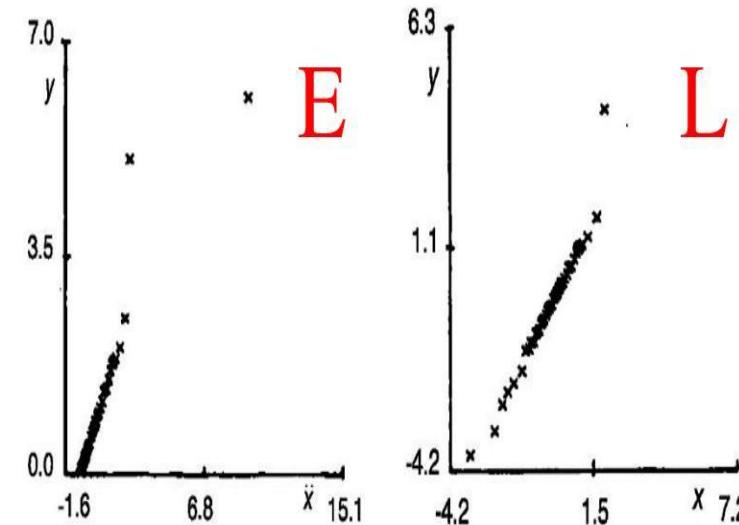
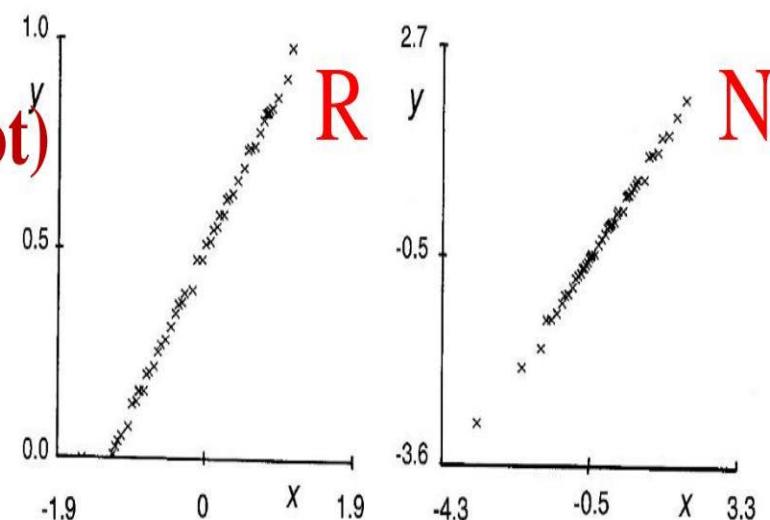
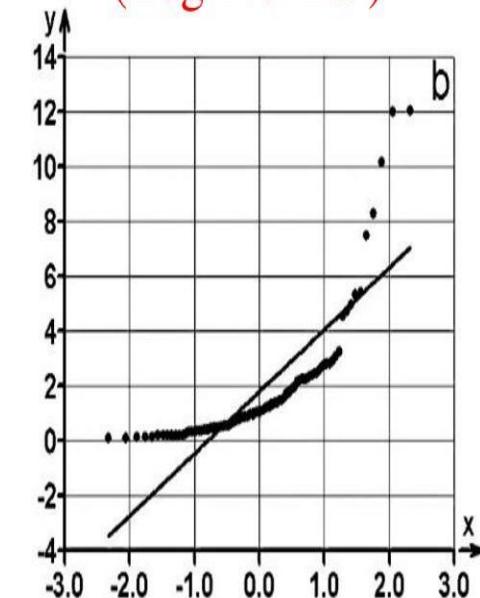


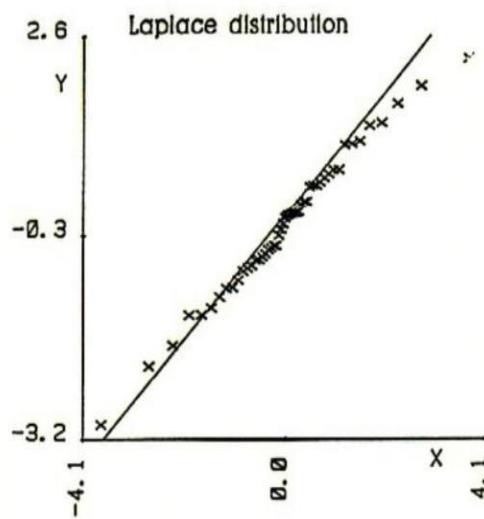
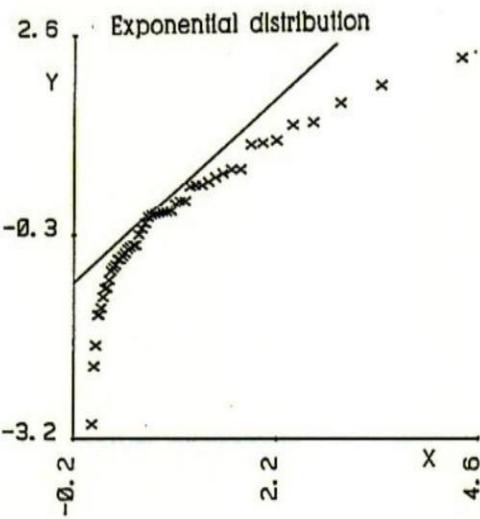
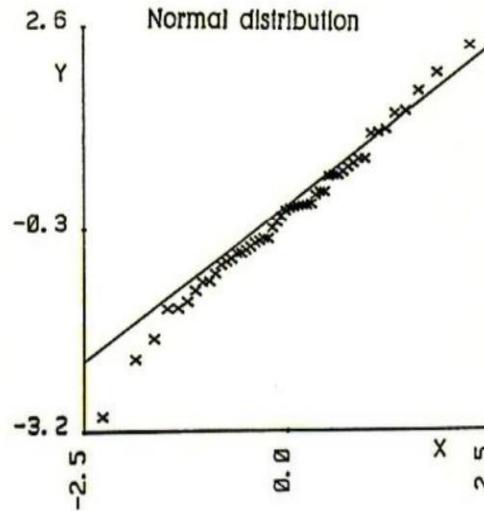
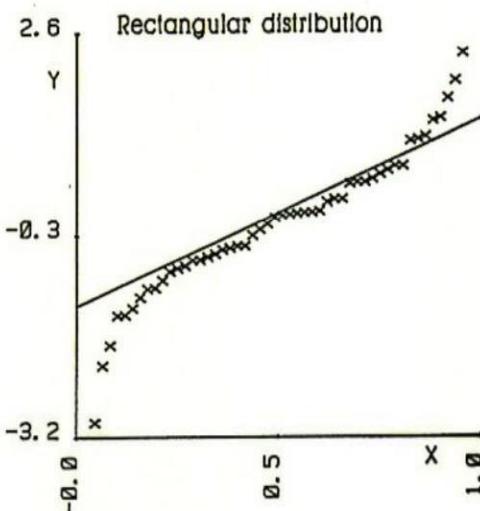
Figure 2.20 The rankit plot G15 (the normal probability plot, x-axis: the standardised normal quantile U_{P_i} , y-axis: the order statistic $x_{(i)}$) for samples with (a) norm, symmetric (Gaussian, normal), and (b) log, asymmetric (log.-normal) distributions, QC-EXPERT.

Figure 2.21 The conditioned rankit plot G16 (x-axis: the function $\Phi^{-1}[U_{(i-1)} + U_{(i+1)}]/2$; y-axis: the order statistic $x_{(i)}$) for samples with (a) rectangular, (b) normal, (c) exponential, and (d) Laplace distributions, QC-EXPERT.

Quantile-quantile plot (Q-Q plot)

(*x-axis*: the quantile $Q_S(P_i)$; *y-axis*: the order statistic $x_{(i)}$)

Quantile-Quantile (Q-Q) Plot



x-axis: the quantile $Q_S(P_i)$,

y-axis: the order statistic $x_{(i)}$

$$x_{(i)} = Q + R Q_S(P_i)$$

Diagnosis:

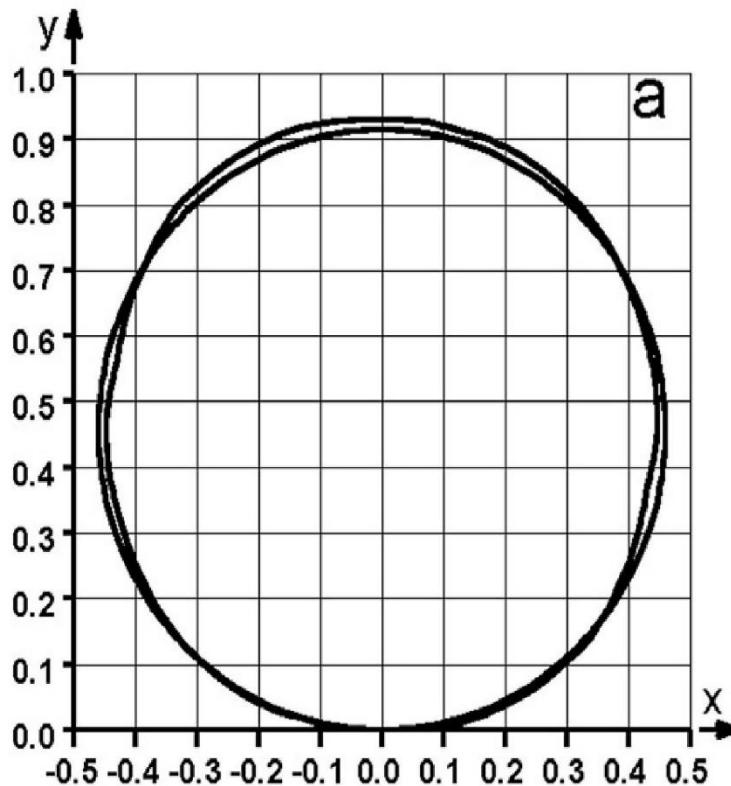
Closeness of the sample to the given theoretical one helps to indicate

an actual distribution.

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	$\Phi(s)$	$(2\pi)^{-1/2} \exp(-0.5s^2)$	$x_{(i)}$	$\Phi^{-1}(P_i)$
Laplace $x < Q$	$0.5 \exp(s)$	$0.5 \exp(s)$	$x_{(i)}$	$\ln(2P_i)$ for $P_i \leq 0.5$
Laplace $x > Q$	$0.5 [2 - \exp(-s)]$	$0.5 \exp(-s)$	$x_{(i)}$	$-\ln(2(1 - P_i))$ for $P_i > 0.5$
Log-normal	$\Phi[\ln(s)]$	$(2\pi)^{-1/2} \exp(-0.5 \ln s^2)$	$x_{(i)}$	$\exp(\Phi^{-1}(P_i))$

Circle Plot

Symmetric distribution



Asymmetric distribution

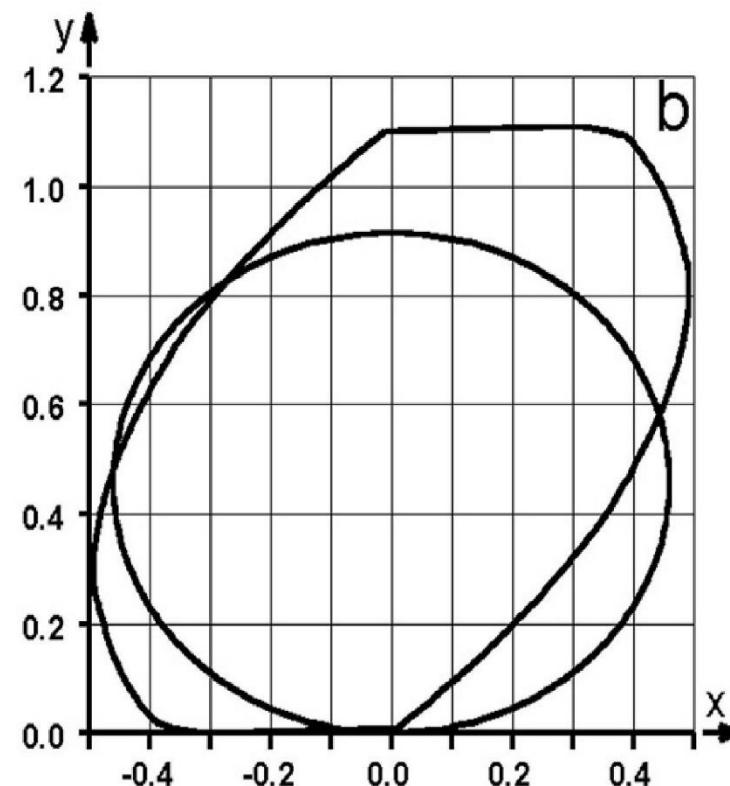


Figure 2.23 The circle plot for samples with (a) *norm*, symmetric (Gaussian, normal), and (b) *log*, asymmetric (log.-normal) distributions, QC-EXPERT.

2nd step:

CDA

Confirmatory Data Analysis:
parameters of location, spread and distribution

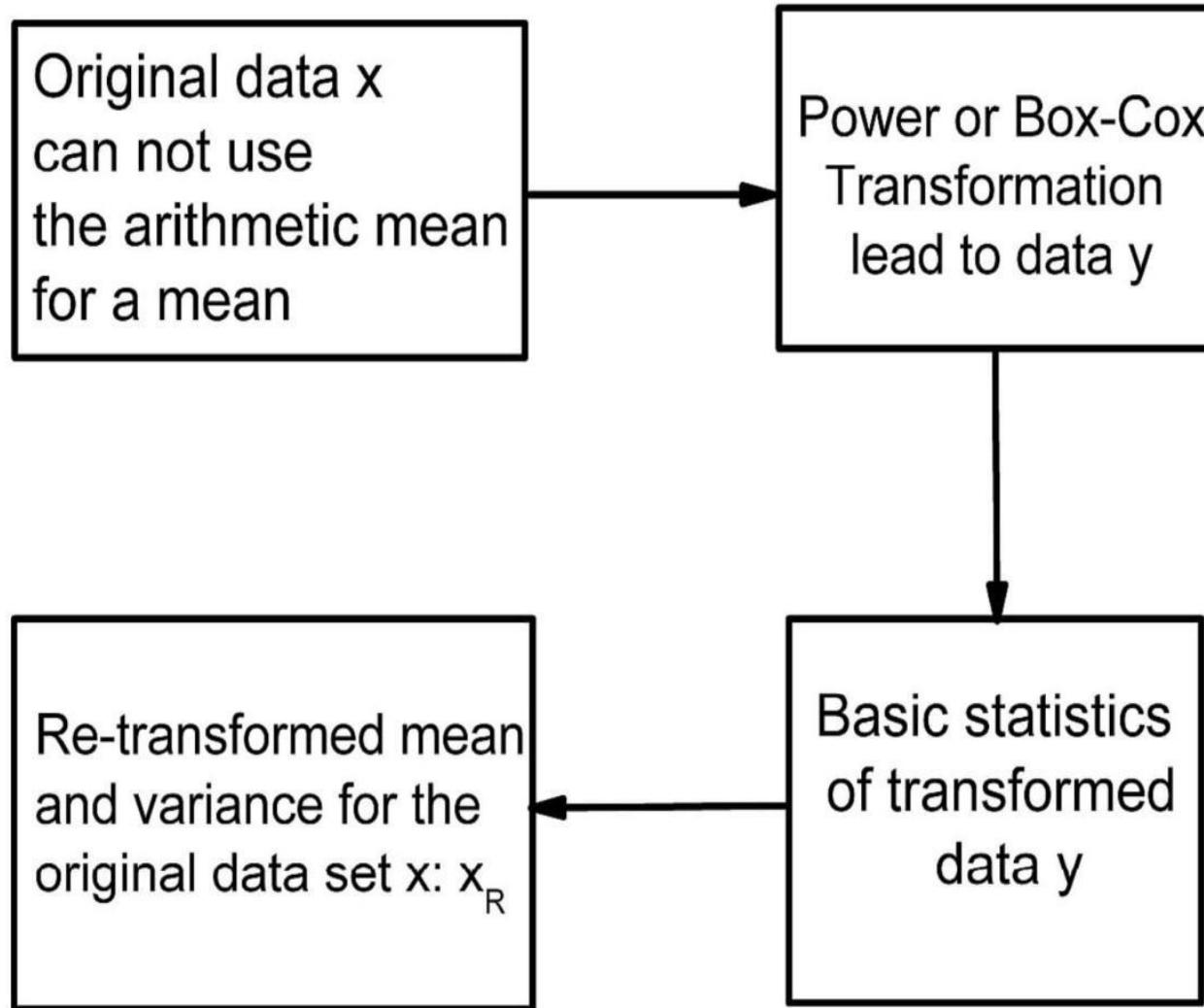
Two rigorous approaches to the parameter estimation

Data Transformation

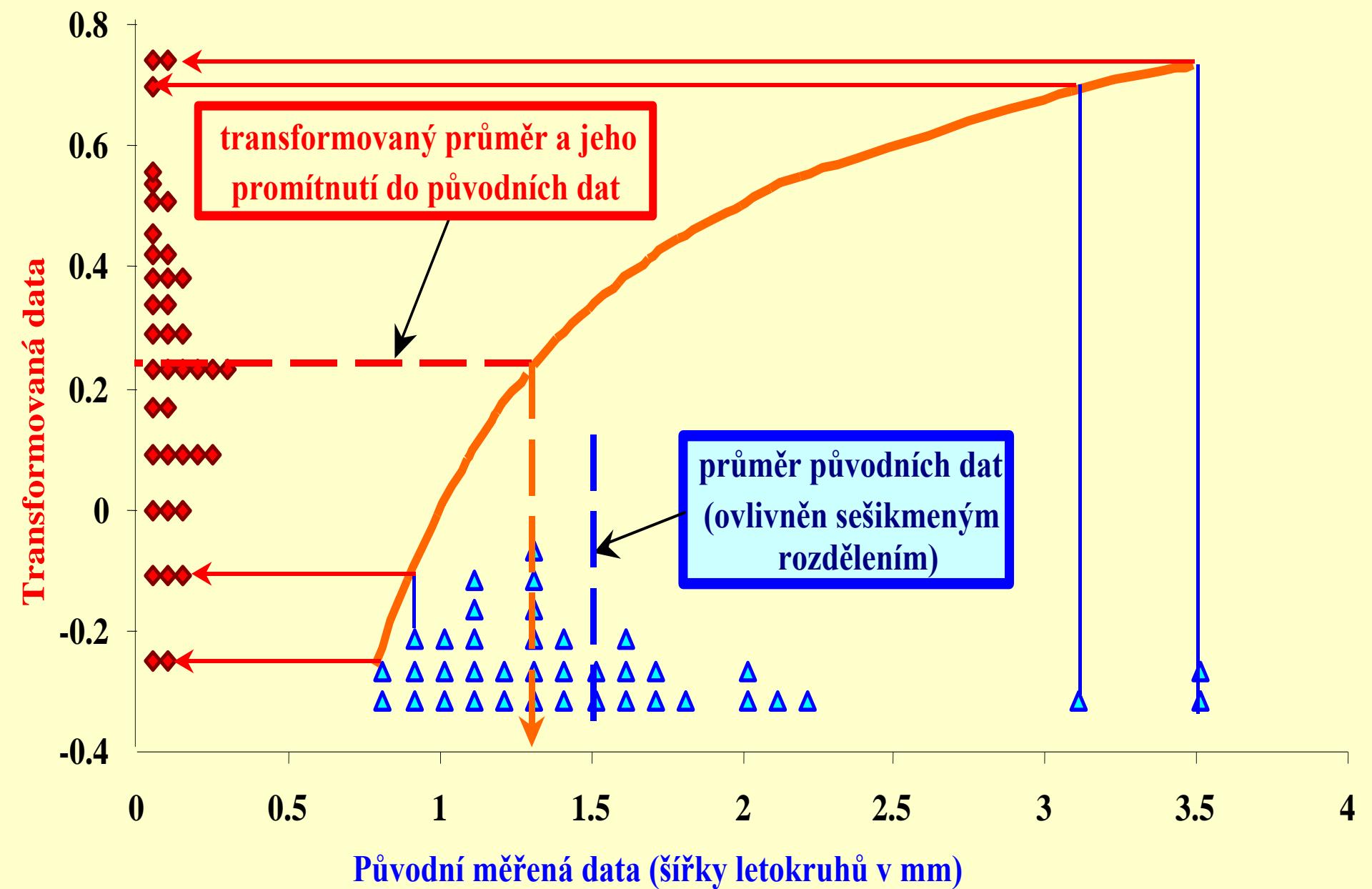
1. Power transformation for symmetry
2. Box-Cox transformation for normality

DATA TRANSFORMATION

Scheme of an application of power and Box-Cox data transformations.



DATA TRANSFORMATION



Transformation for symmetry is carried out by a simple power transformation

$$\begin{aligned}x^\lambda & \quad \text{for parameter } \lambda > 0 \\y = g(x) = \ln x & \quad \text{for parameter } \lambda = 0 \\-x^{-\lambda} & \quad \text{for parameter } \lambda < 0.\end{aligned}$$

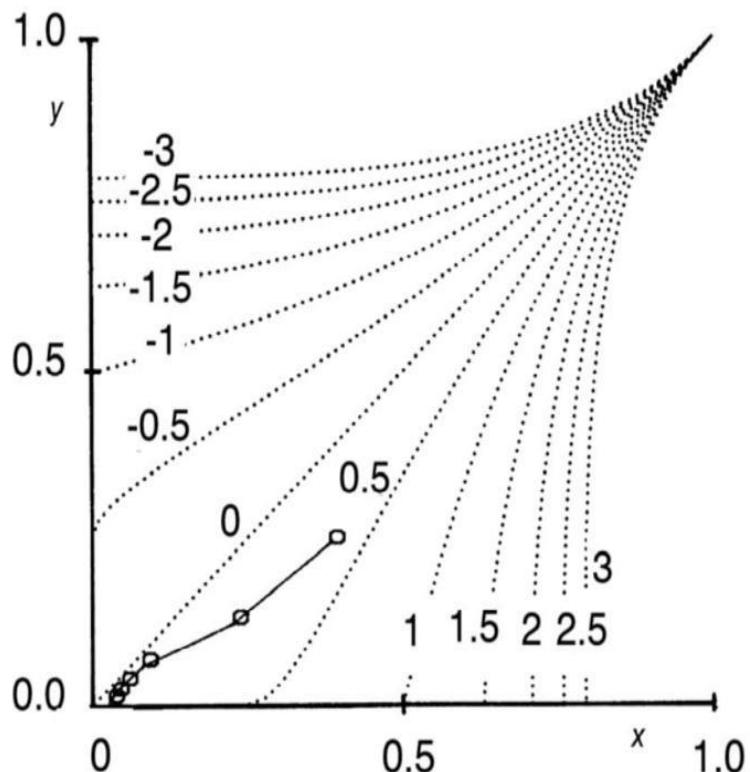
Box-Cox family of transformations defined as

$$y = g(x) = \begin{cases} (x^\lambda - 1) / \lambda & \text{for } \lambda \neq 0 \\ \ln x & \text{for } \lambda = 0 \end{cases}$$

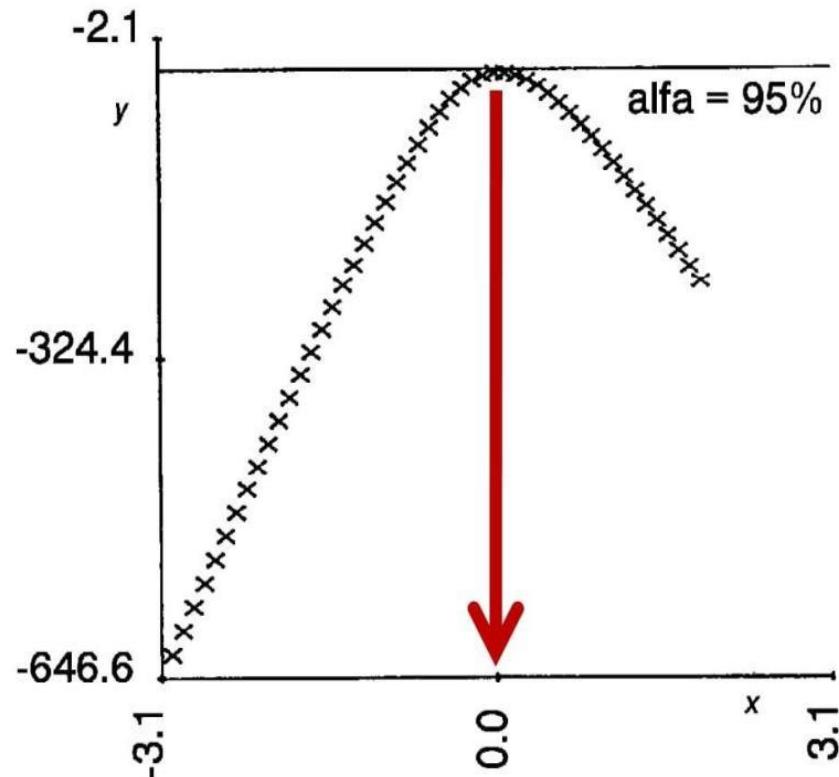
In many sample distributions can be transformed to approximate normality by use of Box-Cox family of transformations.

Hines-Hines selection plot and Plot of the logarithm of the likelihood function

(*x-axis*: the parameter λ ; *y-axis*: the logarithm of the likelihood function $\ln L$)



Determination of λ from a Hines-Hines selection graph (*x-axis*: the ratio , *y-axis*: the ratio).



Plot of the logarithm of maximum likelihood (*x-axis*: the parameter λ ; *y-axis*: the logarithm of the likelihood function $\ln L$).

Rough re-expressions of the mean

$$\bar{x}_R$$

(1) *Rough re-expressions* represent a single reverse transformation $\bar{x}_R = g^{-1}(y)$. This re-expression for a simple power transformation leads to the general mean

$$\bar{x}_R = \bar{x}_\lambda = \left[\frac{1}{n} \sum_{i=1}^n x_i^\lambda \right]^{1/\lambda} \quad (2.36)$$

where for $\lambda = 0$, $\ln x$ is used instead of x^λ and e^x instead of $x^{1/\lambda}$. The re-expressed mean $\bar{x}_R = \bar{x}_{-1}$ stands for the *harmonic mean*, $\bar{x}_R = \bar{x}_0$ for the *geometric mean*, $\bar{x}_R = \bar{x}_1$ for the *arithmetic mean* and $\bar{x}_R = \bar{x}_2$ for the *quadratic mean*.

Rigorous re-expression of the mean and the variance

$$\bar{\bar{x}}_R$$

(2) For $\lambda \neq 0$ and the Box-Cox transformation, Eq. (2.31), the re-expressed mean \bar{x}_R will be represented by one of the two roots of the quadratic equation

$$\bar{x}_{R,1,2} = [0.5(1 + \lambda \bar{y}) \pm 0.5 \left\{ 1 + 2\lambda(\bar{y} + s^2(y)) + \lambda^2(\bar{y}^2 - 2s^2(y)) \right\}^{1/2}]^{1/\lambda} \quad (2.42)$$

which is close to the median $\tilde{x}_{0.5} = g^{-1}(\tilde{y}_{0.5})$. If \bar{x}_R is known, the corresponding variance may be calculated from

$$s^2(x) = \bar{x}^{(-2\lambda+2)} s^2(y) \quad (2.43)$$

Schema mocninné a Box-Coxovy transformace

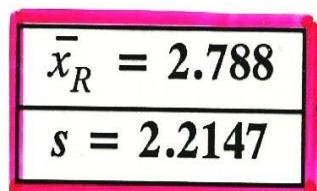
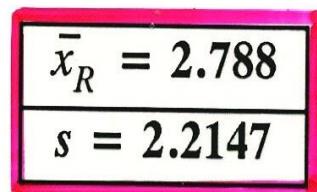
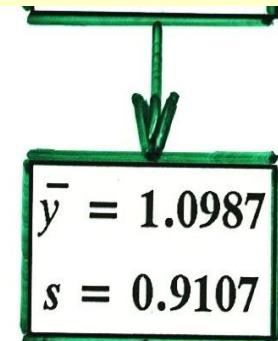
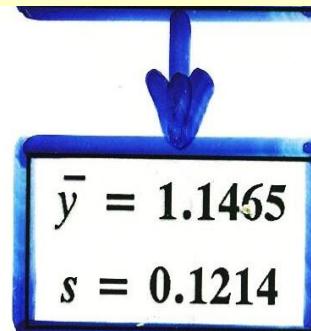
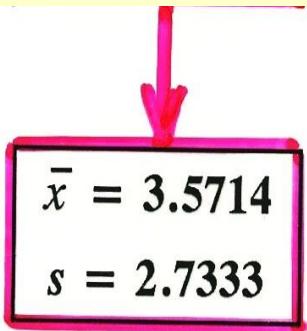
x	y (mocninná)	y (Box-Cox)
0.5	0.9117	-0.6621
0.9	0.9861	-1.0462
1.0	1.0000	0.0000
1.0	1.0000	0.0000
1.1	1.0128	0.0959
1.3	1.0356	0.2670
1.5	1.0556	0.4166
1.5	1.0556	0.4166
1.5	1.0556	0.4166
1.8	1.0815	0.6114
1.9	1.0893	0.6701
2.0	1.0968	0.7262
2.0	1.0968	0.7262
2.7	1.1416	1.0620
3.2	1.1678	1.2582
3.2	1.1678	1.2582
3.3	1.1726	1.2942
3.3	1.1726	1.2942
3.6	1.1862	1.3968
5.2	1.2459	1.8439
5.5	1.2552	1.9140
5.5	1.2552	1.9140
6.0	1.2699	2.0239
6.0	1.2699	2.0239
7.0	1.2962	2.2217
8.0	1.3195	2.3963
8.0	1.3195	2.3963
11.5	1.3849	2.8869

Průměr

Směrodatná odchylka

Retransform. průměr (mocnina)

Retransform. průměr (Box-Cox)



$$\bar{X}_R = ?$$

Problem 2.26 EDA in determination of trace copper in kaolin

Task: Trace copper was determined in a standard sample of kaolin, and the values were arranged in increasing order. Examine the type of sample distribution and decide what type of measures of location and spread should be used.

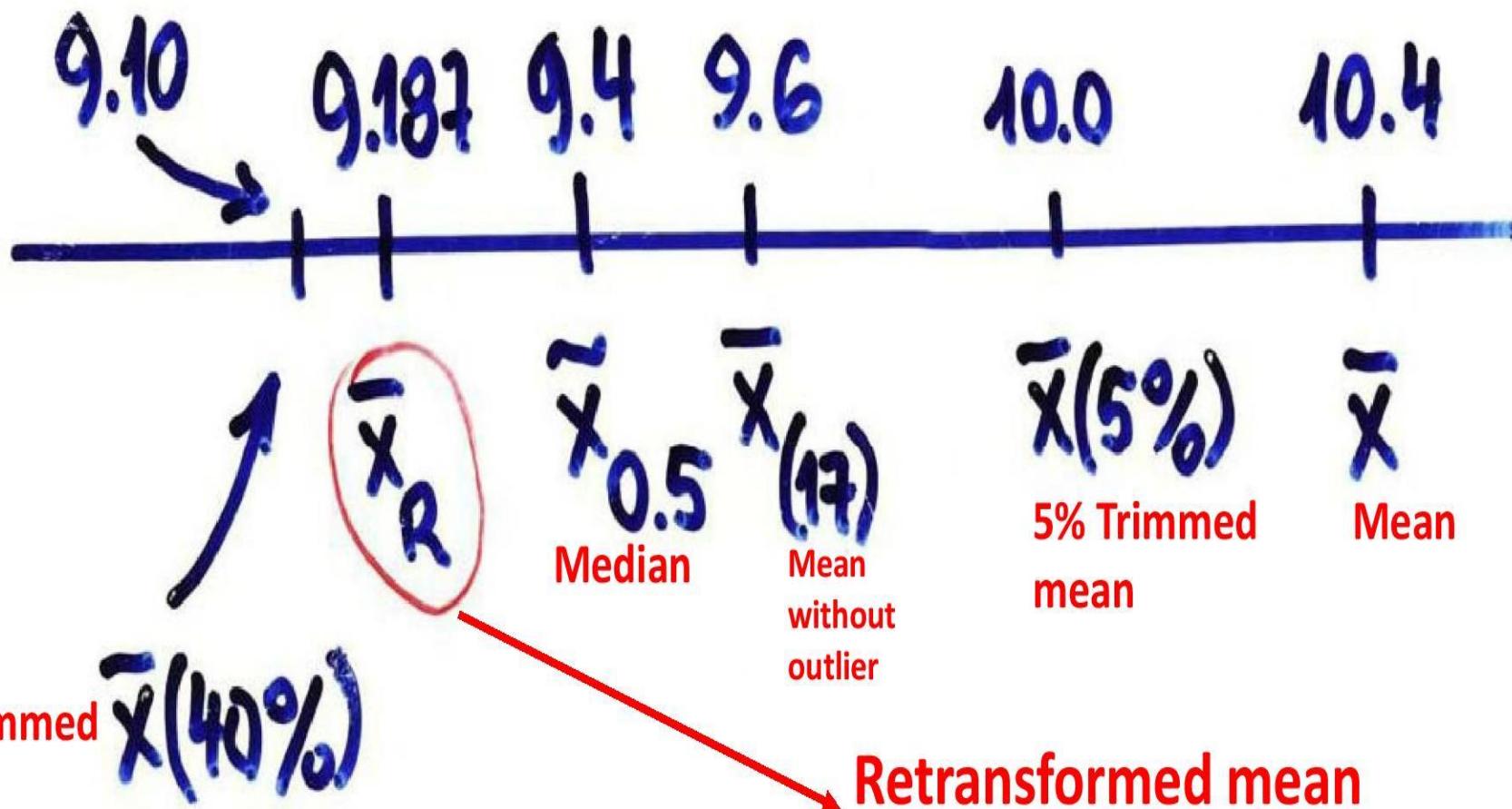
Data: copper concentration [ppm]; $n = 17$,

4, 5, 7, 7, 7, 8, 8.3, 8.4, 9.4, 9.5, 10, 10.5, 12, 12.8, 13, 22, 23.

Program: CHEMSTAT: Basic statistics: Exploratory data analysis.

Conclusion: $\bar{x}_R = ?$

The best estimate of the mean value



Parameters of small samples ($4 \leq n \leq 20$) with

Horn's Procedure

Pivot half-sum (= point and interval estimates of location)

Pivot range (= spread)

Horn's procedure for $4 \leq n \leq 20$

Procedure based on order statistics.

- 1) Write the table of **order statistics**.
- 2) The **pivot depth** is expressed by $H_L = \text{int}[(n + l)/2]/2$ or $H_L = \text{int}[(n + l)/2 + l]/2$ according to which of the H_L is an integer.
- 3) The **lower pivot** is $x_L = x_{(H)}$ and the **upper** one is $x_U = x_{(n+1-H)}$.
- 4) The estimate of the parameter of location is then expressed by the **pivot half sum**

$$P_L = 0.5(x_L + x_U)$$

- 5) The estimate of the parameter of spread is expressed by the
pivot range

$$R_L = x_U - x_L$$

- 6) The **random variable**

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

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

- 7) The **95% confidence interval** of the mean is expressed by pivot statistics as

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

Table :The quantile $t_{L,1-\alpha}(n)$ of the Horn-distribution

$1 - \alpha$ N	0.9	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	0.776
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

Exercise B3.01 Estimate of median value of haptoglobin in human blood serum (Horn)

The concentration of haptoglobin in human blood serum was measured in eight adult individuals. Calculate estimates for median value, parameter of variance, and 95% interval of reliability of median value. Examine whether this sample comes from a logarithmic-normal distribution. Also apply Horn's procedure (pg. 51 in [14]).

Data: Concentration of haptoglobin [g. l⁻¹] in human blood serum:
1.82, 3.32, 1.07, 1.27, 0.49, 3.79, 0.15, 1.98.

1. Order statistics:

i	1	2	3	4	5	6	7	8
x _(i)	0.15	0.49	1.07	1.27	1.82	1.98	3.32	3.79

2. Depth of pivot:

$$n = 8,$$

$$H = \text{integer} \frac{\frac{n+1}{2} + 1}{2} = \text{int}(2.75) \approx 2$$

3. Lower and upper pivot:

$$x_D = x_{(H)} = \\ x_H = x_{(n+1-H)} =$$

$$x_{(2)} = 0.49 \\ x_{(7)} = 3.32$$

4. $P_L = \frac{x_D + x_H}{2} = 1.905$

5. $R_L = x_H - x_D = 3.32 - 0.49 = 2.83$

6. 95% $t_{L, 1-\alpha/2} = 0.564$

$$P_L - R_L t_{L, 1-\alpha/2}(n) \leq \mu \leq P_L + R_L t_{L, 1-\alpha/2}(n)$$

$$1.905 - 2.83 \times 0.564 \leq \mu \leq 1.905 + 2.83 \times 0.564$$

$$0.31 \leq \mu \leq 3.50$$

Horn

Analýza malého výběru

Na vzorové úloze **B3.01 Střední hodnota haptoglobinu v lidském krevním séru** ukážeme Hornův postup analýzy malých výběrů.

Data: Koncentrace haptoglobinu [g l⁻¹]: 1.82 3.32 1.07 1.27 0.49 3.79 0.15 1.98

Řešení: Hornův postup pivotů pro malé výběry (4 < n < 20):

1. Pořadkové statistiky:

i	1	2	3	4	5	6	7	8
x _(i)	0.15	0.49	1.07	1.27	1.82	1.98	3.32	3.79

2. Hloubka pivotu:

n = 8, sudé

$$H = \text{int} \frac{\frac{m+1}{2} + 1}{2}$$

int(2.75) ≈ 2

3. Pivoty: Dolní pivot x_D = x_(H)

x₍₂₎ = 0.49

Horní pivot x_H = x_(n+1-H)

x₍₇₎ = 3.32

4. Pivotová polosuma P_L = $\frac{x_D + x_H}{2}$

= 1.905

5. Pivotové rozpětí $R_L = x_H - x_D$ $3.32 - 0.49 = 2.83$

6. 95%ní interval spolehlivosti střední hodnoty μ : $t_{L, 1-\alpha/2} = 0.564$

$$P_L - R_L t_{L,1-\alpha/2}(n) \leq \mu \leq P_L + R_L t_{L,1-\alpha/2}(n)$$

$$1.905 - 2.83 \times 0.564 \leq \mu \leq 1.905 + 2.83 \times 0.564$$

$$0.31 \leq \mu \leq 3.50$$

7. Závěr: Bodový odhad míry polohy je **1.91 g/l**, míry rozptylení **2.83** a intervalový odhad míry polohy je **$0.31 \text{ g/l} \leq \mu \leq 3.50 \text{ g/l}$** .

Hornův postup u malých výběrů

Analýza malých výběrů

- Závěry jsou vždy zatíženy značnou mírou nejistoty.
- Malých rozsahů jen tam, kde není možné zvýšit počet.

n = 2: 100(1 - α)%ní konfidenční interval střední hodnoty

$$\frac{x_1 + x_2}{2} - T_{\alpha} \frac{|x_1 - x_2|}{2} \leq \mu \leq \frac{x_1 + x_2}{2} + T_{\alpha} \frac{|x_1 - x_2|}{2}$$

- pro normální rozdělení $T_{\alpha} = \text{cotg}(\alpha \pi / 2)$, $T_{0.05} = 12.71$,
- pro rovnoměrné rozdělení $T_{\alpha} = 1/\alpha - 1$, tj. $T_{0.05} = 19$.

n = 3: 100(1 - α)%ní konfidenční interval střední hodnoty

$$\bar{x} - T'_{\alpha} \frac{s}{\sqrt{3}} \leq \mu \leq \bar{x} + T'_{\alpha} \frac{s}{\sqrt{3}}$$

- pro normální rozdělení je $T'_{\alpha} \approx 1/\sqrt{\alpha} - 3\sqrt{\alpha}/4 \dots, T'_{0.05} = 4.30$.
- pro rovnoměrné rozdělení je $T'_{0.05} = 5.74$,

odhadem parametru rozptylení pivotové rozpětí

$$R_L = x_H - x_D$$

Náhodná veličina k testování

$$T_L = \frac{P_L}{R_L} = \frac{x_D + x_H}{2(x_H - x_D)}$$

má přibližně symetrické rozdělení, jehož vybrané kvantily jsou v tabulce.

95%ní interval spolehlivosti střední hodnoty se vypočte

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

4 ≤ n ≤ 20, (Hornův postup):

je založený na pořádkových statistikách.

Hloubka pivotu je $H = (\text{int}((n+1)/2))/2$
nebo $H = (\text{int}((n+1)/2) + 1)/2$,

Dolní pivot je $x_D = x_{(H)}$ a horní pivot $x_H = x_{(n+1-H)}$.

Odhadem parametru polohy je pivotová polosuma

$$P_L = \frac{x_D + x_H}{2}$$

Tabulka 3.11 Kvantity $t_{L,0.975}(n)$ rozdělení T_L

n	1 - α	0.9	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	0.776
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

Úloha C3.11 Test správnosti koncentrace tenzidů (Horn)

Standardní vzorek obsahuje 2.5 mg/l anionaktivních tenzidů. Aplikujte i Hornův postup. Testujte, zda výsledky koncentrace standardu jsou správné. Jde o symetrické rozdělení?

Data: Koncentrace tenzidů [mg/l]: 2.36 2.40 2.48 2.50 2.57 2.62 2.68

[Výsledky: Gauss. rozd., $\bar{x} = 2.52$, $\bar{x}_R = 2.51$, $\tilde{x}_{0.5} = 2.50$, $s = 0.12$,
 $\hat{g}_1 = 0.04$, $\hat{g}_2 = 1.78$, $2.41 < \bar{x} < 2.62$]

Hornův postup:

1. Pořádkové statistiky:

i	1	2	3	4	5	6	7
$x_{(1)}$	2.36	2.40	2.48	2.50	2.57	2.62	2.68

2. Hloubka pivotu: $n = 7$, liché

$$H = \text{integer} \frac{\frac{n+1}{2}}{2} = \text{int}(2.0) \approx 2$$

3. Pivoty:

$$\begin{array}{ll} \text{Dolní pivot } x_D = x_{(H)} = & x_{(2)} = 2.40 \\ \text{Horní pivot } x_H = x_{(n+1-H)} = & x_{(6)} = 2.62 \end{array}$$

4. Pivotová polosúma $P_L = \frac{x_D + x_H}{2} =$ **= 2.51**

5. Pivotové rozpětí $R_L = x_H - x_D =$ **$2.62 - 2.40 = 0.22$**

6. 95%ní interval spolehlivosti střední hodnoty μ : $t_{L, 1-\alpha/2}(7) = 0.720$

$$P_L - R_L t_{L, 1-\alpha/2}(n) \leq \mu \leq P_L + R_L t_{L, 1-\alpha/2}(n)$$

$$2.51 - 0.22 \times 0.72 \leq \mu \leq 2.51 + 0.22 \times 0.72$$

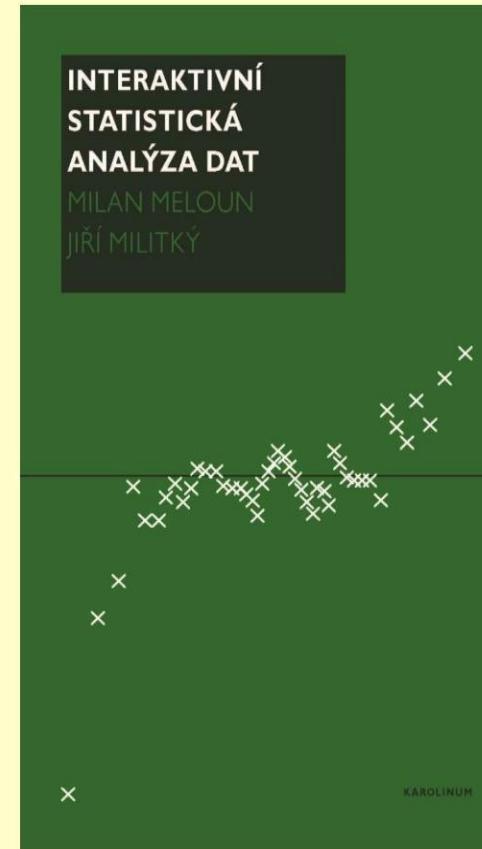
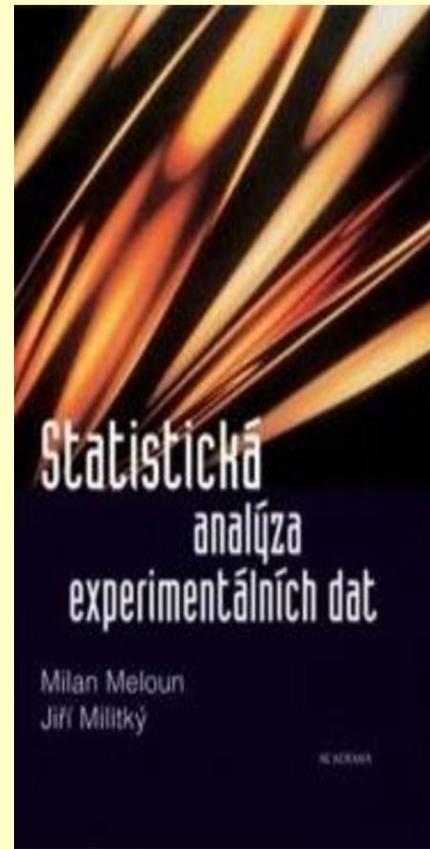
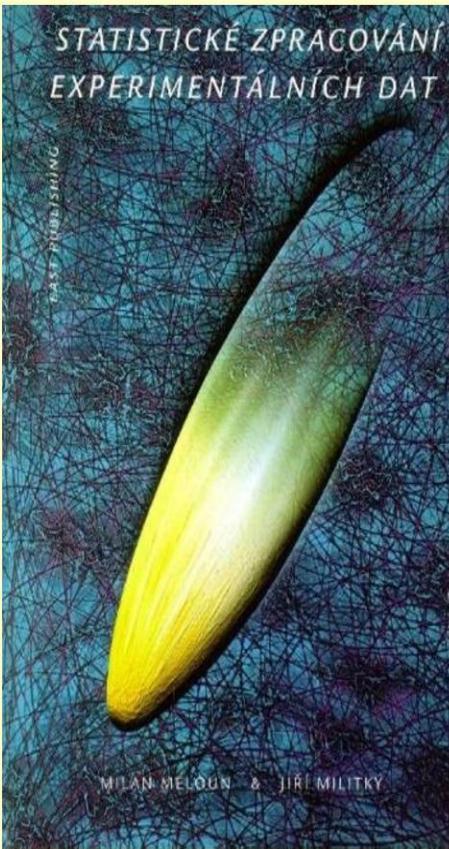
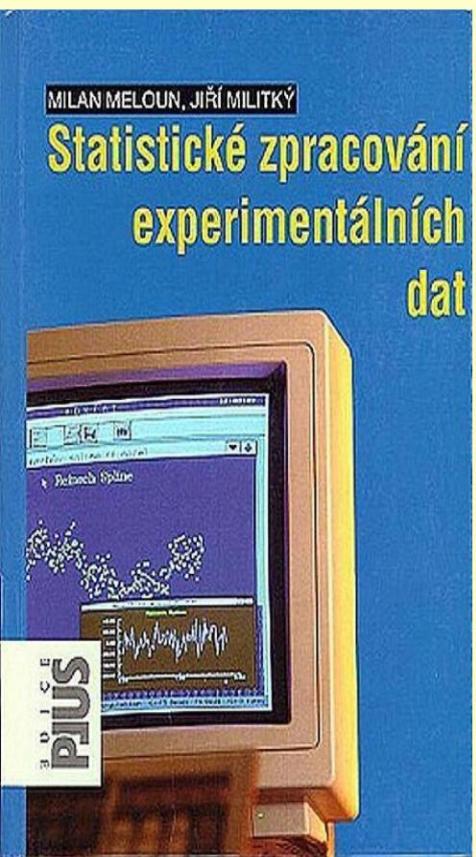
$$\mathbf{2.35 \leq \mu \leq 2.67}$$

Jan Amos Komenský: **Velká didaktika** (1592 – 1670)

“Knihy, jako největší přátelé, rády s námi upřímně, jasně a bez přetvářky hovoří, poučují nás, dávají nám návody, povzbuzují nás a předvádějí i věci našemu zraku velmi vzdálené.“

Our recommended textbooks in czech

Doporučené naše učebnice



1. vydání 1994

2. vydání 1998

3. vydání 2004

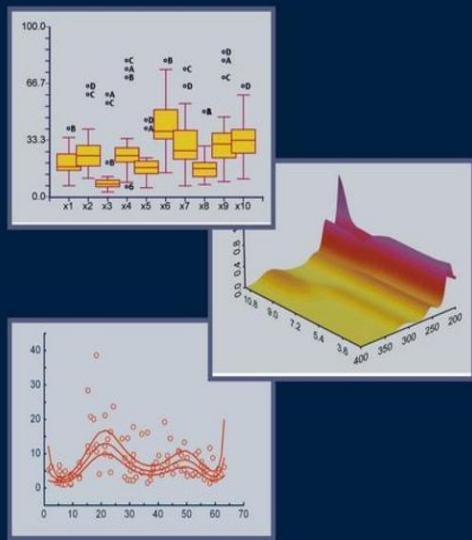
4. vydání 2012

Celostátní učebnice Statistické analýzy jednorozměrných dat

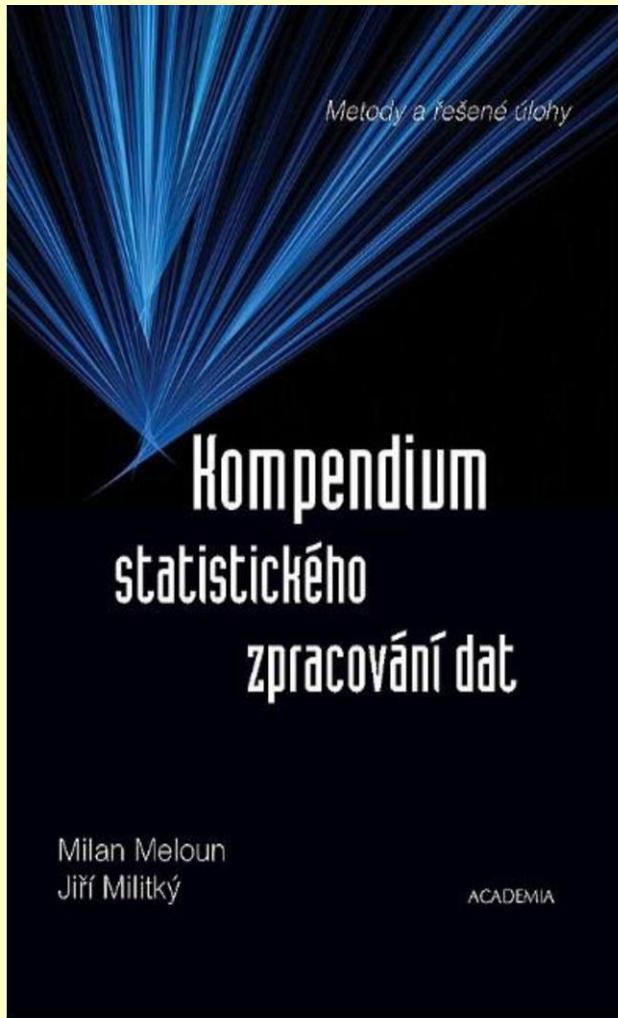
MILAN MELOUN • JIŘÍ MILITKÝ • ACADEMIA

KOMPENDIUM STATISTICKÉHO ZPRACOVÁNÍ DAT

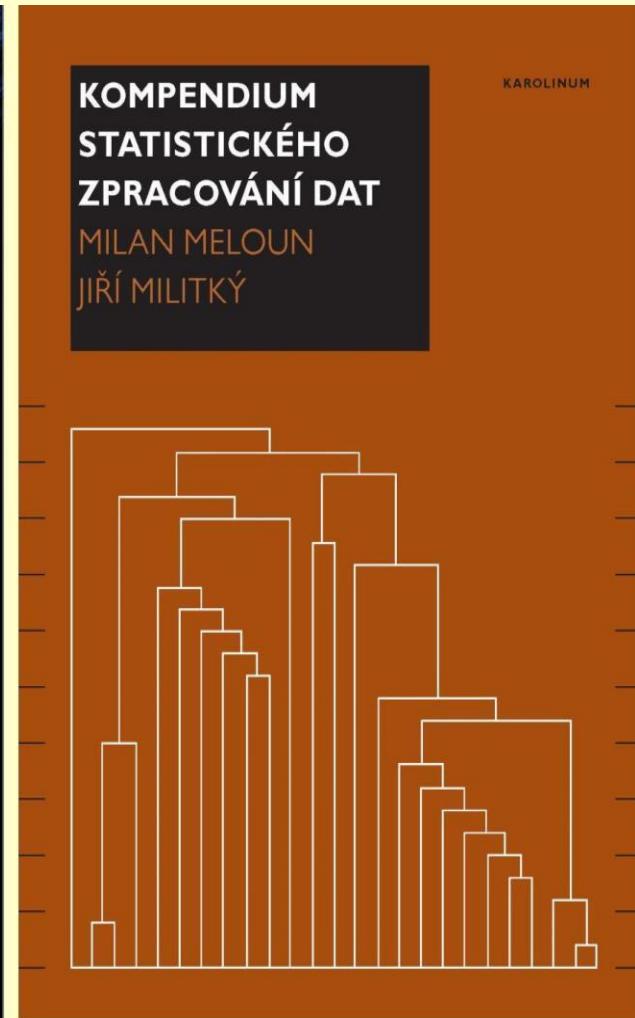
METODY A ŘEŠENÉ ÚLOHY VČETNĚ CD



1. vydání 2002

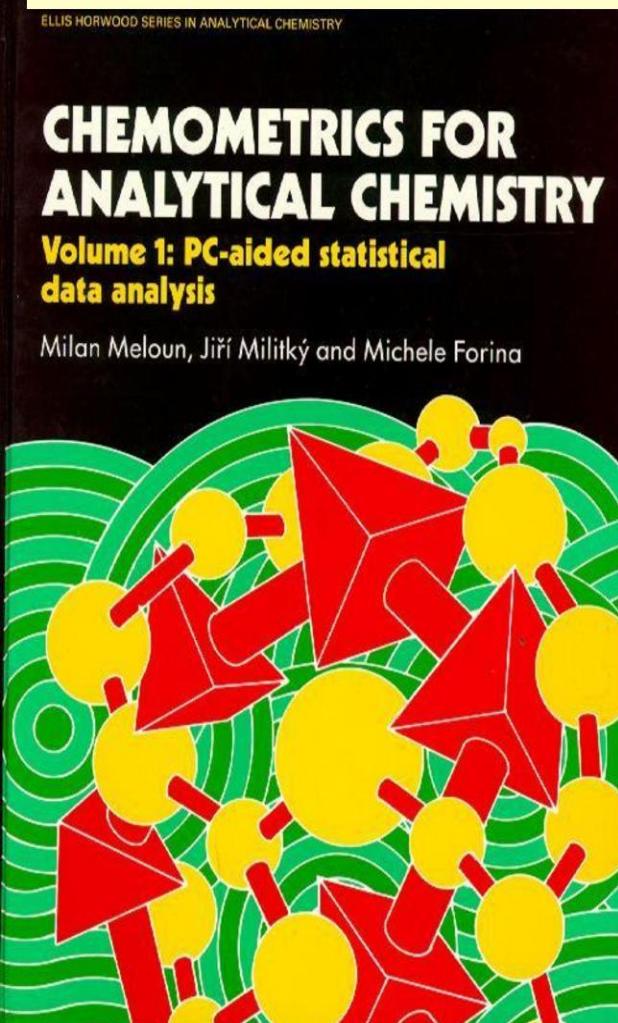


2. vydání 2006

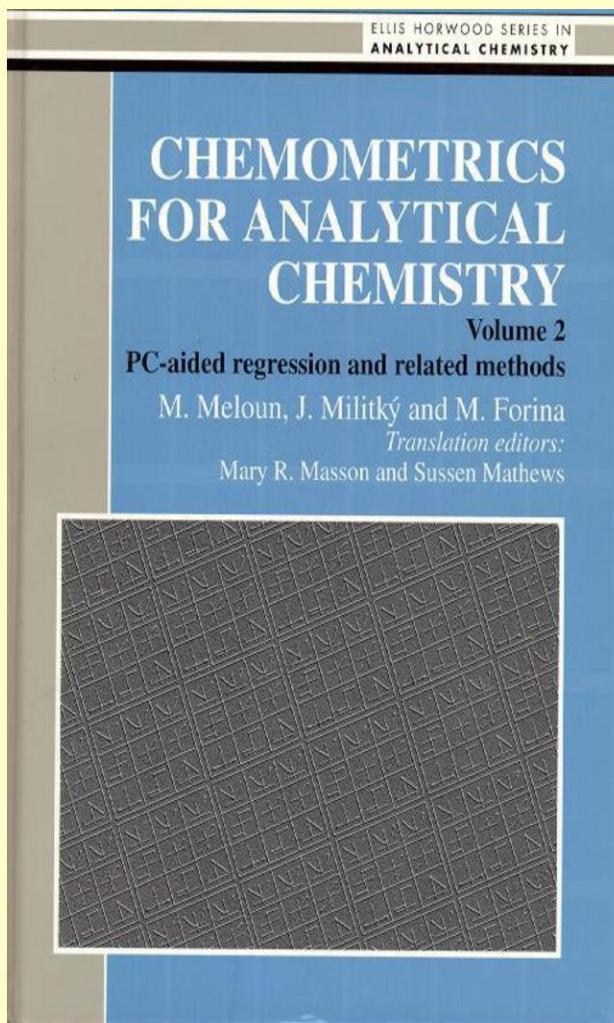


3. vydání 2012

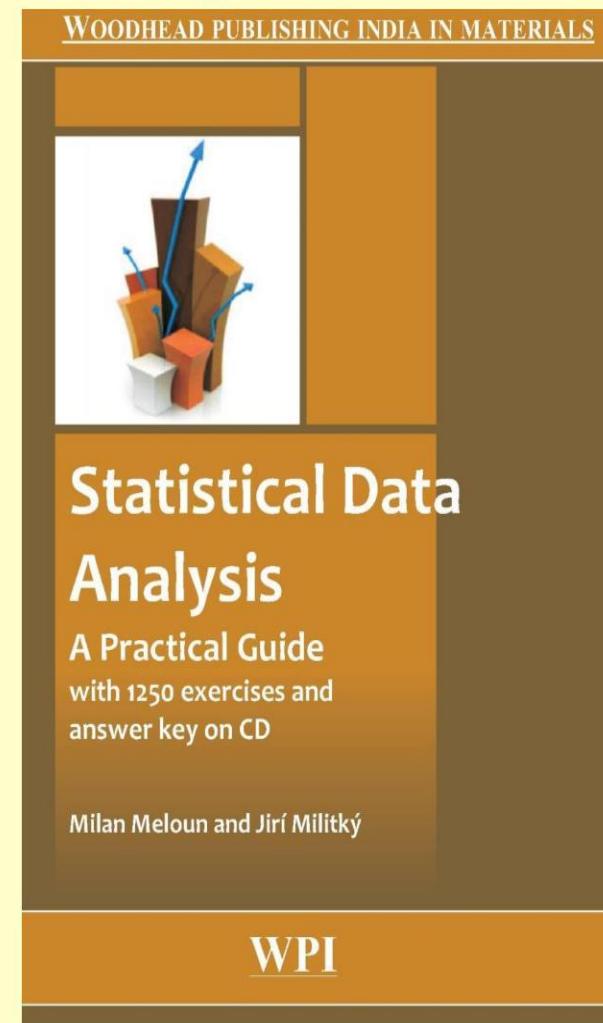
Recommended our textbooks in English



1. vydání 1992

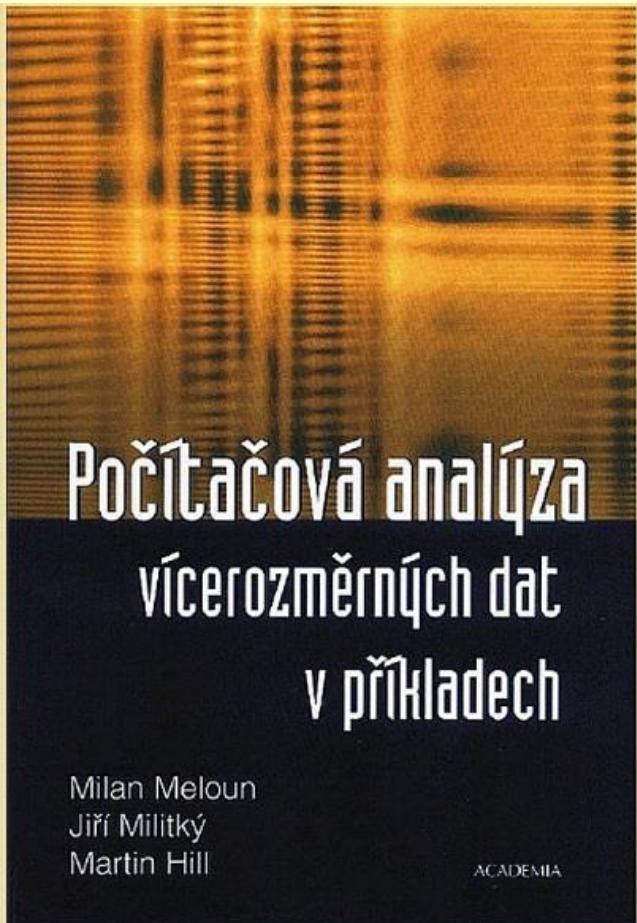


1. vydání 1994

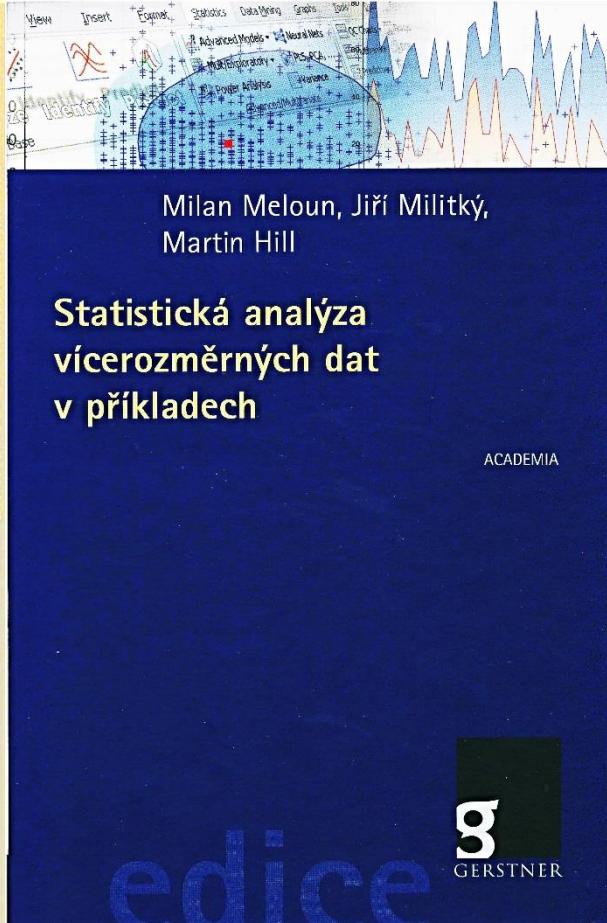


1. vydání 2011

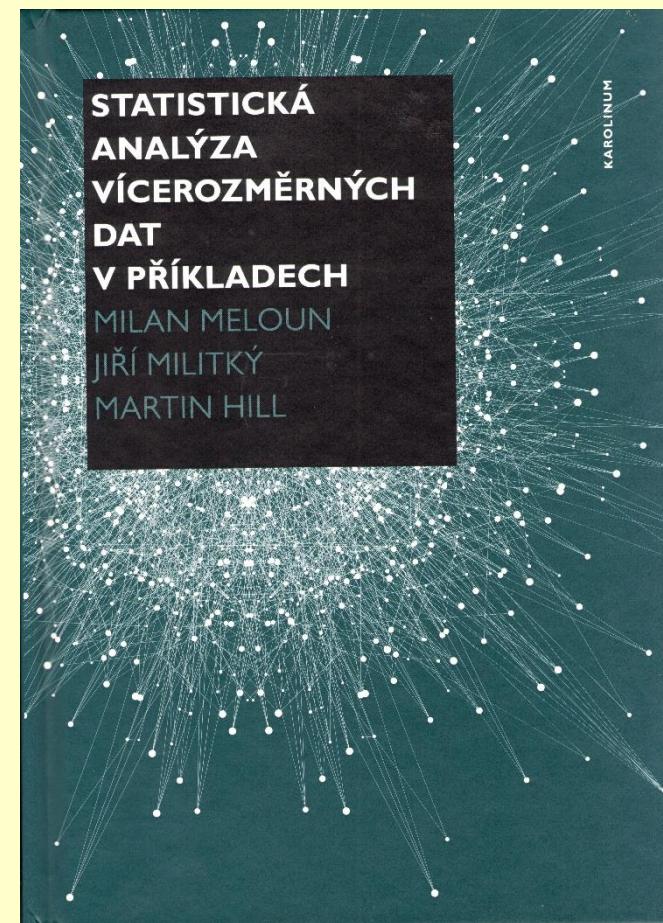
Celostátní učebnice Statistické analýzy vícerozměrných dat



1. vydání Academia 1998

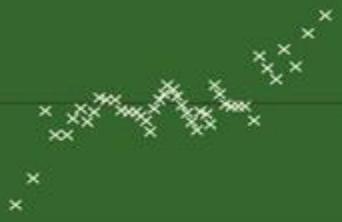


2. vydání Academia 2012



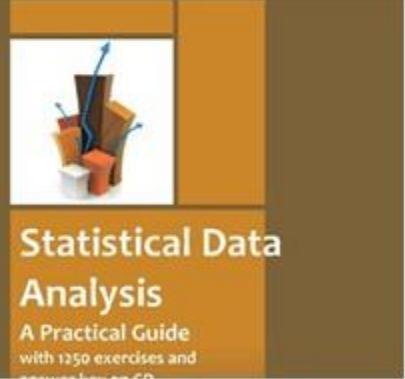
3. vydání Karolinum 2016

INTERAKTIVNÍ
STATISTICKÁ
ANALÝZA DAT
MILAN MELOUN
JIRÍ MILITKÝ



M. Meloun, J. Militký:
Interaktivní statistická analýza dat, Karolinum Praha 2012, 4. vydání, včetně DVD s databazí dat 1700, 955 stran

WOODHEAD PUBLISHING INDIA IN MATERIALS

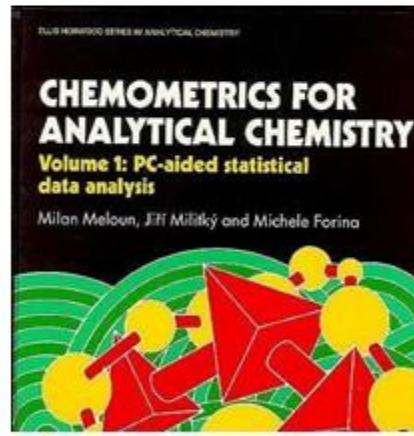


M. Meloun, J. Militký: Statistical Data Analysis, A Practical Guide with 1250 Exercises and Answer key on CD, Woodhead Publishing India, 2011, 1600 pages, ISBN: 978-93-80308-11-1

KOMPENDIUM
STATISTICKÉHO
ZPRACOVÁNÍ DAT
MILAN MELOUN
JIRÍ MILITKÝ

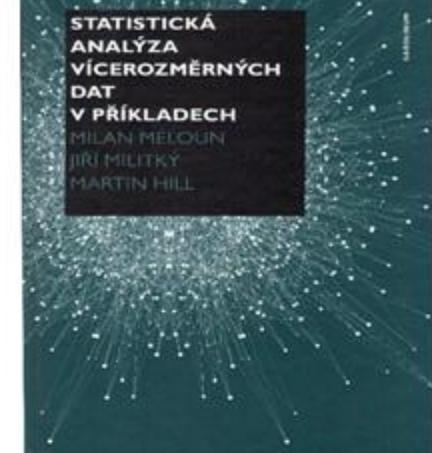


M. Meloun, J. Militký:
Kompendium statistického zpracování dat, Karolinum Praha 2012, 3. vydání, včetně DVD s databazí dat 1700, 985 stran.

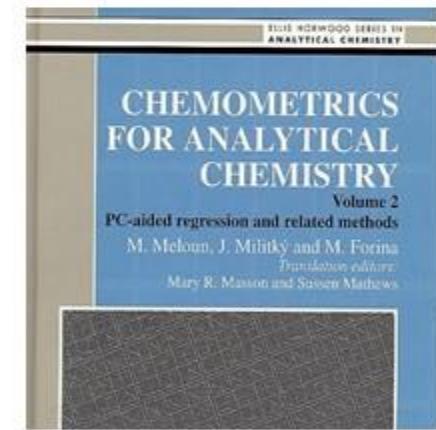


M. Meloun, J. Militký , M. Forina:
CHEMOMETRICS FOR ANALYTICAL CHEMISTRY: Volume 1: PC-Aided Statistical Data Analysis, Ellis Horwood, Chichester 1992, 330 stran, ISBN 0-13-126376-5

STATISTICKÁ
ANALÝZA
VÍCEROZMĚRNÝCH
DAT
V PŘÍKLADECH
MILAN MELOUN
JIRÍ MILITKÝ
MARTIN HIL



M. Meloun, J. Militký, M. Hil:
Statistická analýza vícerozměrných dat v příkladech, Karolinum Praha 2017, 3. vydání, ISBN 978-80-246-3618-4, včetně DVD úloh, 757 stran.



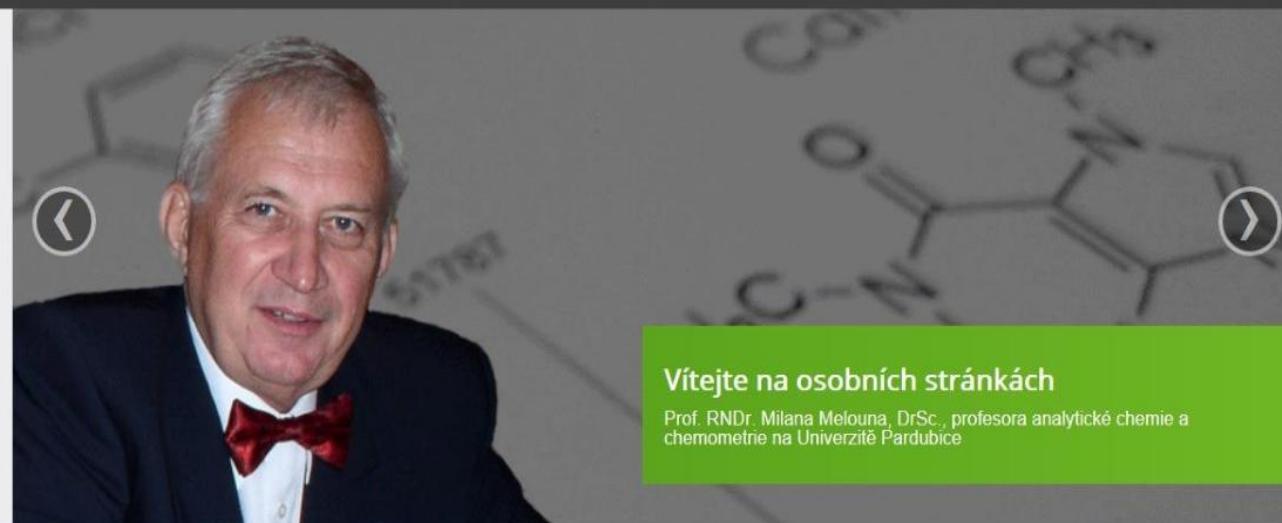
M. Meloun, J. Militký , M. Forina:
CHEMOMETRICS FOR ANALYTICAL CHEMISTRY. Volume 2: PC-Aided Regression and Related Methods, Ellis Horwood, Chichester 1994, ISBN 0-13-123788-7

Klikni na <http://meloun.upce.cz>



Milan Meloun

Profesor analytické chemie and chemometrie na Univerzitě Pardubice.



Vítejte na osobních stránkách

Prof. RNDr. Milana Melouna, DrSc., profesora analytické chemie a chemometrie na Univerzitě Pardubice

Menu

- [Domů](#)
- [Aktuality](#)
- [Životopis](#)
- [Výuka](#)
- [Výzkum](#)
- [Publikace](#)
- [Ke stažení](#)
- [Fotogalerie](#)
- [Užitečné odkazy](#)



Výuka

Blok obsahuje studijní materiály jako snydy přednášek, sylaby ke zkoušce, zkušební otázky, vzorové semestrální práce studentů a doporučenou literaturu k



Výzkum

Blok obsahuje oddíl **Projekty** s názvy grantů a projektů, dále tituly PhD dizertací a názvy diplomových prací ke studiu rovnováh v roztočích. Oddíl **Rovnováhy**



Publikace

Blok obsahuje 6 oddílů publikační aktivity, jako jsou **Kompendium**, **Původní práce**, **Doporučené knihy**, **Konference**, **Patenty** a **Citační index**. Oddíl



Děkuji za pozornost!

<http://meloun.upce.cz>