

The Analysis of Soil Cores Polluted with Certain Metals using the Box-Cox Transformation

Prof. RNDr. Milan Meloun, DrSc.

Department of Analytical Chemistry,
University of Pardubice, CZ532 10 Pardubice, Czech Republic,

milan.meloun@upce.cz

<http://meloun.upce.cz>

High concentrations of **some heavy metals in soils** can cause long-term harm to ecosystems and humans.

Soil survey, monitoring and inventarization define soil properties for a given area or country.

The Register of Contaminated Sites is established as part of the Fertilisers Act (Act No. 156/98 S.B.) and connected decrees in the Czech Republic.

A survey of **the risk element (Cd, Pb, Cr, Hg) content** of agricultural soils on a 1 km² grid was implemented from 1990 to 1993.

The elements **Be, Co, Ni, V and Zn** established a database which has since been continuously filled out by the results of supplementary sampling.

Each sample in the batch is identified by geographical co-ordinates.

A batch of over **40 000 soil samples** has been analysed for the Register database, and the exact sample sizes for each element (2M HNO₃ extraction) are available:

Be: 16544 values,

Cd: 40317 values,

Co: 22176 values,

Cr: 40318 values,

Hg: 32344 values,

Ni: 34989 values,

Pb: 40344 values,

V: 20373 values,

Zn: 36123 values.

Samples from different areas were analysed for the selected range of elements As, Be, Cd, Co, Cr, Cu, Mo, Ni, Pb, V, Zn determined by the AAS or ICP method in 2M HNO₃ extraction, and total Hg content:.

The detection limit [mg.kg⁻¹] and **quantification limit [mg.kg⁻¹]** for the quantitative determination of elements are respectively,

for **As 1.307 [mg.kg⁻¹]** and **4.619 [mg.kg⁻¹]**,
for **Be 0.06 [mg.kg⁻¹]** and **0.197 [mg.kg⁻¹]**,
for **Cd 0.061 [mg.kg⁻¹]** and **0.196 [mg.kg⁻¹]**,
for **Co 0.658 [mg.kg⁻¹]** and **2.203 [mg.kg⁻¹]**,
for **Cr 0.598 [mg.kg⁻¹]** and **2.179 [mg.kg⁻¹]**,
for **Cu 0.515 [mg.kg⁻¹]** and **1.821 [mg.kg⁻¹]**,
for **Hg 0.02 [mg.kg⁻¹]** and **0.06 [mg.kg⁻¹]**,
for **Mo 0.1 [mg.kg⁻¹]** and **0.297 [mg.kg⁻¹]**,
for **Ni 0.574 [mg.kg⁻¹]** and **1.992 [mg.kg⁻¹]**,
for **Pb 0.957 [mg.kg⁻¹]** and **2.916 [mg.kg⁻¹]**,
for **V 1.611 [mg.kg⁻¹]** and **5.764 [mg.kg⁻¹]**,
for **Zn 1.37 [mg.kg⁻¹]** and **3.979 [mg.kg⁻¹]**.

Properly processed analytical data can be used both in research, in government and in legislation:

- (a) Results may serve as **a national database** characterising the degree of pollution of agricultural soils;
- (b) The appropriate parts of such a database may be distributed to local offices, (**environmental sections**) to be available to the regional and local government (e.g. in **urban planning, privatisation projects, changes in land use**, the application of **sewage sludge** or **sediment to agricultural soil**);
- (c) Results may be used in the process of constructing legislative measures concerning the **limit values of harmful substances in soil**;
- (d) A database can serve as one source for calculating critical loads and balances of **risk elements in agroecosystems**.

Proposed procedure of statistical data treatment

Step 1: *Survey of descriptive statistics*

Step 2: *Basic diagnostic plots in the exploratory data analysis*

Step 3: *Determination of sample distribution*

Step 4: *Tests of basic assumptions about the sample*

Step 5: *Data transformation (Power or Box-Cox)*

Step 1: *Survey of descriptive statistics*

Estimate of	Beryllium	Cadmium	Cobalt	Chromium	Mercury	Nickel	Lead	Vanadium	Zinc
Sample size n	16544	40317	22176	40318	32344	34989	40344	20373	36123
Minimum x_1	0	0	0.2	0.1	0	0.1	0.17	0.37	0.7
Maximum x_n	9.33	28.1	110.5	1577.4	69.086	662.0	1121.0	86.0	2070.0

Classical estimates of location, scale and shape

Sample mean	0.470±0.004	0.238±0.003	5.593±0.039	7.104±0.170	0.105±0.006	6.033±0.081	18.637±0.299	10.878±0.083	19.354±0.234
Stand. deviation	0.264	0.300	2.930	17.35	0.534	7.728	30.594	6.015	22.73
Skewness	5.99	30.74	4.19	40.09	107.88	34.49	19.77	2.16	34.20
Kurtosis	119	2123.1	89.85	2608.52	12963.7	2298.8	528.2	12.41	2265.0

Robust estimates of location

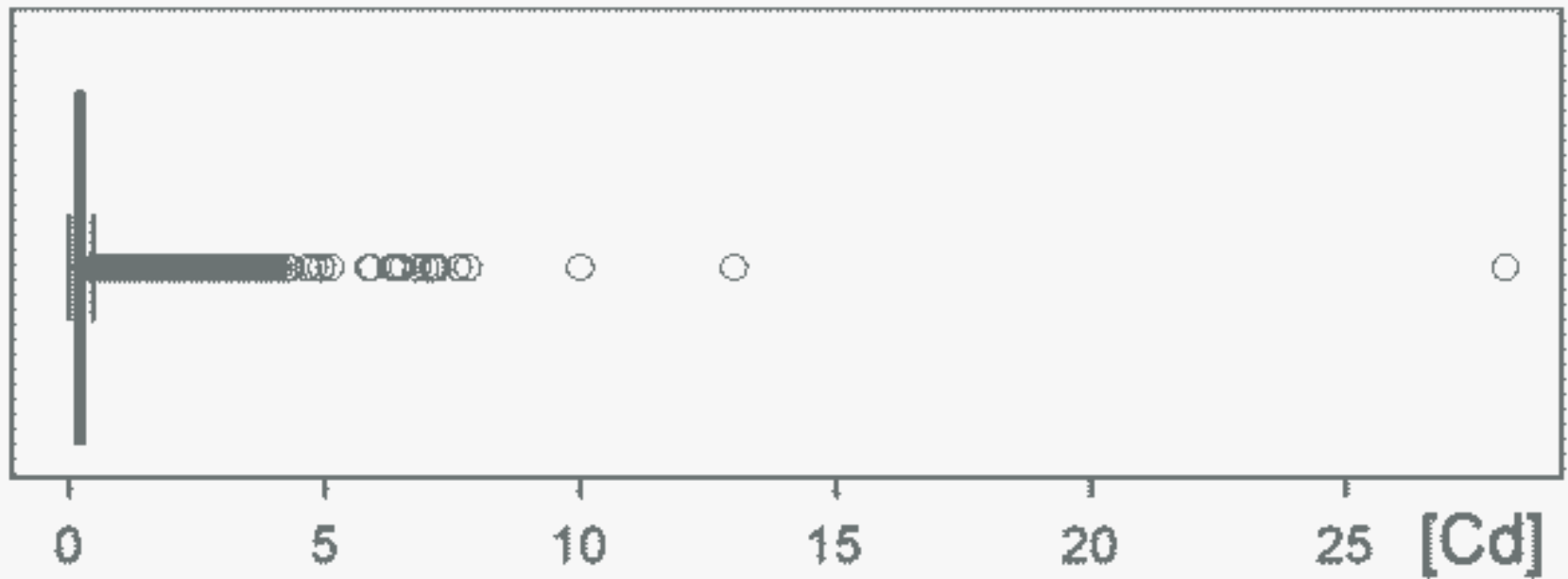
Median	0.43±0.01	0.19±0.00	5.0±0.0	4.60±0.05	0.08±0.00	4.70±0.05	14.90±0.05	9.60±0.10	16.0±0.05
Trimmed mean	0.449±0.003	0.210±0.001	5.356±0.033	5.361±0.040	0.086±0.001	5.320±0.039	15.860±0.067	10.320±0.074	17.446±0.089

Step 2: *Basic diagnostic plots in the exploratory data analysis (EDA)*

For a graphical visualization of data the EDA-plots are used:

- (a) the box-and-whisker plot
and the notched box-and-whisker plot,
- (b) the quantile plot.

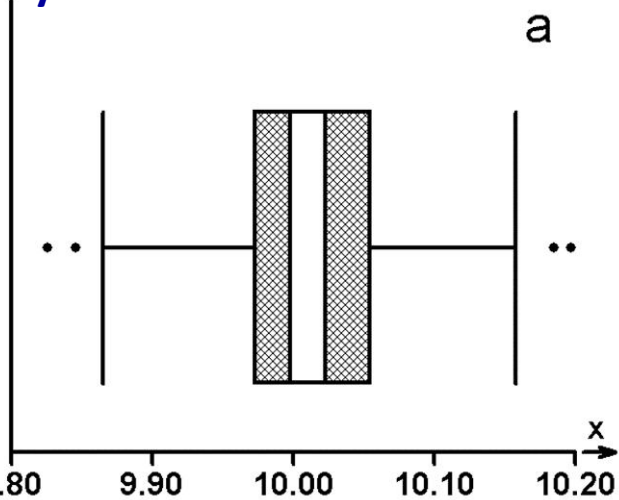
Results: Since EDA-diagnostics, the skewness and kurtosis prove that the sample distribution strongly differs from a normal one, the data should be examined to find the transformation leading to symmetric distribution, stabilizing variance and making the distribution closer to normal.



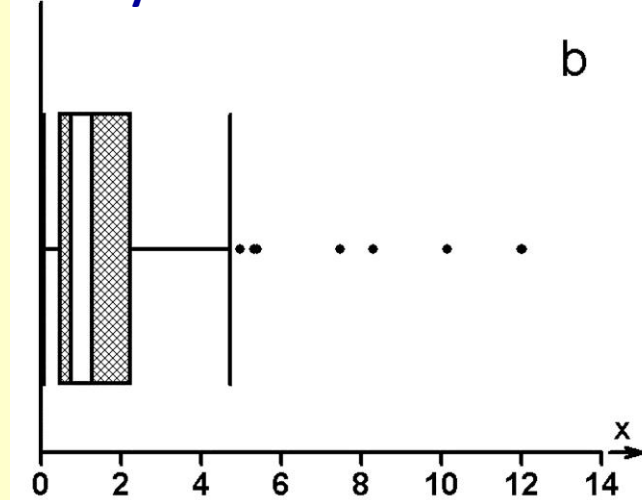
The box-and-whisker plot of the Cd sample data indicates too many outliers.

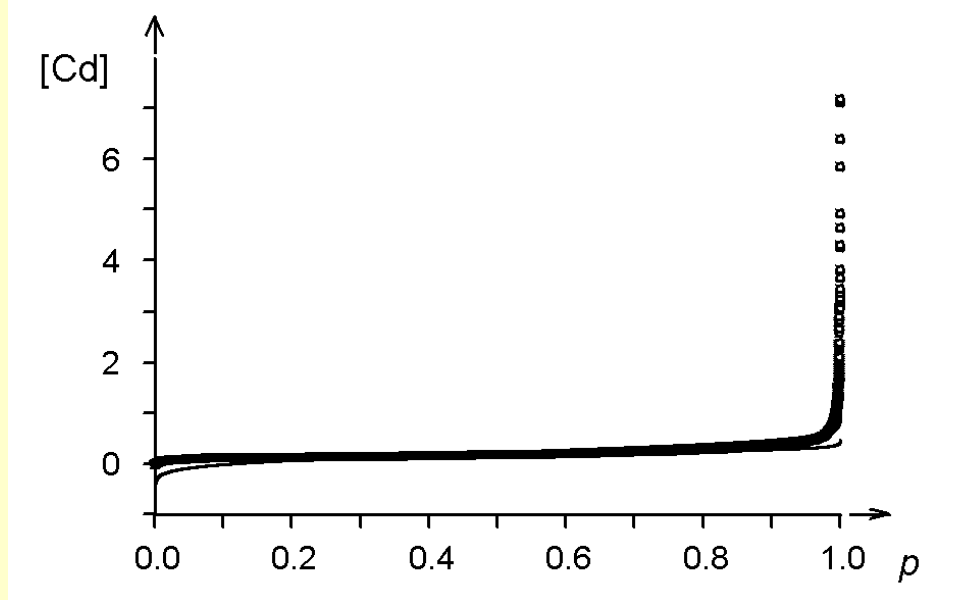
Interpretation of the diagnostic tool:

Symmetric distribution



Asymmetric distribution

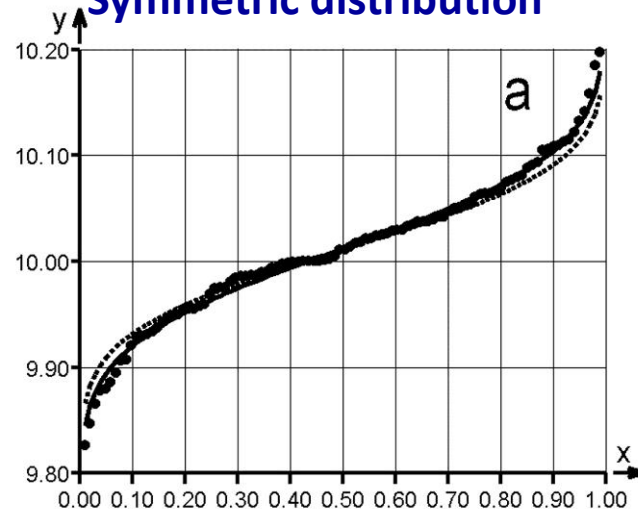




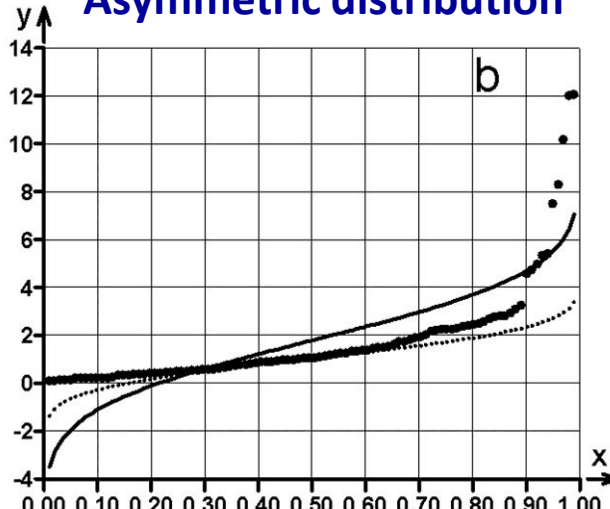
The quantile plot of the Cd sample data indicates strongly asymmetric, skewed distribution.

Interpretation of the diagnostic tool:

Symmetric distribution



Asymmetric distribution

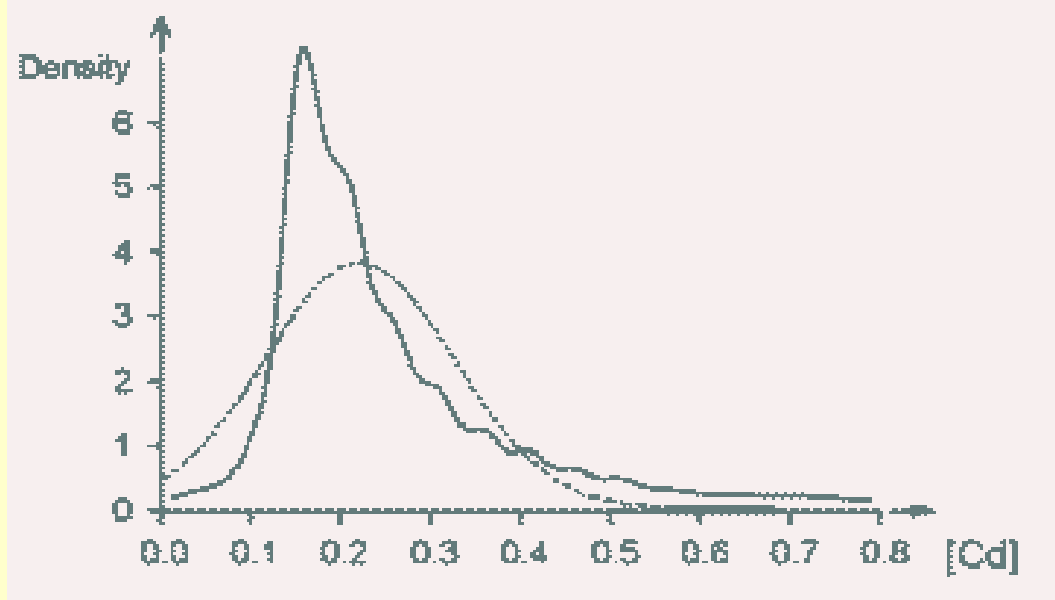


Step 3: *Determination of sample distribution*

The sample distribution represented by symmetry, skewness and kurtosis is examined by two diagnostic tools:

(c) the kernel density estimator of **the probability density function** ,

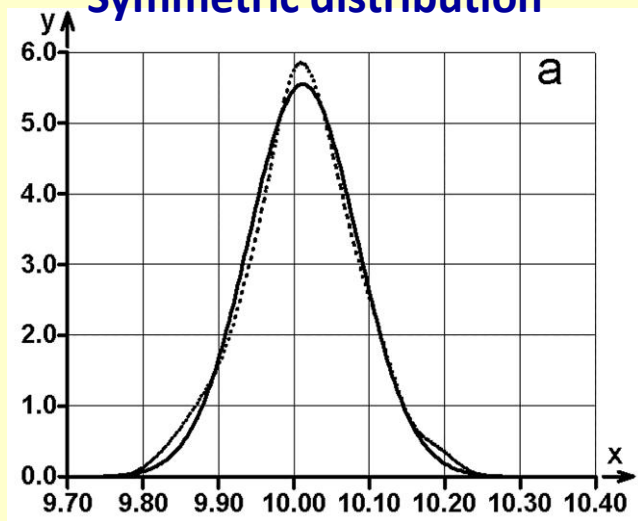
(d) **the quantile-quantile plot**, which is used for comparison of the actual with the theoretical sample distribution.



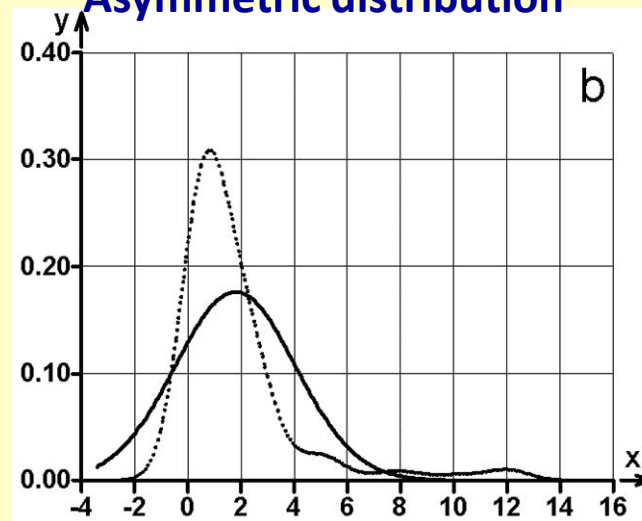
The kernel density estimator of the probability density function of the Cd data shows that the sample points are located in one class and the plot indicates strongly skewed distribution.

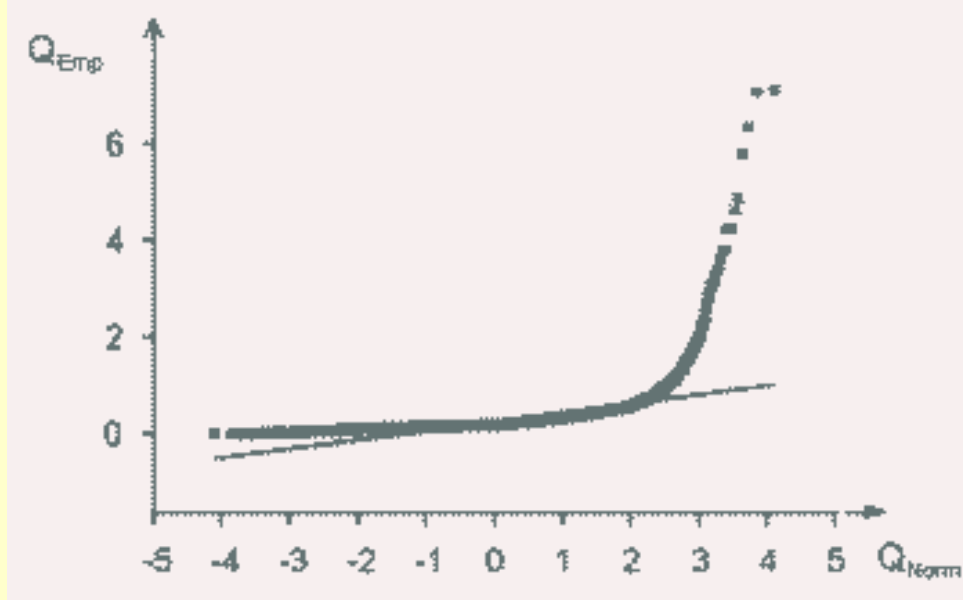
Interpretation of the diagnostic tool:

Symmetric distribution



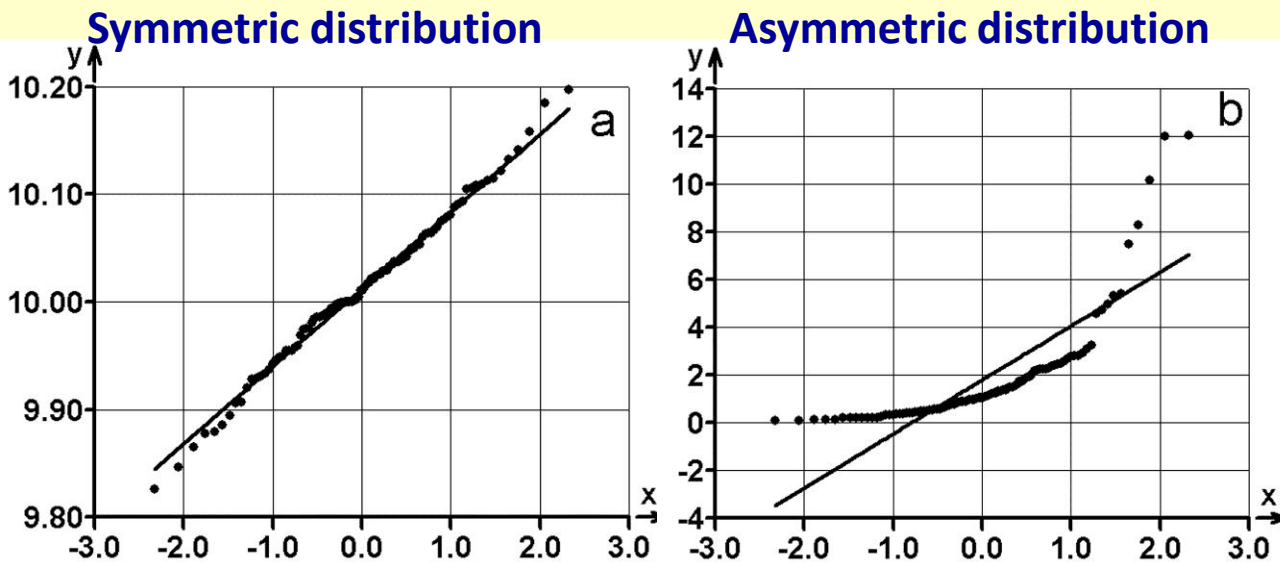
Asymmetric distribution





The quantile-quantile plot (for normal distribution called the rankit plot) of the Cd sample data does not exhibit close agreement of the sample points with a straight line.

Interpretation of the diagnostic tool:



Step 4: Tests of basic assumptions about the sample

Jarque-Berra normality test, critical value ($\alpha = 0.05$) is $\chi^2_{0.95}(2) = 5.99$

Estimate of	Beryllium	Cadmium	Cobalt	Chromium	Mercury	Nickel	Lead	Vanadium	Zinc
Sample size n	16544	40317	22176	40318	32344	34989	40344	20373	36123
Test. criterion	157.1,	291.1,	145.9,	311.1,	382.0,	294.3,	259.3,	111.4,	294.9,
Conclusion	Reject	Reject	Reject	Reject	Reject	Reject	Reject	Reject	Reject

Homogeneity test

Outliers	265	2095	496	2285	1180	1128	1359	486	961
----------	-----	------	-----	------	------	------	------	-----	-----

Conclusion of EDA:

- 1) A combined sample skewness and kurtosis test according to Jarque-Berra does not prove normality of the sample distribution for all elements analyzed.
- 2) Transformation of sample data is necessary.
- 2) Power transformation or Box-Cox transformation may be used.

Step 5: Data transformation (Power or Box-Cox)

The transformation leads to a symmetric data distribution, stabilizes the variance or makes the distribution closer to normal.

(i) **Power transformation** for variance stabilization implies ascertaining the transformation $y = g(x)$ in which the variance $\sigma^2(y)$ is constant.

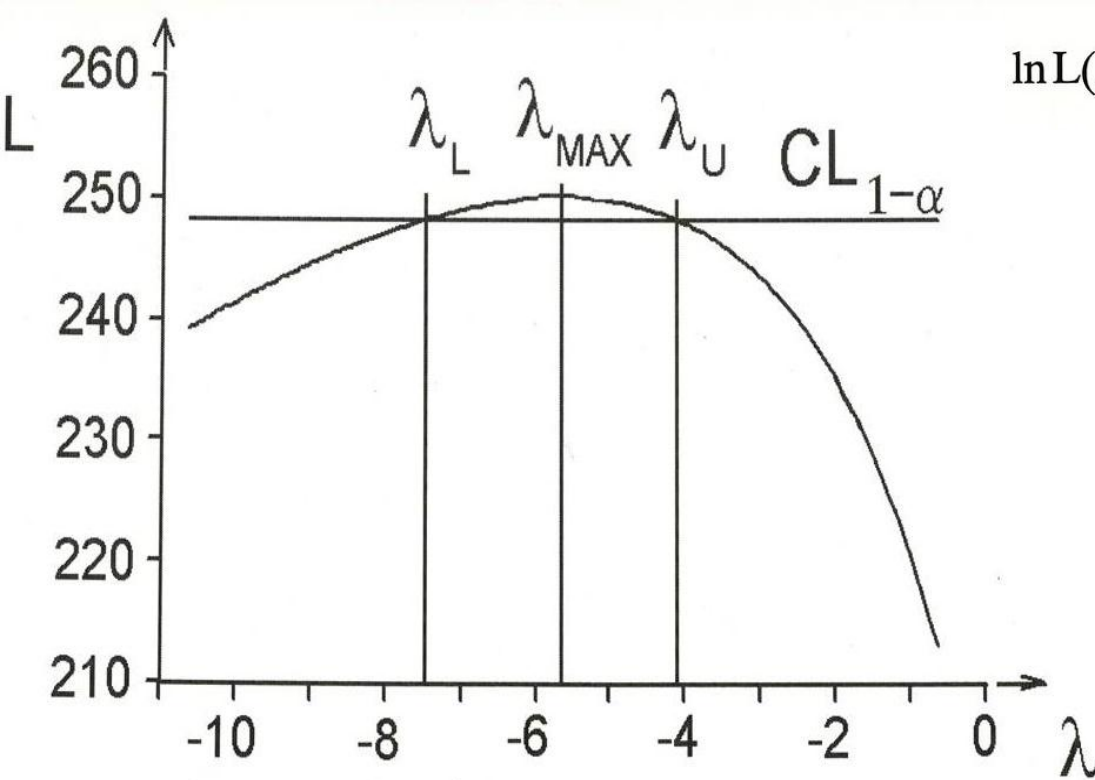
$$y = g(x) = \begin{cases} x^\lambda & \lambda > 0 \\ \ln x & \text{pro } \lambda = 0 \\ -x^{-\lambda} & \lambda < 0 \end{cases}$$

(ii) Transformation leading to approximate normality may be carried out by the **Box-Cox transformation family**. The Box-Cox transformation can be applied only to positive data.

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

Results: While classical measures of location, spread and shape for the original data are **0.238 mg.kg⁻¹**, $s(x) = 0.300 \text{ mg.kg}^{-1}$, the skewness 30.74 and kurtosis 2123.04 are out of statistical significance and may be taken as **false estimates of location**.

The Box-Cox t. with $\lambda = -0.0556$ calculates the corrected mean **0.187 ± 0.001 mol.dm⁻³**.



$$\ln L(\lambda) = -\frac{n}{2} \ln s^2(y) + (\lambda - 1) \sum_{i=1}^n \ln x_i$$

For 100(1-a)% confidence interval of λ it is valid that

$$2 [\ln L(\hat{\lambda}) - \ln L(\lambda)] \leq \chi^2_{1-\alpha}(1)$$

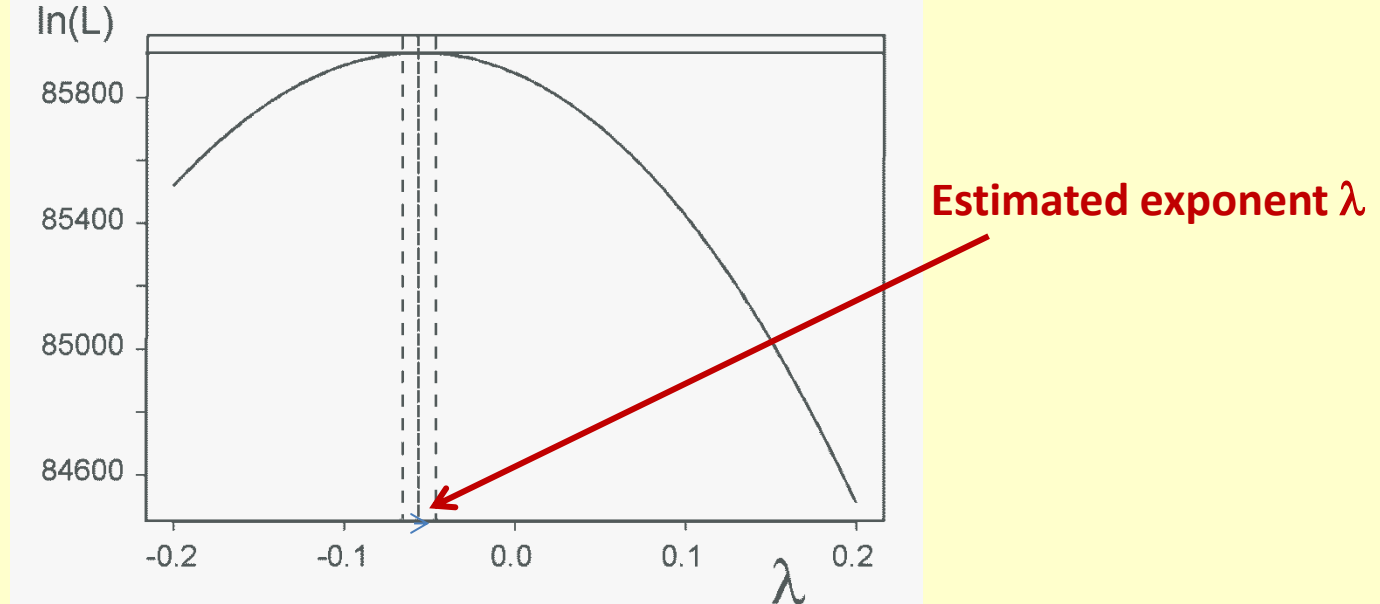
In the confidence interval there are all λ for which it is valid

$$\ln L(\lambda) \geq \ln L(\hat{\lambda}) - 0.5 \chi^2_{1-\alpha}(1)$$

The function $\ln L = f(\lambda)$ is expressed graphically for a suitable interval, for example, $-3 \leq \lambda \leq 3$.

The **maximum on this curve** represents the maximum likelihood estimate.

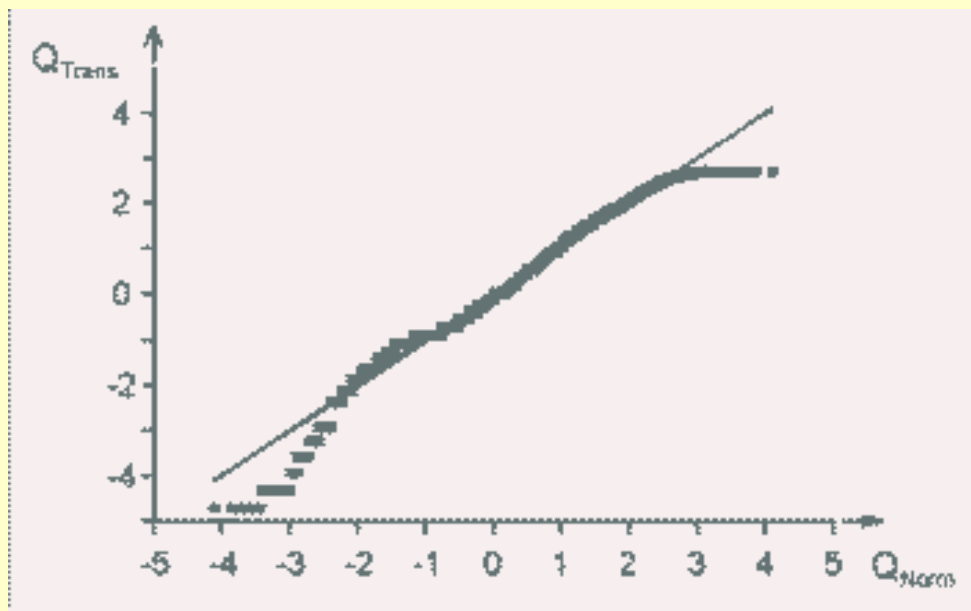
When the value $\lambda = 1$ is also covered by this confidence interval, the **transformation is not efficient**.



The plot of the logarithm of the maximum likelihood for the Cd sample data with Box-Cox transformation enables estimation of the power λ .

From the plot of the logarithm of the likelihood function for the Box-Cox transformation the maximum of the curve is at $\lambda = -0.0556$.

The corresponding 95% confidence interval does not contain the exponent value $\lambda = 1$, so **all transformations are statistically significant**.



The quantile-quantile plot for the Cd sample data with Box-Cox transformation shows that most sample points are on the straight line and the measure of location is more reliable now.

Re-transformed estimation is calculated using the Taylor equation

$$\bar{x}_R \approx g^{-1} \left[\bar{y} - \frac{1}{2} \frac{d^2 g(x)}{dx^2} \left(\frac{dg(x)}{dx} \right)^{-2} s^2(y) \right]$$

$$s^2(x_R) \approx \left(\frac{dg(x)}{dx} \right)^{-2} s^2(y) .$$

Estimate of	Beryllium	Cadmium	Cobalt	Chromium	Mercury	Nickel	Lead	Vanadium	Zinc
Sample size n	16544	40317	22176	40318	32344	34989	40344	20373	36123
Outliers	265	2095	496	2285	1180	1128	1359	486	961

Classical and robust estimates of location

Sample mean	0.470±0.004	0.238±0.003	5.593±0.039	7.104±0.170	0.105±0.006	6.033±0.081	18.637±0.299	10.878±0.083	19.354±0.234
Median	0.43±0.01	0.19±0.00	5.0±0.0	4.60±0.05	0.08±0.00	4.70±0.05	14.90±0.05	9.60±0.10	16.0±0.05
Trimmed mean	0.449±0.003	0.210±0.001	5.356±0.033	5.361±0.040	0.086±0.001	5.320±0.039	15.860±0.067	10.320±0.074	17.446±0.089

Box-Cox transformation of location estimate

Re-transformed	0.427	0.187	5.078	4.922	0.082	4.797	15.172	9.611	16.360
mean	±0.003	±0.001	±0.030	±0.023	±0.001	±0.020	±0.050	±0.050	±0.050

The most rigorous estimate of location is represented by the re-transformed mean after Box-Cox transformation of original data.

This estimate can be taken as the best for each element studied here.

Conclusion:

1) All EDA display techniques prove that the sample distribution is **asymmetric, skewed with many outliers**.

2) Therefore, the sample distribution does **not come** from a population with a normal distribution.

3) The classical measures \bar{x} and s are strongly corrupted with outliers and cannot be used here. The arithmetic mean does not represent an objective measure of location,

$0.238 \pm 0.003 \text{ mg.kg}^{-1}$, and can not be used.

4) On the basis of the quantile-quantile plot the Box-Cox transformation is considered the most rigorous technique to estimate a measure of location, with the corrected mean value

$0.187 \pm 0.001 \text{ mol.dm}^{-3}$.

Estimate of	Beryllium	Cadmium	Cobalt	Chromium	Mercury	Nickel	Lead	Vanadium	Zinc
Sample size n	16544	40317	22176	40318	32344	34989	40344	20373	36123
Classical estimates of location, scale and shape									
Sample mean	0.470±0.004	0.238±0.003	5.593±0.039	7.104±0.170	0.105±0.006	6.033±0.081	18.637±0.299	10.878±0.083	19.354±0.234
St. deviation	0.264	0.300	2.930	17.35	0.534	7.728	30.594	6.015	22.73
Skewness	5.99	30.74	4.19	40.09	107.88	34.49	19.77	2.16	34.20
Kurtosis	119	2123.1	89.85	2608.52	12963.7	2298.8	528.2	12.41	2265.0
Robust estimates of location									
Median	0.43±0.01	0.19±0.00	5.0±0.0	4.60±0.05	0.08±0.00	4.70±0.05	14.90±0.05	9.60±0.10	16.0±0.05
Trimmed mean	0.449±0.003	0.210±0.001	5.356±0.033	5.361±0.040	0.086±0.001	5.320±0.039	15.860±0.067	10.320±0.074	17.446±0.089
Jarque-Berra normality test, critical value for $\alpha = 0.05$ is $\chi^2_{0.95}(2) = 5.99$ and Homogeneity test									
Testing criterion	157.1, Reject	291.1, Reject	145.9, Reject	311.1, Reject	382.0, Reject	294.3, Reject	259.3, Reject	111.4, Reject	294.9, Reject
Outliers	265	2095	496	2285	1180	1128	1359	486	961
Box-Cox transformation									
Re-transformed	0.427	0.187	5.078	4.922	0.082	4.797	15.172	9.611	16.360

http://meloun.upce.cz

Site Feed | Site map

A⁺ A A⁻



MILAN MELOUN



Home & News

Personal

Teaching

Research

Publishing

Download

Photo gallery

Useful links

Www visitors

MAIN MENU

- Home & News
- Personal
- Teaching
- Research
- Publishing
- Download
- Photo gallery
- Useful links
- Wwv visitors

We have 17 guests online

search...

Prof. RNDr. Milan Meloun, DrSc.



professor of analytical chemistry and chemometrics

Address: [Department of Analytical Chemistry, University of Pardubice](#)
Studentská 573, blok HB/D, 5. patro, 53210 Pardubice, Czech Republic

Telephone: + 420-46 603 7026

Fax: + 420-46 603 7068

E-mail: milan.meloun@upce.cz

NEWS: RECOMMENDED BOOKS / NAŠE DOPORUČOVANÉ KNIHY



NEWS: FORMS OF POSTGRADUATE STUDY / FORMY DALŠÍHO VZDĚLÁVÁNÍ

- [Statistické zpracování dat - nejbližší termín](#)
[Intenzivního týdenního kurzu počítačové](#)
[analýzy dat v úlohách](#)
- [Další soustředění nového 13. licenčního studia](#)
[ARCHIMEDES \(Statistické zpracování dat a](#)
[informatika\)](#)
- [Další soustředění 12. licenčního studia](#)
[PYTHAGORAS \(Statistické zpracování](#)
[experimentálních dat\)](#)
- [1. soustředění ARISTOTELES nejdříve v červnu](#)
[2011 \(dle naplnění počtu\)](#)



MAIN MENU

- Home & News
- Personal
- Teaching
- Research
- Publishing
- Download
- Photo gallery
- Useful links
- Wwv visitors

We have 17 guests online

Analytical Chemistry

Chemometrics

Chemometrics I

Chemometrics II

Chemometrics III


Methods

License study -
PYTHAGORASLicense study -
ARCHIMEDESLicense study
ARISTOTELES

Week-Courses

Prof. RNDr. Milan Meloun, DrSc.

professor of analytical chemistry and chemometrics

Address: [Department of Analytical Chemistry, University of Pardubice](#)
Studentská 573, blok HB/D, 5. patro, 53210 Pardubice, Czech Republic**Telephone:**  + 420-46 603 7026 **Fax:** + 420-46 603 7068**E-mail:** milan.meloun@upce.czRECOMMENDED BOOKS / NAŠE
DOPORUČOVANÉ KNIHYNEWS: FORMS OF POSTGRADUATE STUDY /
FORMY DALŠÍHO VZDĚLÁVÁNÍ [Statistické zpracování dat - nejbližší termín](#)
Intenzivního týdenního kurzu počítačové

Materiály k výuce pro licenční studium, kurzy a studenty 4. ročníku.
Slides and tutorials for the Post-Graduate Licence Study and graduates.



Home & News

Personal

Teaching

Research

Publishing

Download

Photo gallery

Useful links

Www visitors

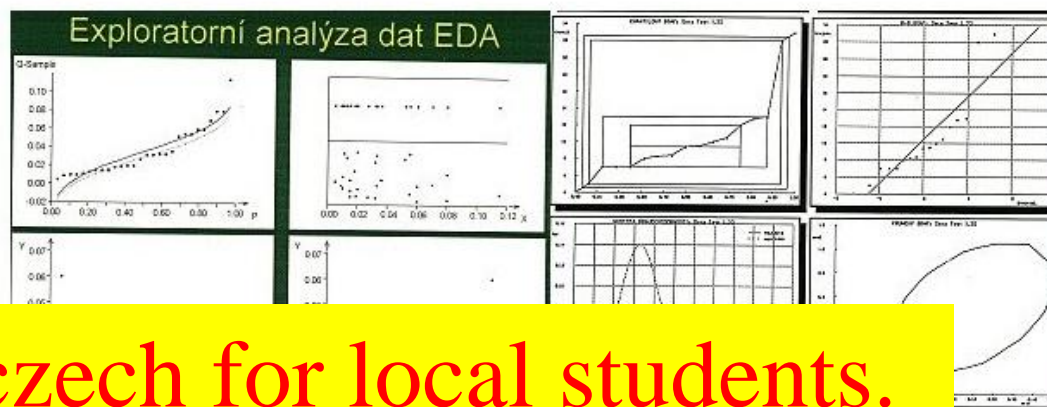
MAIN MENU

- Home & News
- Personal
- Teaching
 - Analytical Chemistry
 - Chemometrics
 - Chemometrics I
 - Chemometrics II
 - Chemometrics III
 - Methods
 - Unpublished

CHEMOMETRICS I. (CHEMOMETRIE I.)



Jednorozměrná data



Učený člověk má své
bohatství vždy v
sobě.

HOMO DOCTUS IN
SE SEMPER
DIVITIAS HABET.
Phaedrus (kol. 15
př.n.l. - 50 n.l.)
Fabulae Aesopiae IV,
22, 1.

Mostly in czech for local students.

Zkoušení: vždy 7.30 až 10 hodin písemně v učebnách nového areálu ChTF

A. Potřebné učebnice a vzory semestrálních prací:

1. M. Meloun, J. Militký: [STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT](#), Plus, Praha 1994 (1. vydání) nebo East Publishing Praha 1998 (2. vydání), Academia 2004.
2. M. Meloun, J. Militký: [Kompedium statistického zpracování dat](#), Učebnice s CD, Academia Praha 2002 (1. vydání), Academia Praha 2006 (2. vydání).
3. M. Meloun, J. Militký: [Sbírka úloh - STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT](#) Univerzita Pardubice 1997
4. [Vzory semestrálních prací studentů v řádném studiu.](#)
5. [Vzory semestrálních prací studentů v licenčním studiu.](#)



Thank you for your attention!

<http://meloun.upce.cz>

Your comments are welcomed:

1) Now, please, if you ask me?

2) Or via email contact: milan.meloun@upce.cz