# EXPLORATORY DATA ANALYSIS IN ANALYTICAL CHEMISTRY OF SMALL AND TRACE CONCENTRATIONS

**Milan Meloun and Jiří Militký** [1)]
Department of Analytical Chemistry, University Pardubice,
532 10 Pardubice, Czech Republic
and
[1)]Department of Textile Materials, Textile Faculty,
Technical University, 461 17 Liberec, Czech Republic,

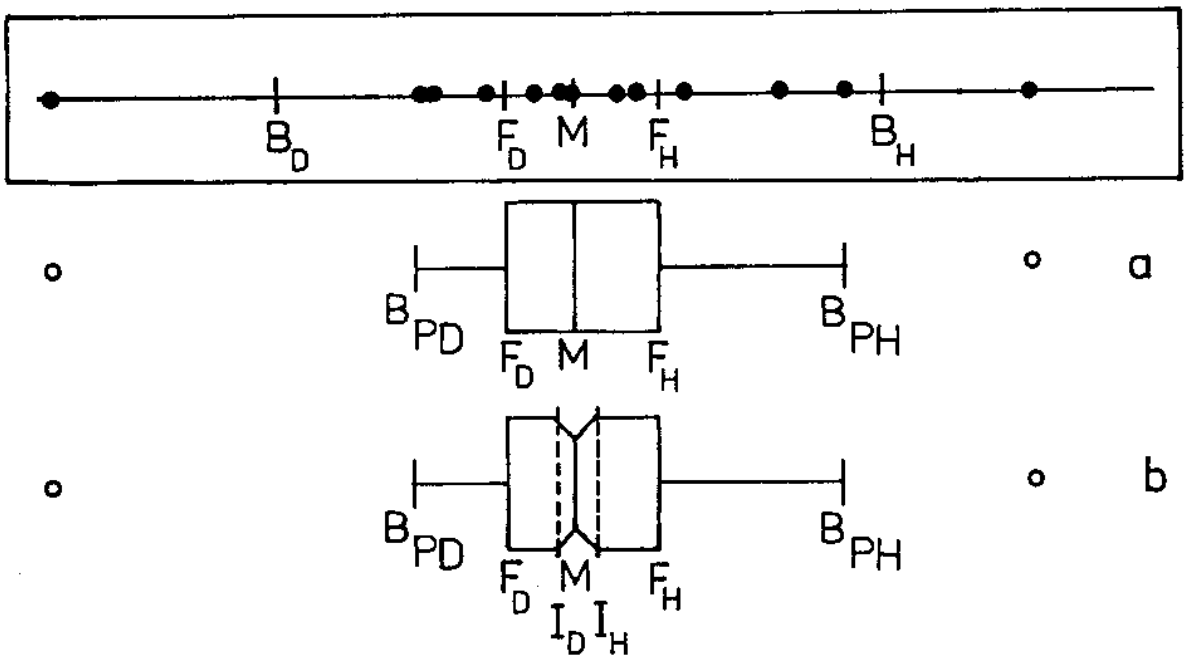
*Summary: Exploratory data analysis (EDA) isolates certain statistical features and patterns of data with the use of the quantile plot, the dot and jittered dot diagrams, the (notched) box-and-whisker plot, the midsum plot, the symmetry plot, the curtosis plot, the differential quantile plot and the quantile-box plot, the kernel estimation of probability density function, the histogram, the frequency polygon, the bar chart, the rootgram, the quantile-quantile plot, the rankit plot and the conditioned rankit plot. The power and Box-Cox transformation is often used.*

The first step of univariate data analysis called an exploratory data analysis (EDA) isolates certain basic statistical features and patterns of data while the second step, a confirmatory data analysis (CDA) stresses evaluation of an available evidence by tests of probability models. According to Tukey the EDA is a "detective work". Its tools are various descriptive graphically oriented techniques which are typically free of strict statistical assumptions about data. The EDA-techniques are often called "distribution-free" and are based on a continuity and differentiability of underlying density only. The EDA techniques are quite effective for an investigation of statistical behavior of data from new or non-standard analytical procedures.

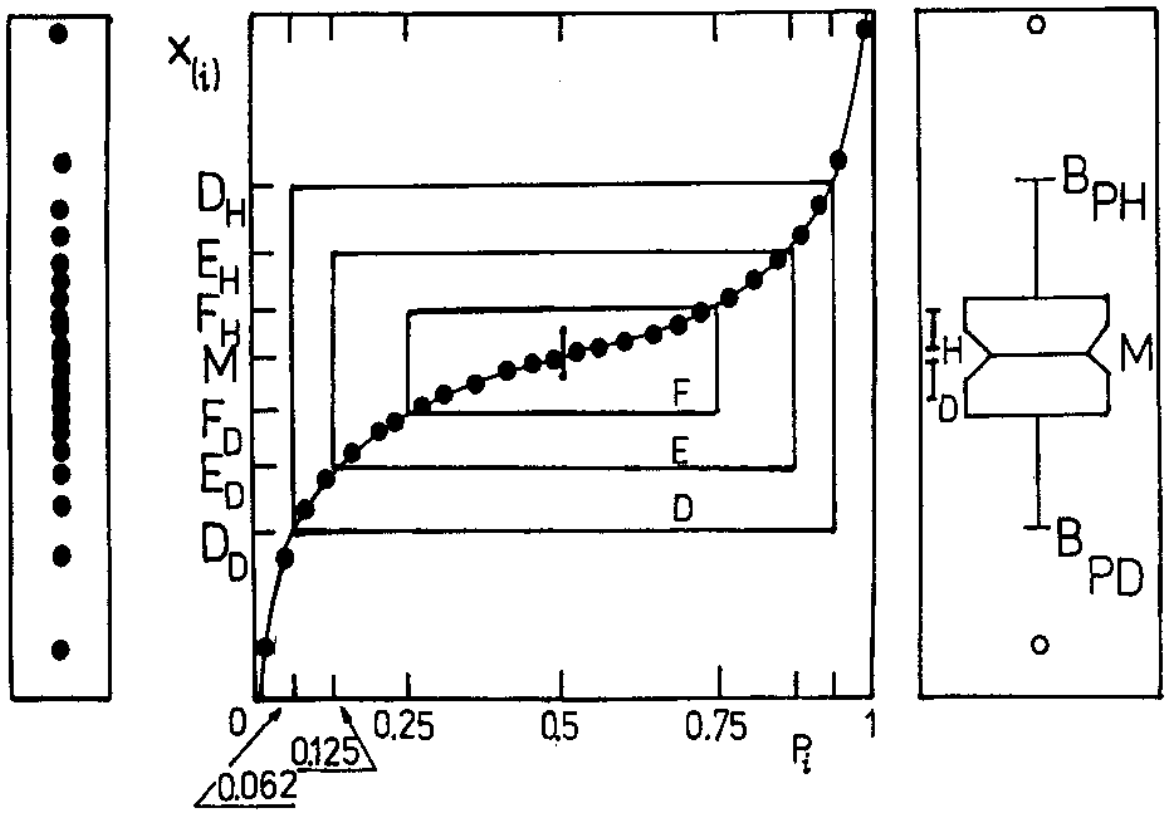
For graphical visualization of data the EDA uses *the quantile plot, the dot and jittered dot diagrams, the (notched) box-and-whisker plot* while the sample distribution is investigated by *the midsum plot, the symmetry plot, the curtosis plot, the differential quantile plot and the quantile-box plot.* The construction of sample distribution i.e. the estimation of probability density function is done by *the stem-and-leaf display, the kernel estimation of probability density function, the histogram, the frequency polygon, the bar chart, the rootgram, the quantile-quantile plot, the rankit plot and the conditioned rankit plot.*

When an exploratory data analysis (EDA) finds that the sample distribution differs from a normal one or when a confirmatory data analysis (CDA) does not prove a sample independence and a sample homogeneity, the original data should be transformed. The power transformation and the Box-Cox transformation improves a sample symmetry and stabilizes a sample variance. *The Hines-Hines selection graph* and *the plot of logarithm of maximal likelihood function* enables to find an optimum power transformation. Diagnostics of the EDA and CDA about assumptions of the actual data set are also examined.

According to results of an examination about sample assumptions the classical, robust or adaptive estimates of location and spread are calculated.




*Construction of (a) the box-and-whisker plot and (b) the notched box-and-whisker plot from the dot diagram. Empty circles indicate outliers*




*An example of a quantile-box plot. The dot diagram (left) and the notched box-and-whisker plot (right) are given for comparison*

# UNIVARIATE DATA ANALYSIS

## 1st stage: EXPLORATORY DATA ANALYSIS (EDA)

EDA is "detective work" which indicates certain statistical features and paterns of data (symmetry, kurtosis, dispersion, outliers, etc.) by the distribution-free technique.

**1. EDA DIAGNOSTIC PLOTS AND DISPLAYS:** Quantile plot, Jittered dot diagram, Box-and-whisker plot, Midsum plot, Symmetry plot, Kurtosis plot, Differential quantile plot, Quantile-box plot.

**2. EXAMINING A SAMPLE DISTRIBUTION:** Q-Q plot, Rankit plot, Conditioned rankit plot.

**3. DATA TRANSFORMATION:** Power and Box-Cox transformations.

## 2nd stage: CONFIRMATORY DATA ANALYSIS (CDA)

CDA is judical in nature and tests four basic assumptions about data. CDA determines parameters of location, spread and distribution shape.

**1. TEST OF BASIC ASSUMPTIONS ABOUT DATA**

(a) A test for minimal sample size
(b) A test for independence of sample elements
(c) A test for homogeneity of sample
(d) A test for normality

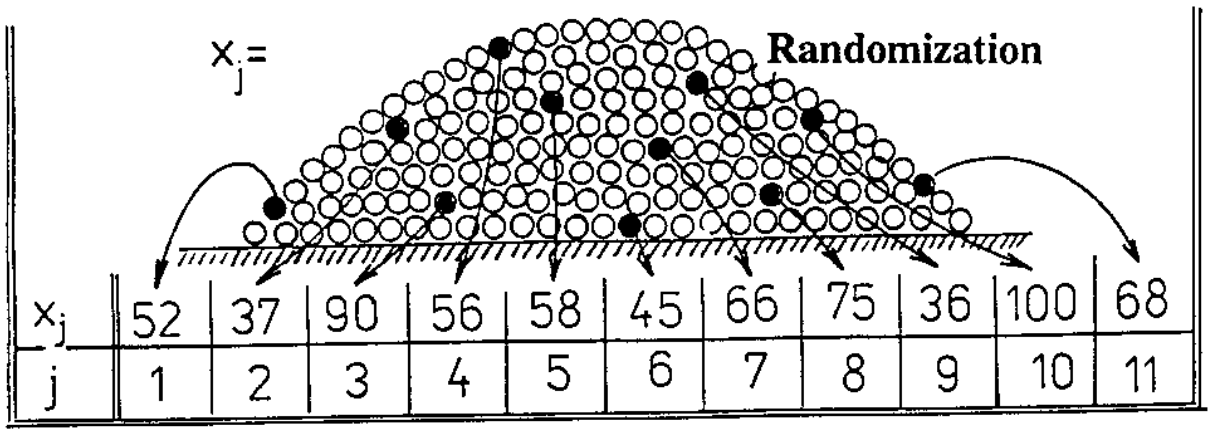
**2. CONSTRUCTION OF PROBABILITY DENSITY FUNCTION:** Stem-and-leaf display, Kernel estimation of probability density, Histogram, Frequency polygon, Bar chart, Rootogram.

**3. POINT ESTIMATES FOR PARAMETERS OF LOCATION, SPREAD AND SHAPE**

**4. INTERVAL ESTIMATES FOR PARAMETERS OF LOCATION AND SPREAD**

**5. STATISTICAL HYPOTHESIS TESTING**

# Sampling, Sorting, Ranking and Sampling Depth




**Sampling**

| $x_j$ | 52 | 37 | 90 | 56 | 58 | 45 | 66 | 75 | 36 | 100 | 68 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| j | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Sorting** | $x_{(i)}$ | 36 | 37 | 45 | 52 | 56 | 58 | 66 | 68 | 75 | 90 | 100 |
| **Ranking** | $R_{Pi}$=i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| | $K_{Pi}$ | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
| **Sample depth** | $H_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 5 | 4 | 3 | 2 | 1 |

***Representative sample*** gives a sufficiently complete view of the population involved.

***Sorting*** is putting a set of numbers into order (increasing or decreasing): sample $x_1, x_2, \ldots x_n, \ldots\ldots$ ordered sample $x_{(1)}, x_{(2)}, \ldots x_{(n)}$

***Ranking:*** we count sample elements from the smallest to the highest and the first of these yields the observattion's **upward rank** $R_{Pi} = i$, in general

$x_{(i)}$ has **upward rank i**

and counting down from the largest yields an observation's **downward rank** $K_{Pi} = n + 1 - i$,
in general

$x_{(i)}$ has **downward rank n + 1 -i.**

***Depth of the ith element*** in a sample is the smaller of its upward rank and its downward rank or it is a value how far it is from the low end or high end of the sample,

$H_i = \min (R_{Pi}, K_{Pi})$

***Cumulative (rank, order) probability*** is given by $\mathbf{P_i = i/(n + 1)}$ and then $\mathbf{P_i.100\%}$**-sample quantile** means the value of x below or at which 100 $P_i$% of the sample values lie.

**IMPORTANT QUANTILES** are for P = 25%, 50% and 75%

25th quantile (or percentile) is the **lower (first) quartile,** $\mathbf{F_L}$

50th quantile (or percentile) is the **median (second) quartile, M**

75th quantile (or percentile) is the **upperr (third) quartile,** $\mathbf{F_U}$

**LETTER VALUES are for** $\mathbf{P_i = 2^{-i}}$**, i = 1, 2, 3 ...**

For i = 1 $P_1 = 2^{-1} = 0.500$ is the 1st quantile or **median M**

For i = 2 $P_2 = 2^{-2} = 0.250$ is the 2nd quantile or **quartile F**

For i = 3 $P_3 = 2^{-3} = 0.125$ is the 3rd quantile or **octile E**

For i = 4 $P_4 = 2^{-4} = 0.062$ is the 4th quantile or **sedecile D**

**LOWER AND UPPER QUANTITY OF LETTER VALUES:**

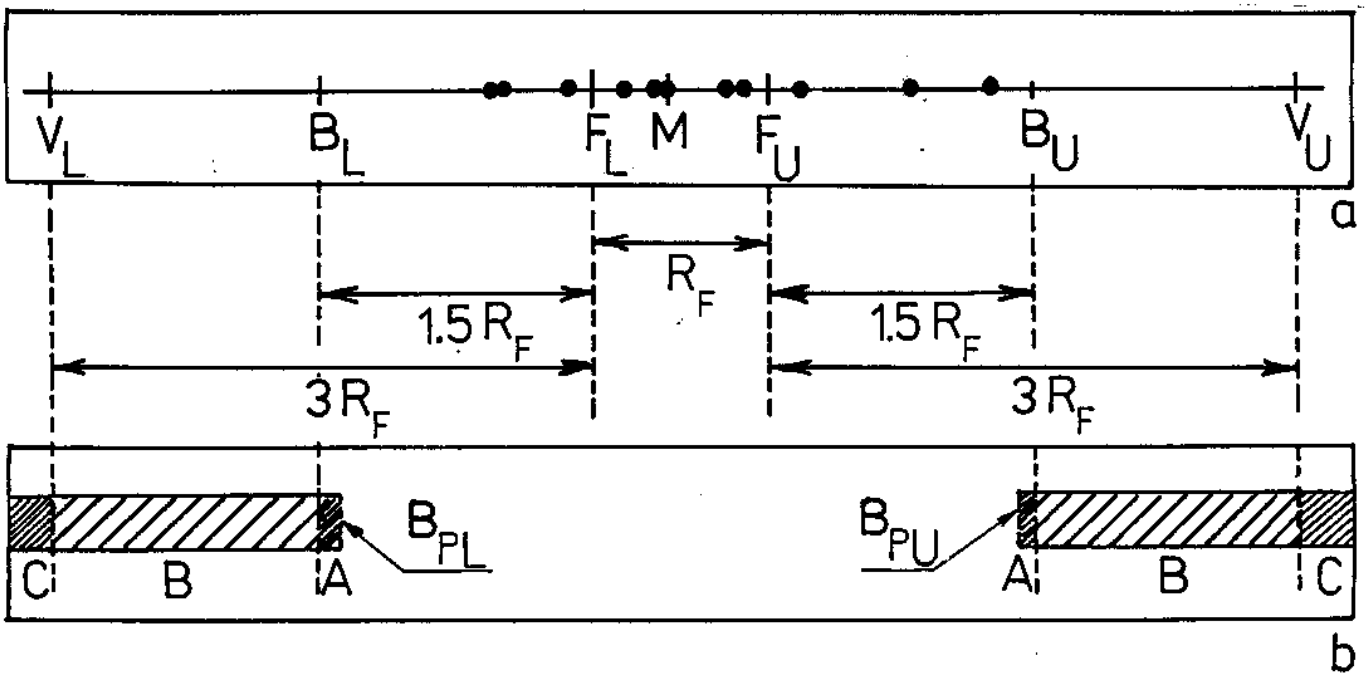
$P_1 = 2^{-1}$ is **median M** and no lower and upper quantity.

$P_2 = 2^{-2} = 0.250$ is the **lower quartile** $\mathbf{F_L}$ and 1 - 0.250 is the **upper quartile**

$P_3 = 2^{-3} = 0.125$ is **lower octile** $\mathbf{E_L}$ and 1 - 0.125 is the **upper octile.**

$P_4 = 2^{-4} = 0.062$ is **lower sedecile** $\mathbf{D_L}$ and 1 - 0.062 is the **upper sedecile.**

# Letter Values and Bounds




**QUARTILE RANGE:** $R_F = F_U - F_L$

**INNER BOUNDS:** $B_U = F_U + 1.5\ R_F$ upper inner bound

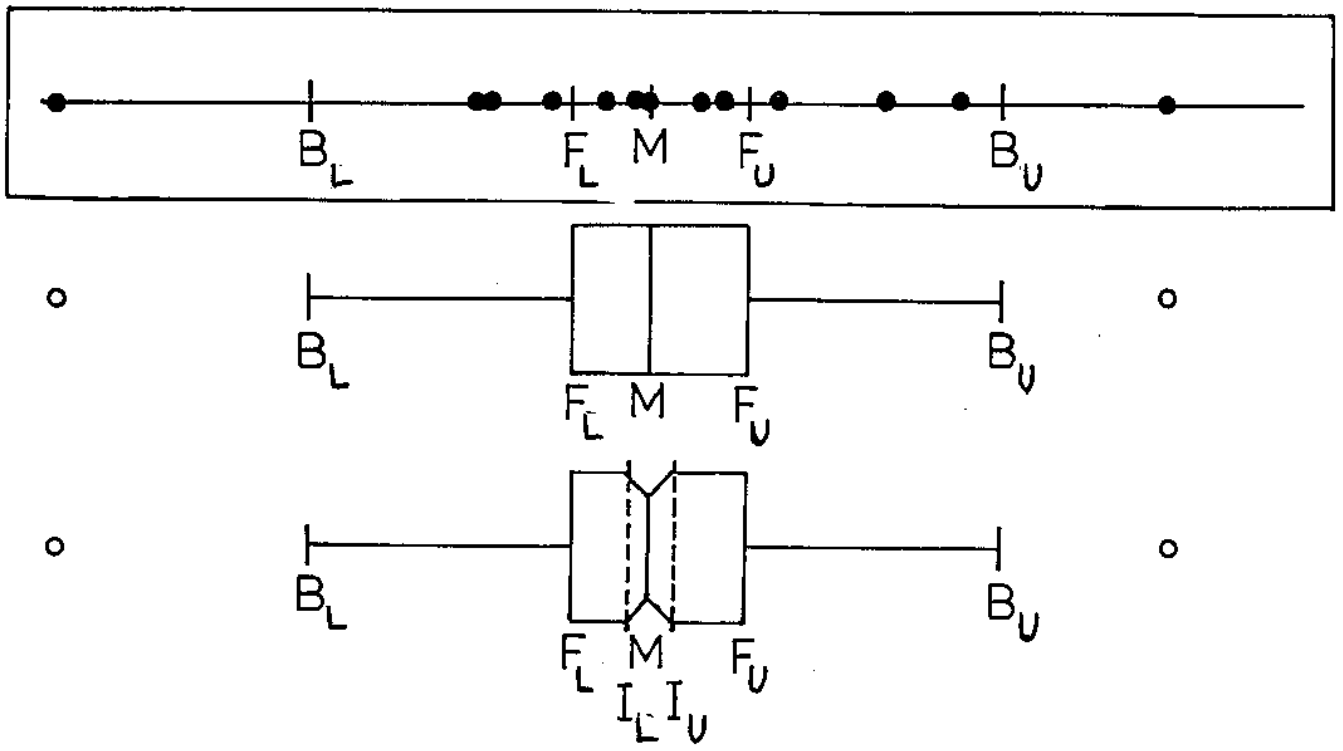
$B_L = F_L - 1.5\ R_F$ lower inner bound

**OUTER BOUNDS:** $V_U = F_U + 3\ R_F$ upper outer bound

$V_L = F_L - 3\ R_F$ lower outer bound

**OUTLIERS:** points outside interval $(V_L, V_U)$

# (Notched) Box-and-Whisker Plot




***x-axis:*** x values;

***y-axis:*** any suitable interval

**Diagnosis:**

(1) Robust estimate of median M;

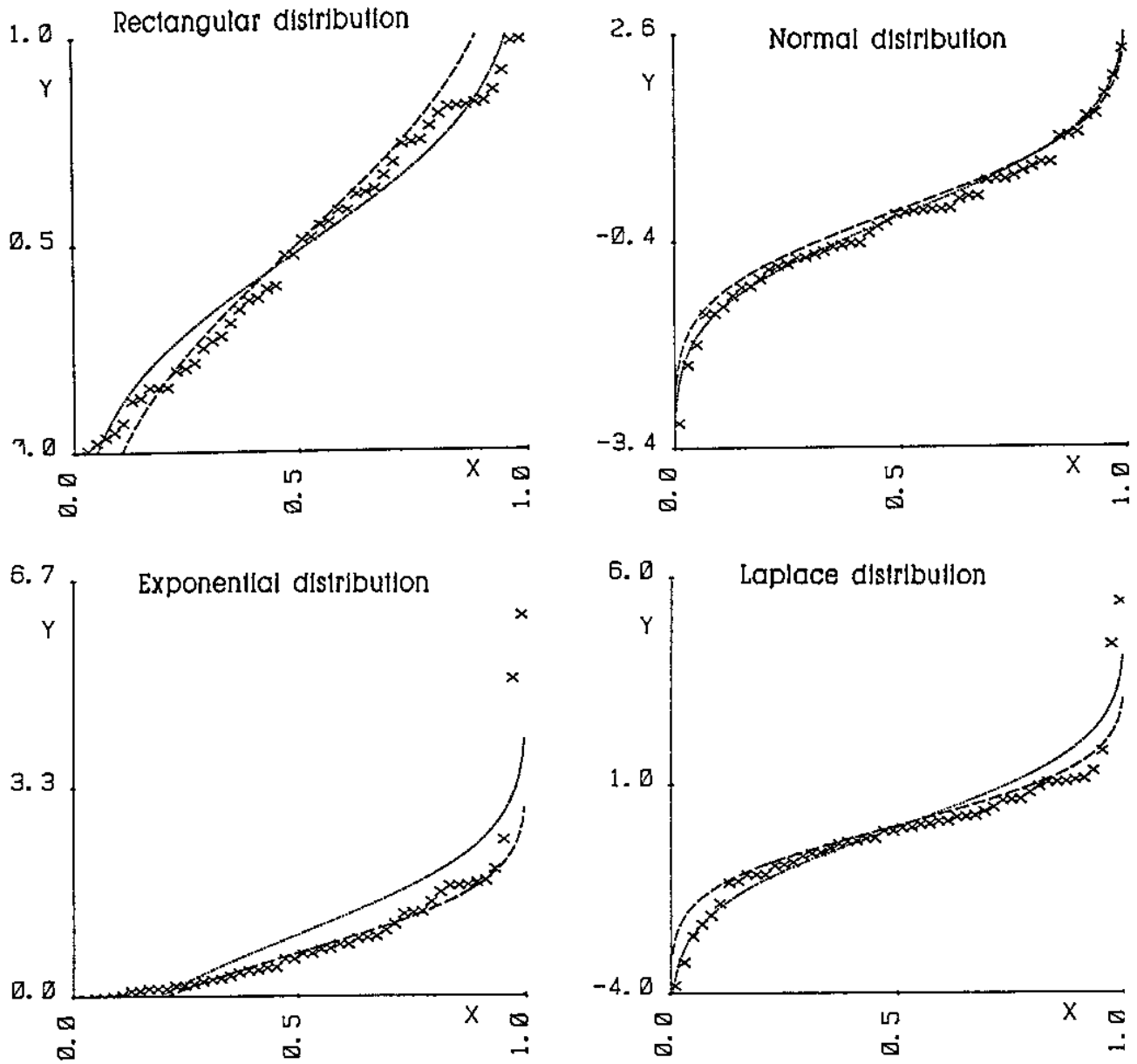
(2) Illustration of the spread and skewness of sample distribution;

(3) Examination of a symmetry and length of tails;

(4) Detection of outliers;

(5) Confidence interval of median by two notches $I_L < M < I_U$.

# Classical and Robust Quantile Plots




***x-axis:*** the cumulative (order) probability $P_i$,

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

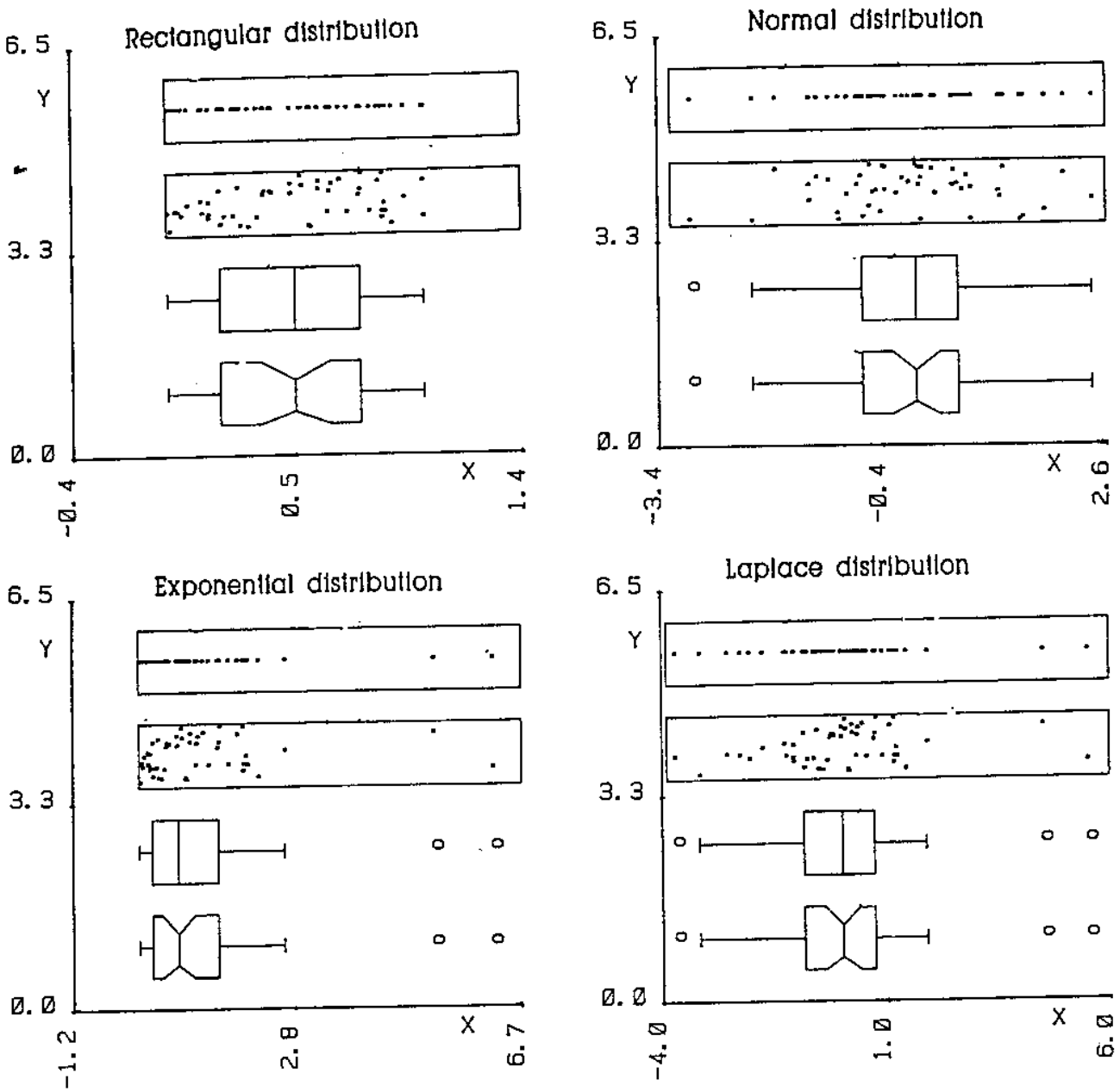
**Diagnosis:**

(1) Different tails length for R, N and E distributions;

(2) Significant skewing to lower values for E;

(3) For distribution with longer lengths (E) it can be not recognized because of outliers;

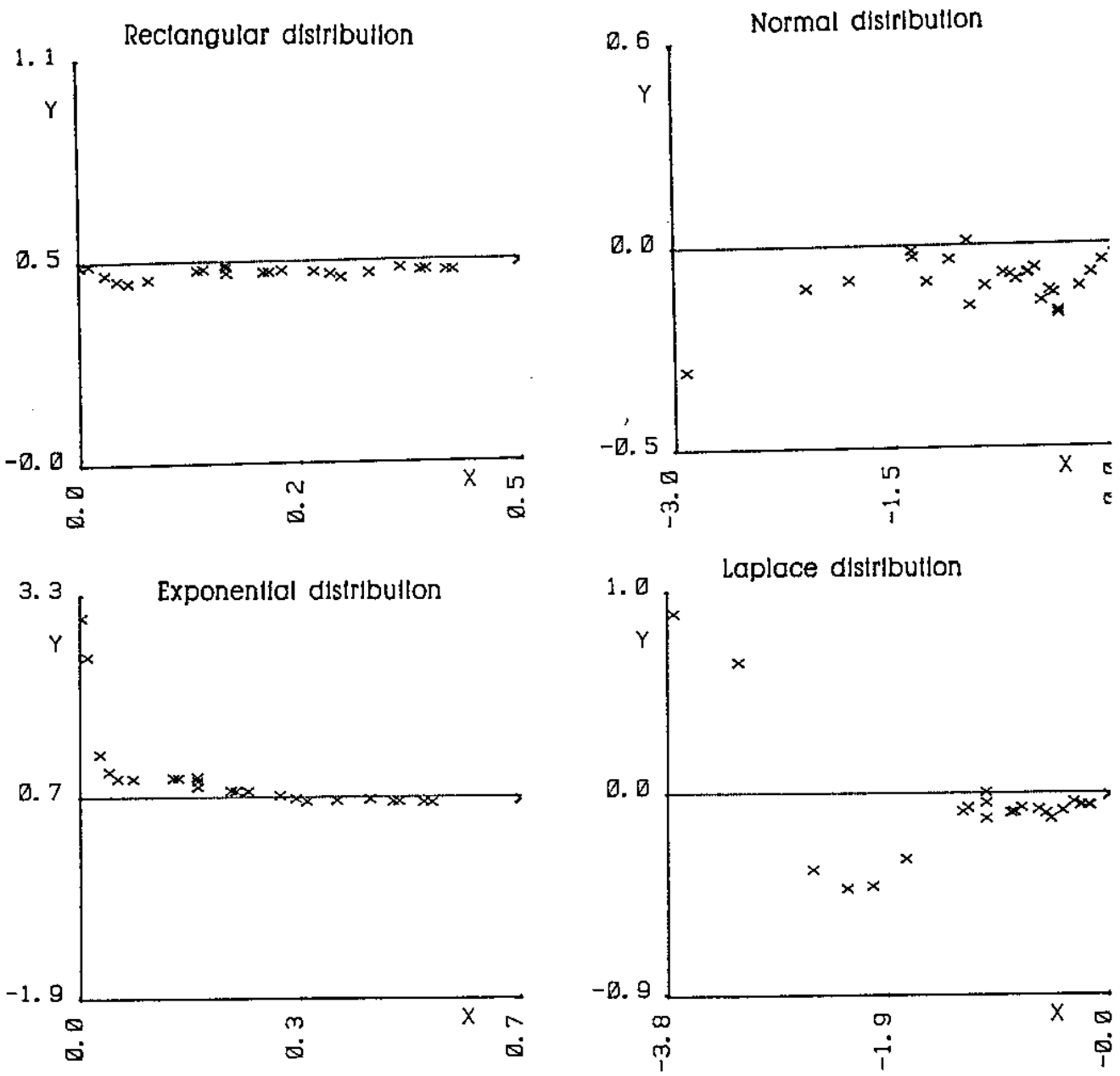
# Dot Diagrams and Box-and-Whisker Plots




## Diagnosis:

It shows asymmetry of sample distributions and outliers in data.

# Midsum Plot



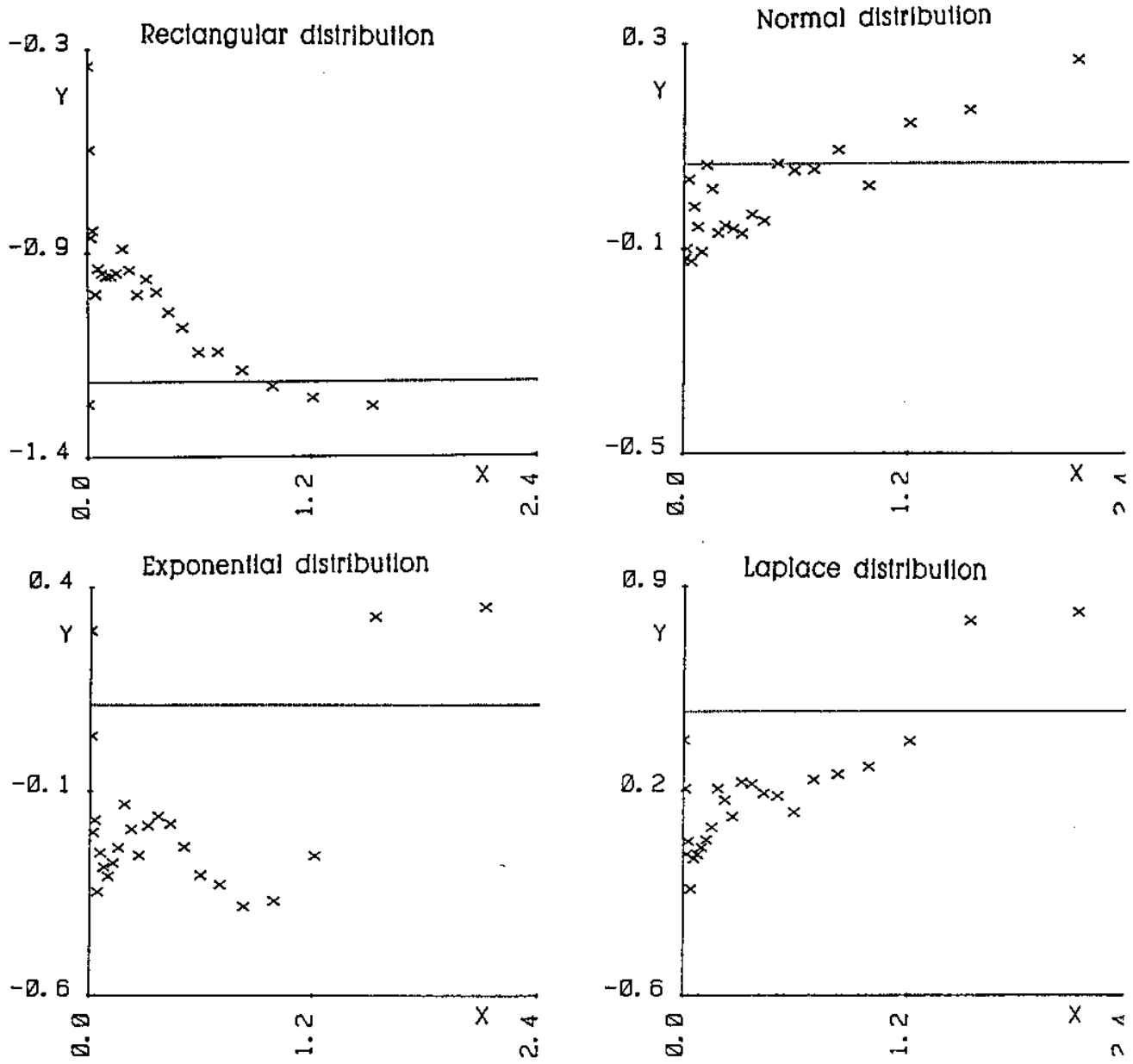

***x-axis:*** the order statistic $x_{(i)}$;

***y-axis:*** the midsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2$

**Diagnosis:**

For symmetrical distribution, the midsum plot forms a horizontal line $y = M$.

# Curtosis Plot




***x-axis:*** the quantile $u^2_{P(i)}/2$ for $P_i = i/(n + 1)$;

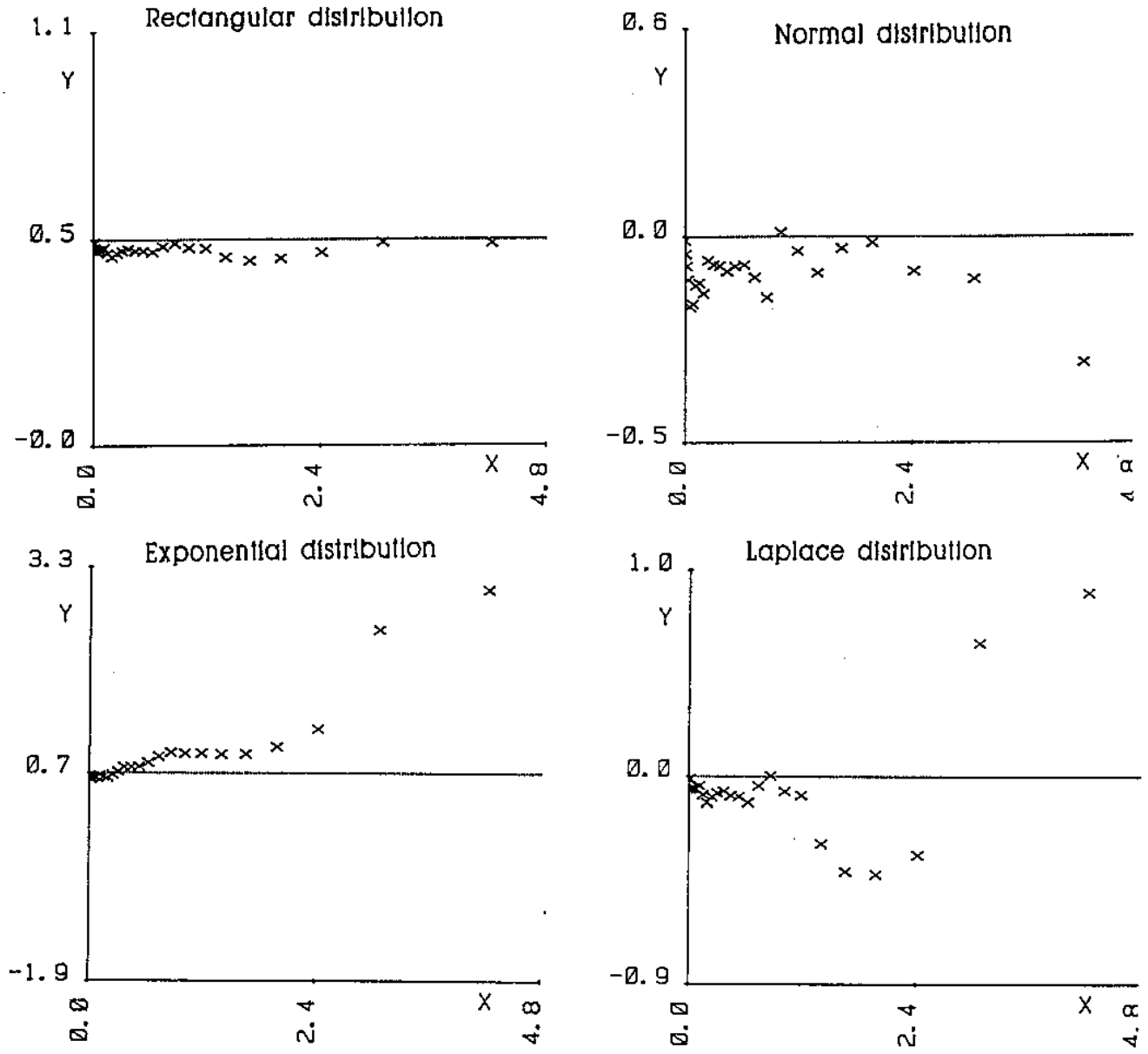
***y-axis:*** the quantity $\ln[(x_{(n+1-i)} + x_{(i)})/(-2u_{P(i)}]$

**Diagnosis:**

(1) For normal distribution, the kurtosis plot forms a horizontal line.

(2) When the line has non-zero slope, the slope gives an estimate of kurtosis.

# Symmetry Plot




***x-axis:*** the quantile $u^2_{P(i)}$;

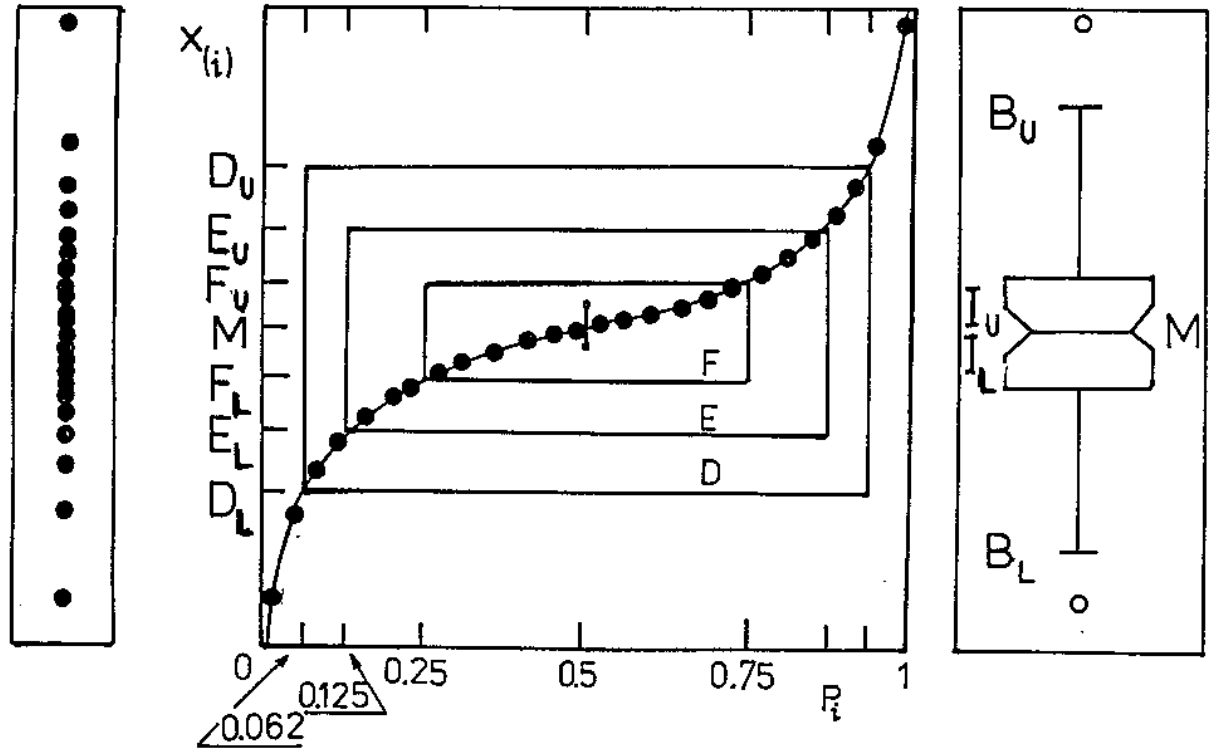
***y-axis:*** the midsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2$

**Diagnosis:**

(1) For symmetrical distribution, the symmetry plot forms a horizontal line $y = M$.

(2) For asymmetrical distributions, the line has non-zero slope, the slope gives an estimate of skewness.

# Construction of a Quantile-Box Plot




***x-axis:*** the cumulative (order) probability $P_i$,

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

**Diagnosis:**

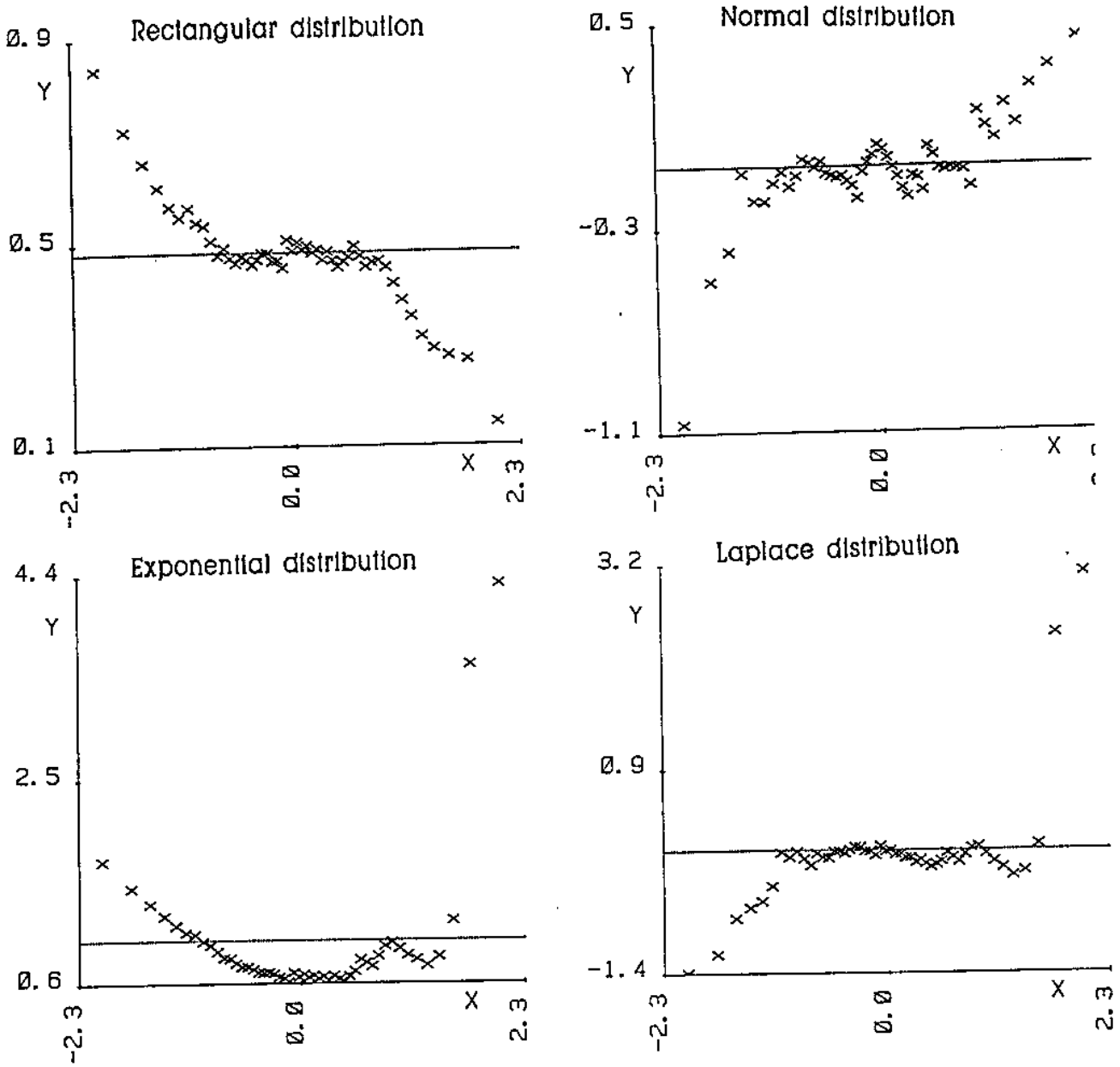
(1) A symmetric unimodal sample distribution contains individual boxes arranged inside one another.

(2) For an asymmetric distribution there are significantly shorter distances between the lower than between the upper parts of boxes.

(3) Outliers are indicated by a sudden increase of the quantile function outside the F box.

(4) A multimodal sample distribution is indicated by several parts of the quantile function inside box F reaching zero slope.

# Differential Quantile Plot



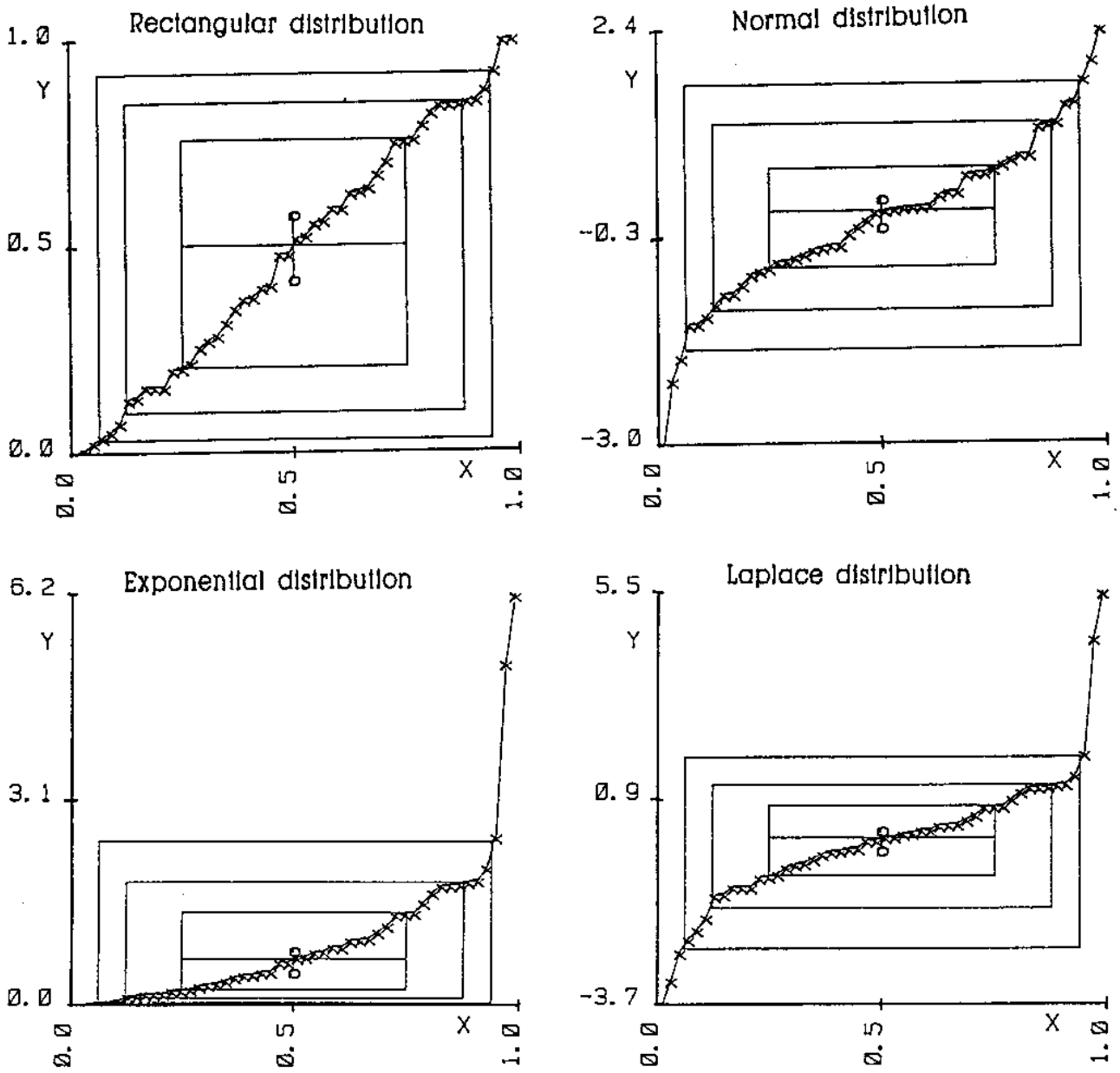

***x-axis:*** the quantile $u_{Pi}$;

***y-axis:*** the deviation of order statistics $d_{(i)} = x_{(i)} - s.\ u_{Pi}$.

**Diagnosis:**

It compares sample distribution with the normal one. A horizontal line indicates a symmetrical distribution with tails similar to the normla one.

# Quantile-Box Plot



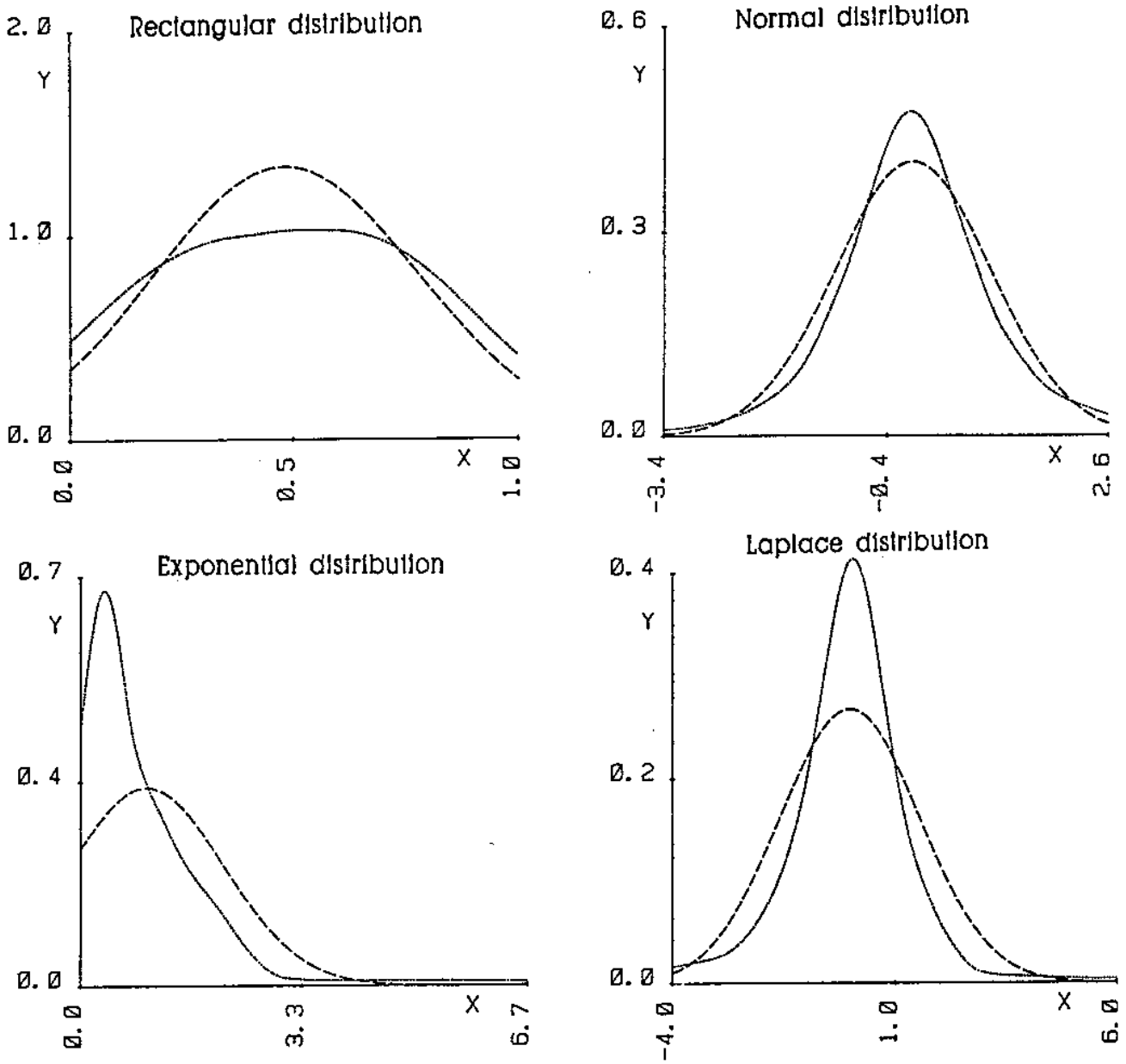

***x-axis:*** the cumulative (order) probability $P_i$,

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

**Diagnosis:**

Significant differences in the tail lengths of symmetrical distributions (R, M, L) and obvious skewing in the case of exponential distribution (E) can be observed.

# Plot of Probability-Density Function




***x-axis:*** the variable x;

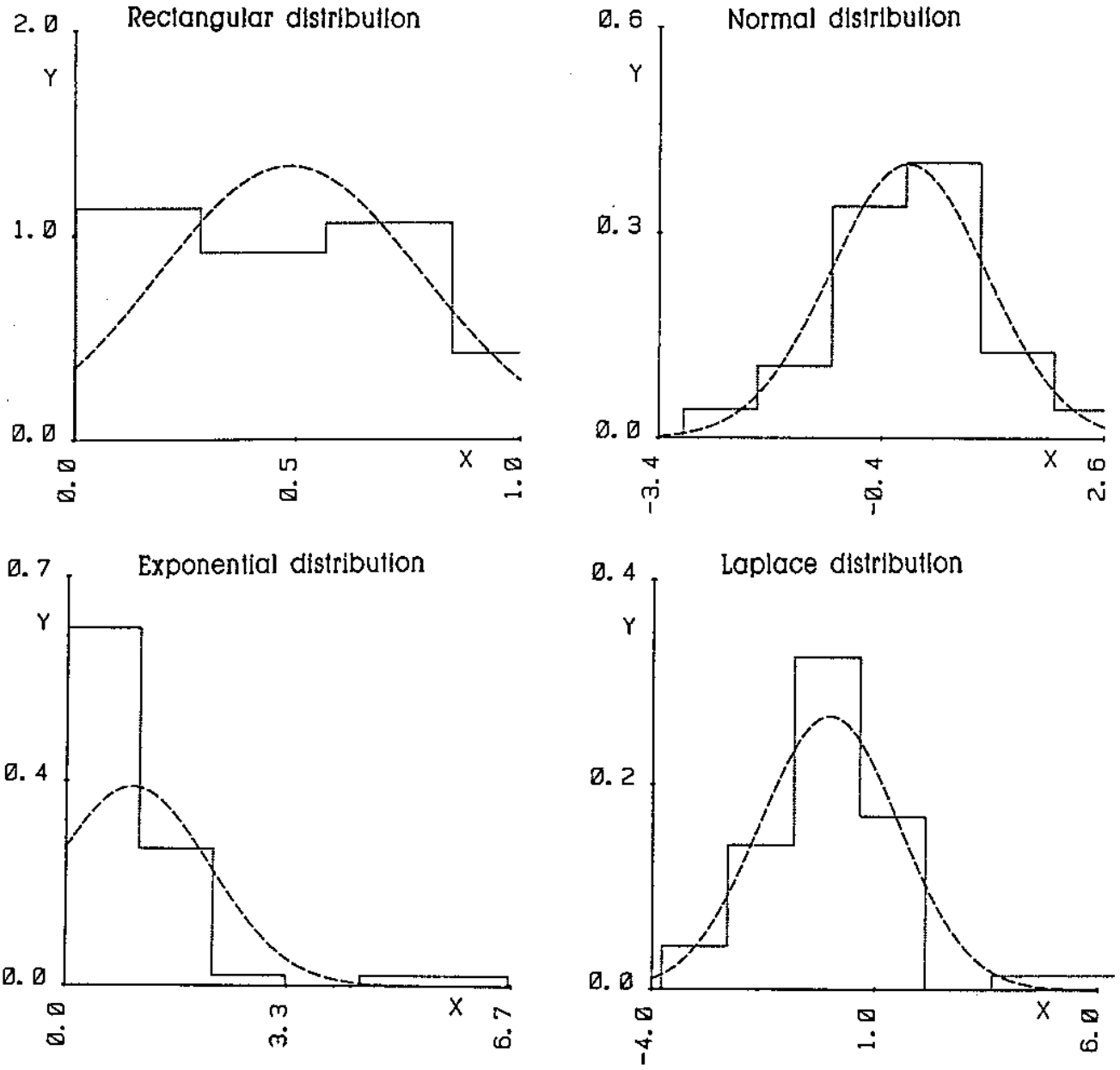
***y-axis:*** the probability density function PDF f(x);

curves: classical (Gaussian) PDF ----, robust (empirical) PDF .....

**Diagnosis:**

Kernel estimation of probability density PDF discovers an actual sample distribution. Classical and robust approaches are in good agreement.

# Histogram and Probability-Density Function




***x-axis:*** the variable x;

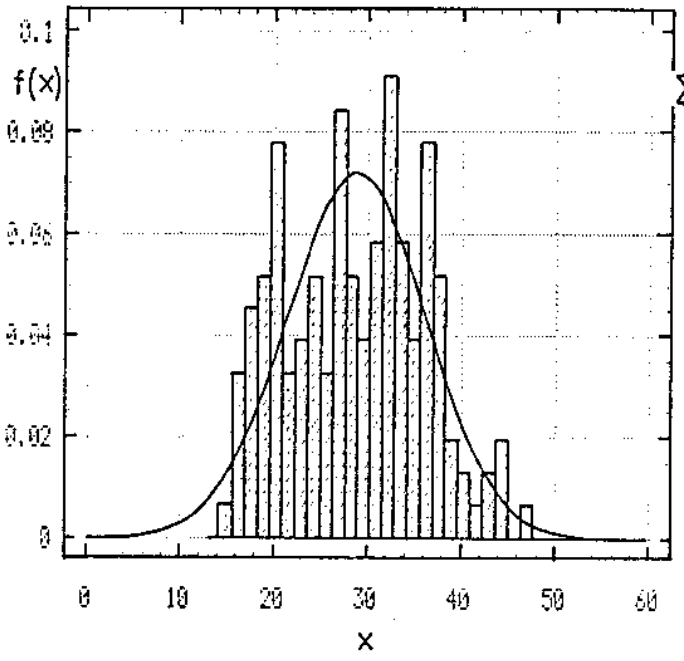
***y-axis:*** the probability density function PDF f(x);

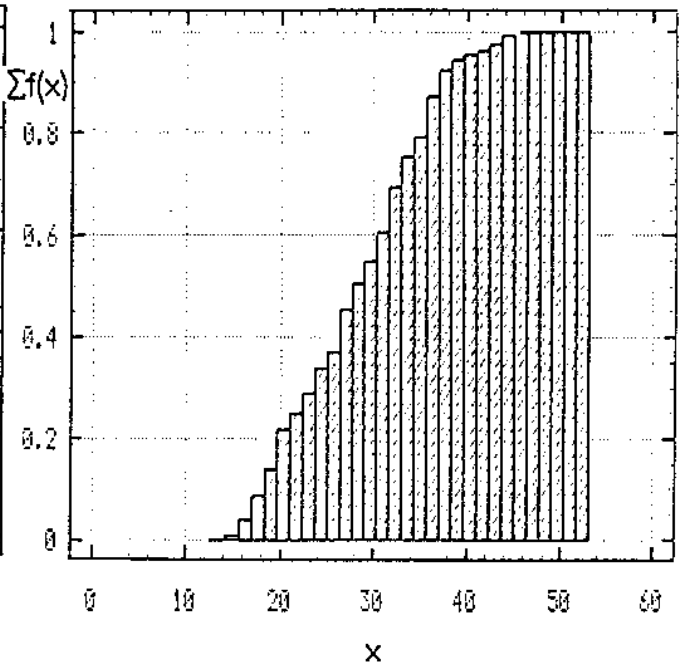
curves: classical (Gaussian) PDF ----, histogram .....

**Diagnosis:**

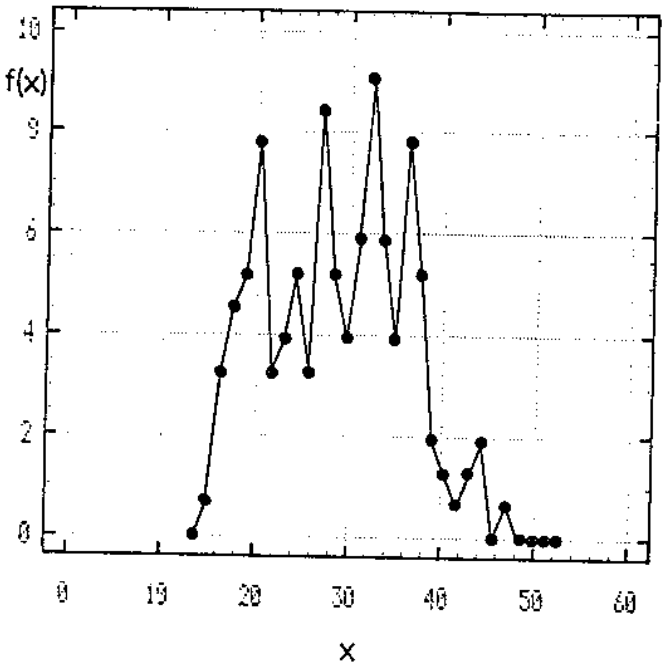
Histograms indicate quite obviously the type of distribution the sample was taken from.

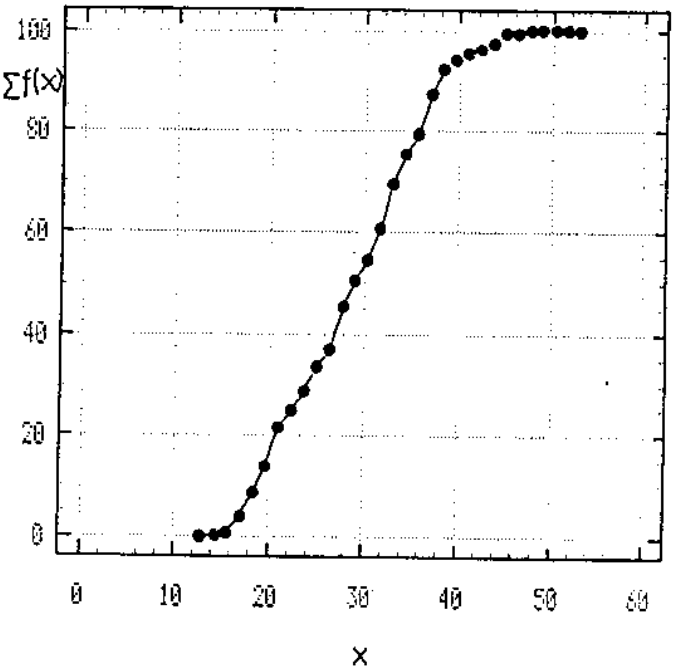
# HISTOGRAM



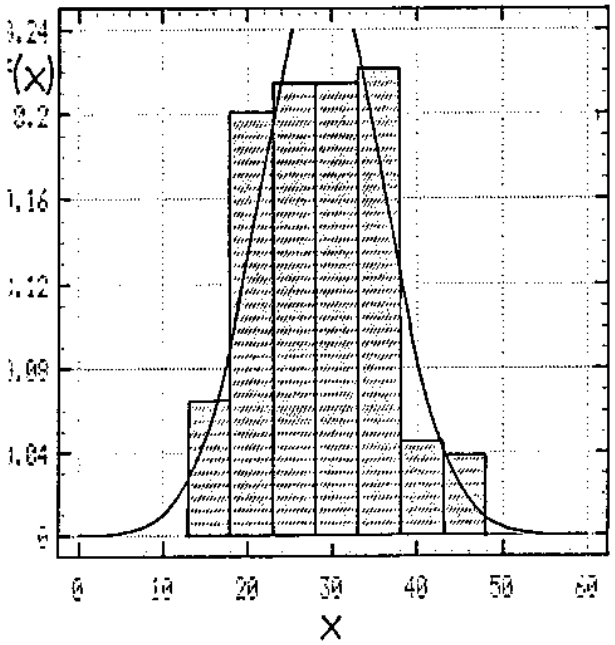



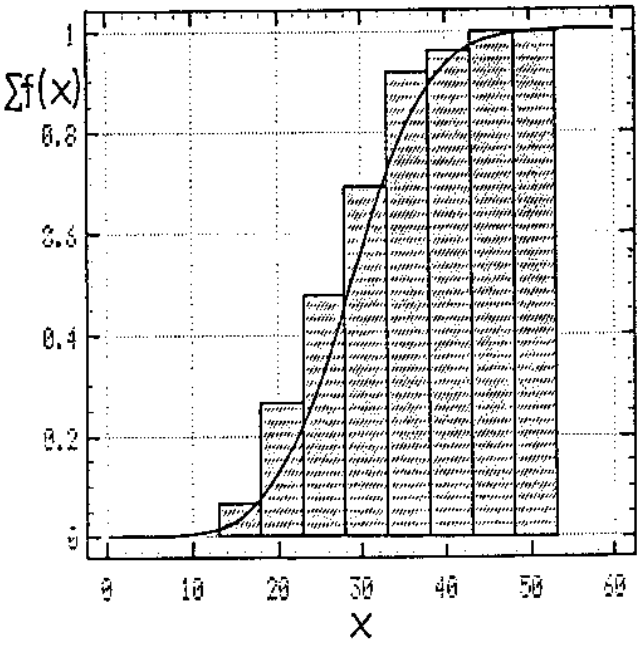

# FREQUENCY POLYGON



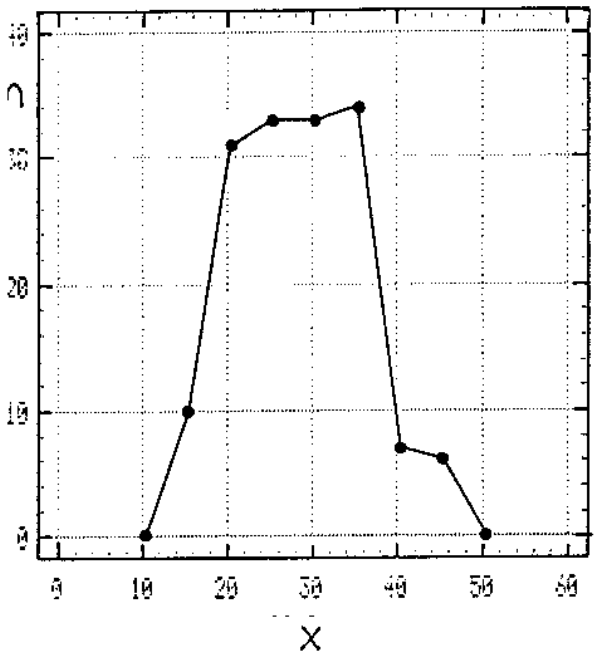



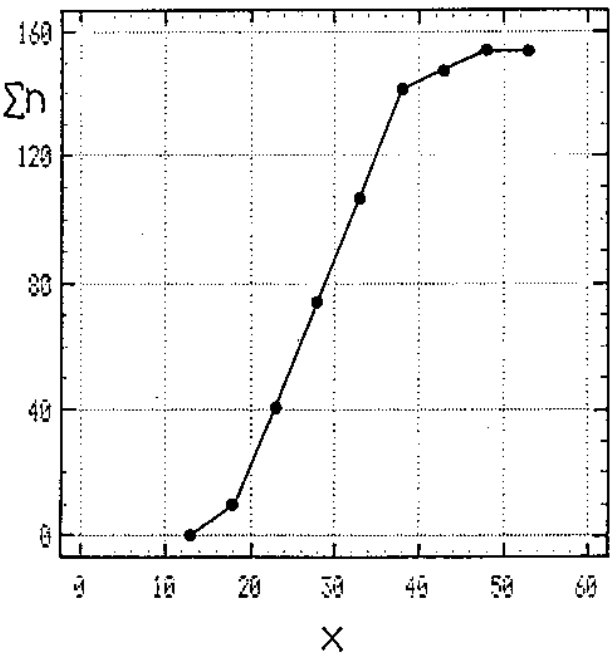
## HISTOGRAM



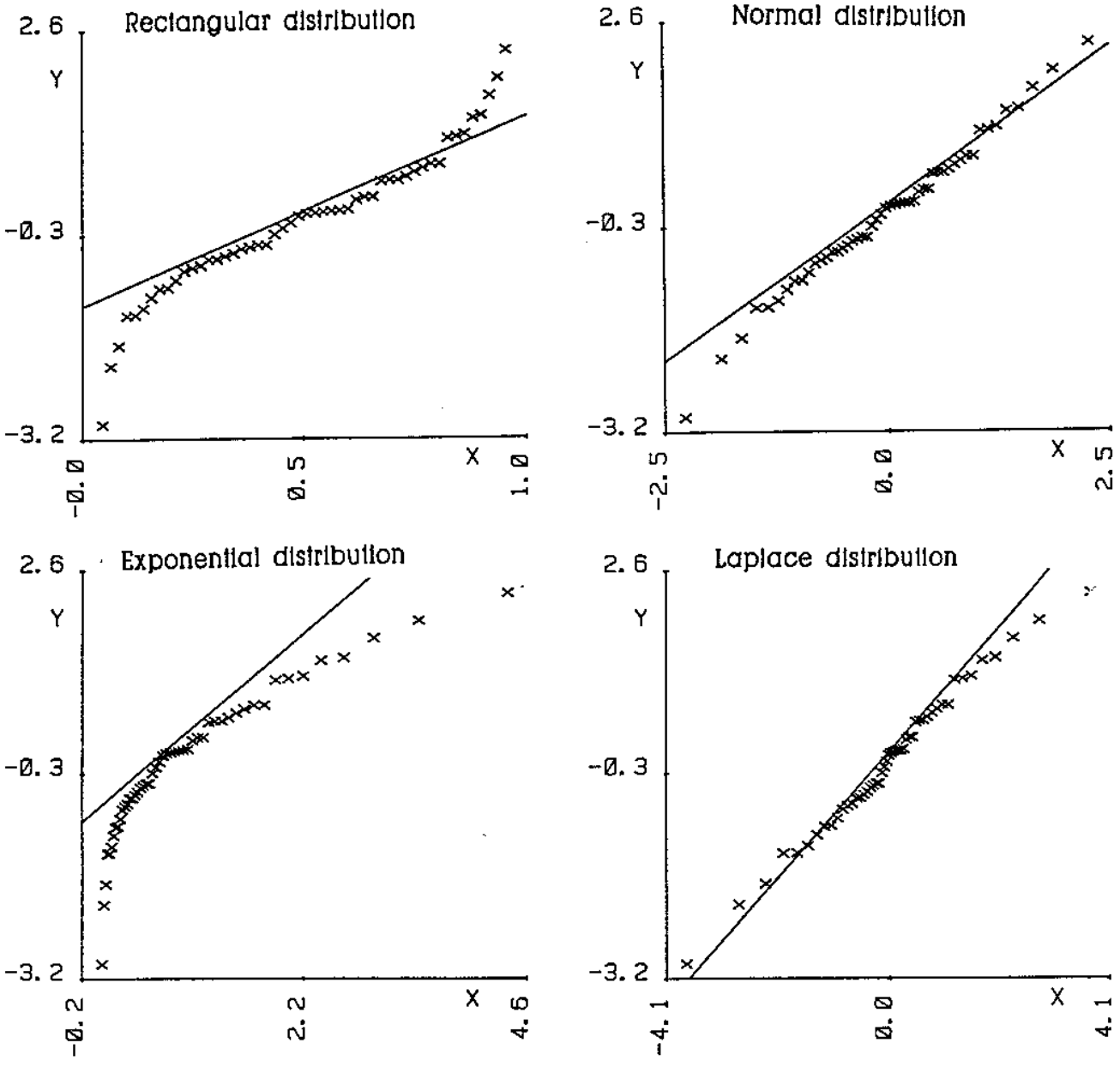




## FREQUENCY POLYGON



# Quantile-Quantile (Q-Q) Plot



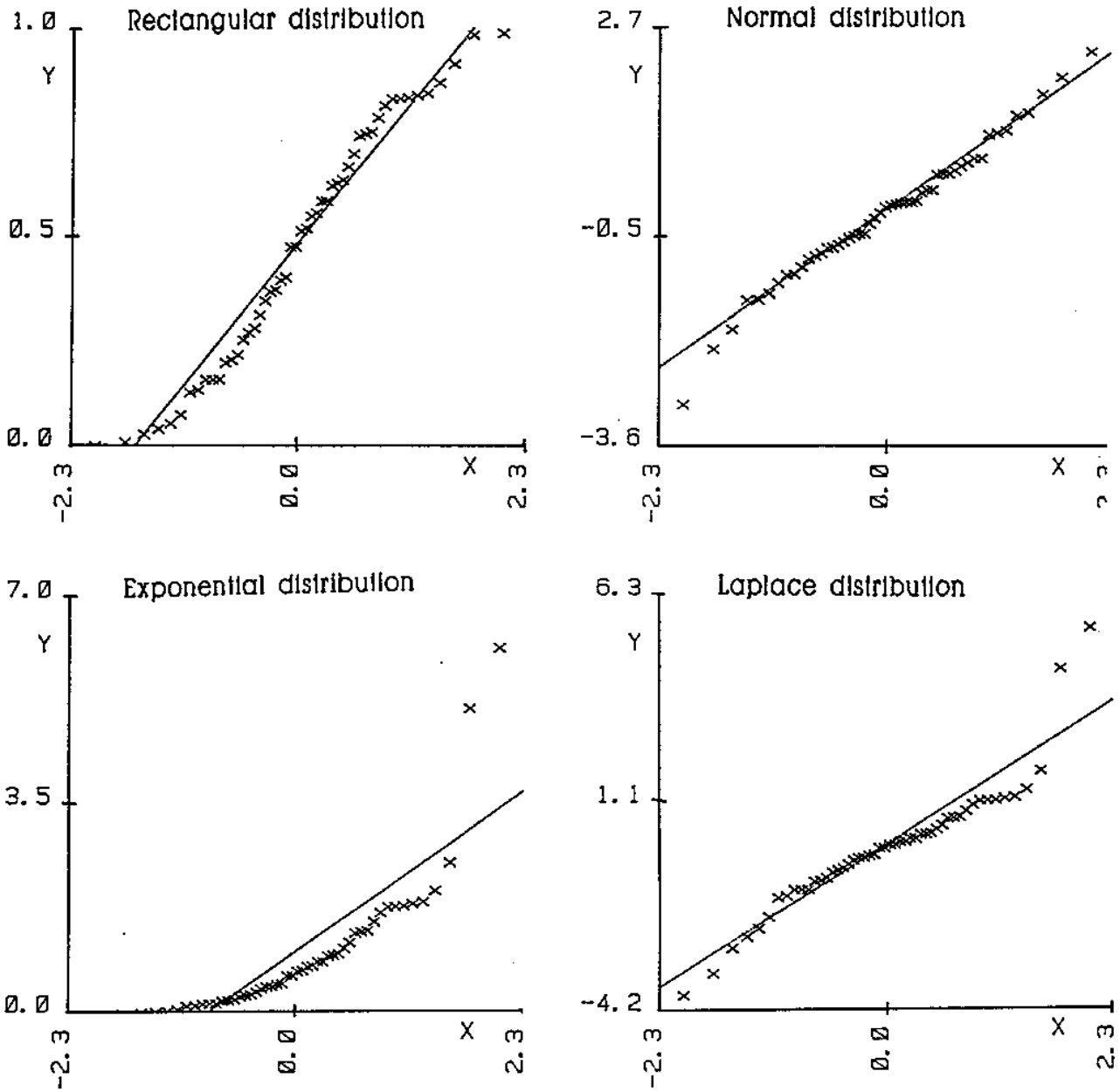

***x-axis:*** the quantile $Q_s(P_i)$,

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

**Diagnosis:**

Closeness of the sample to the given theoretical one helps to indicate an actual distribution.

# Rankit Plot



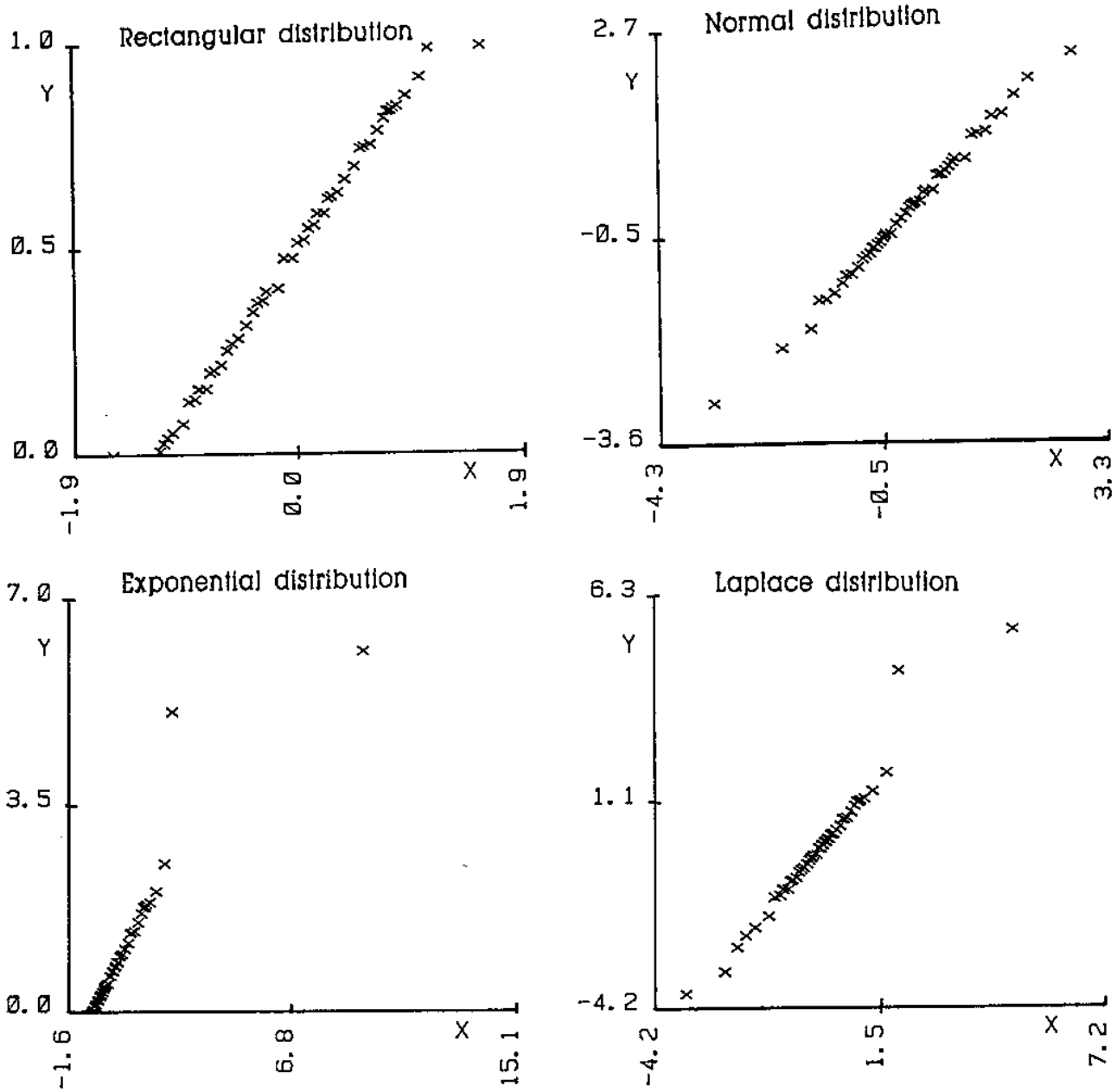

***x-axis:*** the standardized normal quantile $u_{Pi}$;

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

**Diagnosis:**

Plot enables classification of a sample distribution according to its skewness, kurtosis and tail length: a convex or concave shape indicates a skewed distribution, a sigmoidal shape indicates difference from the normal one.

# Conditioned Rankit Plot




***x-axis:*** the standardized quantile function of normal distribution $F[(U_{(i-1)} + U_{(i+1)})/2]$;

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

**Diagnosis:**

A linear plot proves the normal distribution.

## 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 concentraction [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.

***Solution:***

**E D A**

***(1) EDA: Diagnostic plots and displays:***

1. Quantile plot
2. Jittered dot diagram and Box-and-whisker plot
3. Midsum plot
4. Symmetry plot
5. Skewness plot
6. Kurtosis plot

***(2) EDA: Examining a sample distribution:***

1. Probability density plot
2. Quantile-box plot
3. Rankit plot
4. Conditioned Q-Q plot

***(3) EDA: Data transformation by Power and Box-Cox method:***

1. Hines-Hines plot
2. Maximum likelihood plot
3. Quantile plot for original data
4. Quantile plot for power transformation
5. Quantile plot for Box-Cox transformation
6. Q-Q plot for original data
7. Q-Q plot for power transformation
8. Q-Q plot for Box-Cox transformation

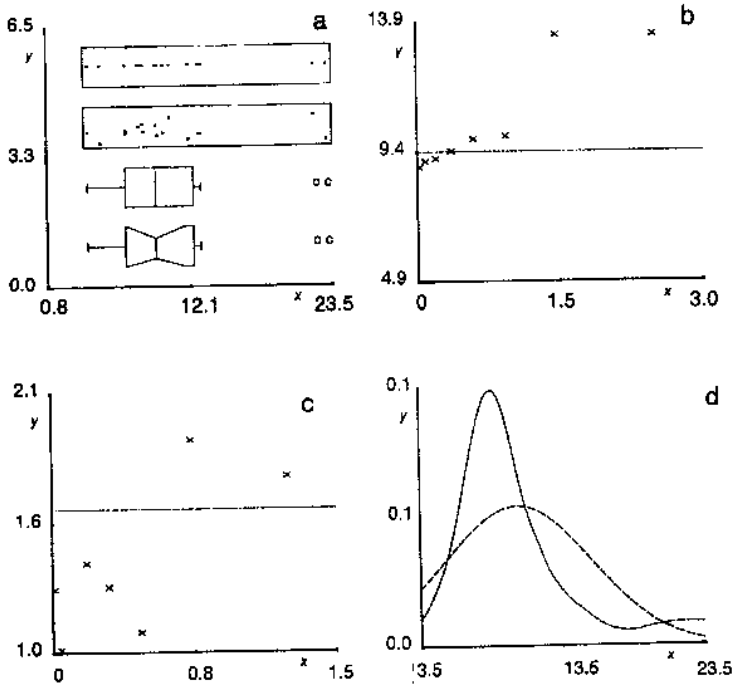
# C D A

*(1) CDA: Tests of basic assumptions about data*

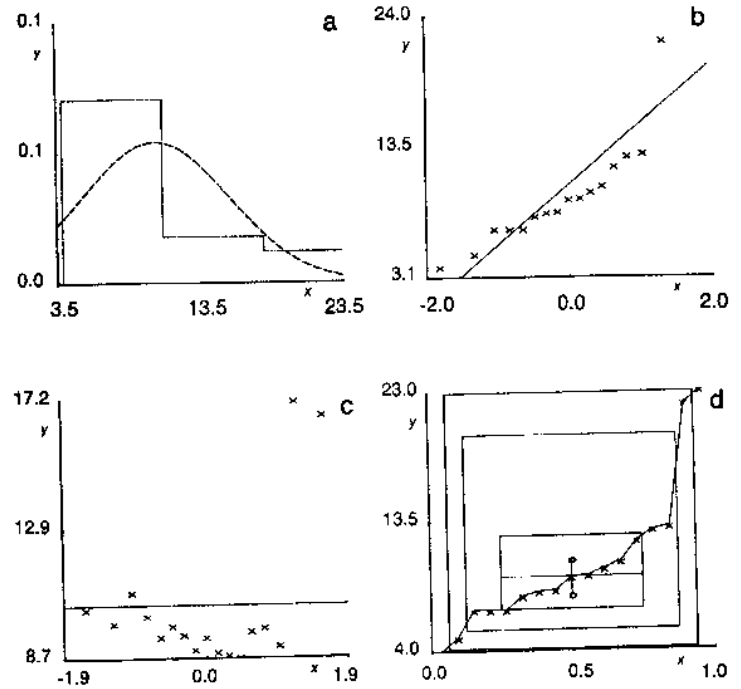
1. A test for minimal sample size
2. A test for independence of sample size
3. A test for homogeneity of sample
4. A test for normality

*(2) Point and interval estimates of location, spread and shape*

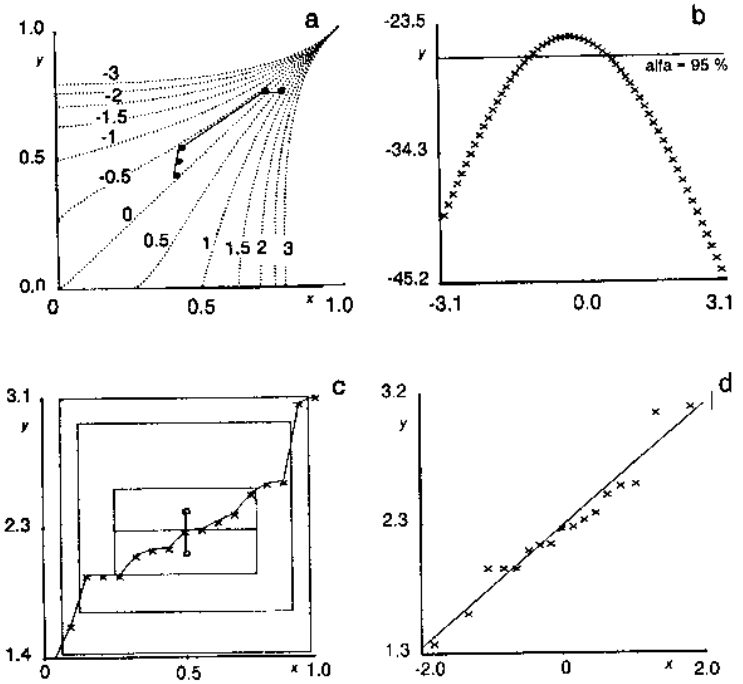
1. Classical statistics
2. Robust statistics
3. Adaptive statistics




Graphical Diagnostics of Exploratory Data Analysis:
(a) Dot Diagrams and Box-and-Whisker Plots,
(b) Symmetry plot, (c) Curtosis Plot,
and (d) Plot of Probability-Density Function




Graphical Diagnostics of Exploratory Data Analysis:
(a) Histogram, (b) Rankit Plot,
(c) Conditioned Rankit Plot, and (d) Quantile-Box Plot




Graphical Diagnostics of Exploratory Data Analysis:
(a) Hines-Hines Selection Graph,
(b) Graph of Logarithm of Likelihood Function,
(c) Quantile-Box Plot, and (d) Rankit Plot

Estimated measures of location, spread aand shape for the original data are $\bar{x}$ = 10.406, $s^2(x)$ = 26.834, $\hat{g}_1(x)$ = 1.399, $\hat{g}_2(x)$ = 4.272, and after logarithmic transformation ($\lambda$ = 0) $\bar{y}$ = 2.243, $s^2(y)$ = 0.203, $\hat{g}_1(y)$ = 0.304 a $\hat{g}_2(y)$ = 3.070 while after power transformation or Box-Cox transformation $\lambda$ = -0.27 it is $\bar{y}$ = 0.554, $s^2(y)$ = 0.043, $\hat{g}_1(y)$ = 0.048 a $\hat{g}_2(y)$ = 3.071.

Non-correct re-transformation leads to estimates $\bar{x}_R = \exp(\bar{x}^*)$ = 9.337 with the lower and upper 95% confidence limits $I_D$ = 7.742 a $I_H$ = 11.878 where $t_{0.975}(16)$ = 2.12. Rigorous re-transformation leads ro estimates $\bar{x}_R$ = 9.167 and $I_D$ = 8.272 a $I_H$ = 13.147.

# Conclusion

In low concentrations and trace analysis the statistical procedures of the mean value estimation contains the exploratory data analysis. Exploratory data analysis (EDA) isolates certain basic statistical features and patterns of data. The EDA techniques are quite effective for an investigation of statistical behavior of data from new or non-standard analytical procedures.

# References

M. Meloun, J. Militký, M. Forina: ***CHEMOMETRICS FOR ANALYTICAL CHEMISTRY, Volume 1: PC-Aided Statistical Data Analysis,*** and ***Volume 2: PC-Aided Regression and Related Methods***, Ellis Horwood Chichester 1992 and 1994.