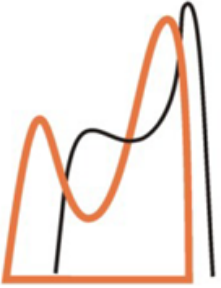


ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК

ЮЖНЫЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

Preliminary Analysis of Data

Stochastic modeling and statistical data processing

Kurbatova Natalia Victorovna

nvkurbatova@sfedu.ru

Methodical material

- Theoretical material: Lecture 2 §§ 1-3:
- Data source: crabs.xlsx, DataForAnalysis.mat
- Examples (MatLab scripts):

  InputOutput.m, preprocessing.m, StatPlot.m
- Statistics Toolbox in MatLab: StMatLabHelpStatistics.pdf

  Table methods and functions.pdf

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Current Goal:

Determine the nature of the data, using the functionality of applied packages (MatLab, Maple, R-lang, Python) that is

- present statistical plots: scattering diagram (outliers), bar, histograms, glyph(for visualizing multivariate data) and so on
- define descriptive statistics: mean, median, variance, standard deviation, high order moments
- calculate consistency measures: correlation coefficient, covariance coefficient, conjugacy*
- use statistical criteria on estimation statistical characteristics

* conjugacy[ˈkɒndʒʊgəsɪ] сопряженность

Experiment

It should be ensured:

- correctness of the experiment – i.e. regularity of feedback (control): measurement, weighing, etc.
- repeatability;
- the same initial conditions.

The main problems of database formation:

- various data storage formats;
- different units of measurement;
- irregularity of collection;
- missing data;
- outliers.

Ensured [ɪnˈʃʊəd] гарантированный

Forms of accumulation, storage and exchange of Data

- Table: Excel or other formats of full Data (*xlsx; *txt, *mat, … etc.)
- Matricies: including mixed data
- Matlab converters: builtin table
- I/O: data handling

* See script InputOutput.m

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Example of data representation, crabs.xlsx:

y	color	spine conditio	width	sattelts	weight	catwidth
1	medium	bothworn	28,3	8	3,05	28,75
0	darkmed	bothworn	22,5	0	1,55	22,75
1	lightmed	bothgood	26	9	2,3	25,75
0	darkmed	bothworn	24,8	0	2,1	24,75
1	darkmed	bothworn	26	4	2,6	25,75
0	medium	bothworn	23,8	0	2,1	23,75
0	lightmed	bothgood	26,5	0	2,35	26,75
0	darkmed	oneworn	24,7	0	1,9	24,75
0	medium	bothgood	23,7	0	1,95	23,75
0	darkmed	bothworn	25,6	0	2,15	25,75
0	darkmed	bothworn	24,3	0	2,15	24,75
0	medium	bothworn	25,8	0	2,65	25,75
1	medium	bothworn	28,2	11	3,05	27,75
0	dark	oneworn	21	0	1,85	22,75
1	medium	bothgood	26	14	2,3	25,75
1	lightmed	bothgood	27,1	8	2,95	26,75
1	medium	bothworn	25,2	1	2	24,75
1	medium	bothworn	29	1	3	28,75

Data Features:

- One-dimensional and multivariate Data
- Data of different volumes
- Data having different units of measurement
- Collinear Data
- Mixed Data

dichotomic[daɪkə(ʊ)ˈtɒmɪk]- дихотомический, nominal data,

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Exploratory Analysis of Data

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

- Data Centering and Standardization
- Analysis of missing data and correction of the database
- Visual data analysis using graphical functions
- Analysis of outliers
- Deleting collinear and deterministic data (multivariate case)
- Determination the basic statistical characteristic

preliminary [prɪˈlɪmɪnərɪ]-предварительный

Data generators

For computer modeling, some data generators can be used:

rand - uniform distribution over the interval [0, 1];

randi – integers data, uniform distributed over the interval [a, b];

randn - normal standard distribution;

chi2rnd – of Chi-square distribution;

frnd - Fisher;

trnd - Student, etc.

Data preprocessing

- Centering :

  $Y=X-\mu$,

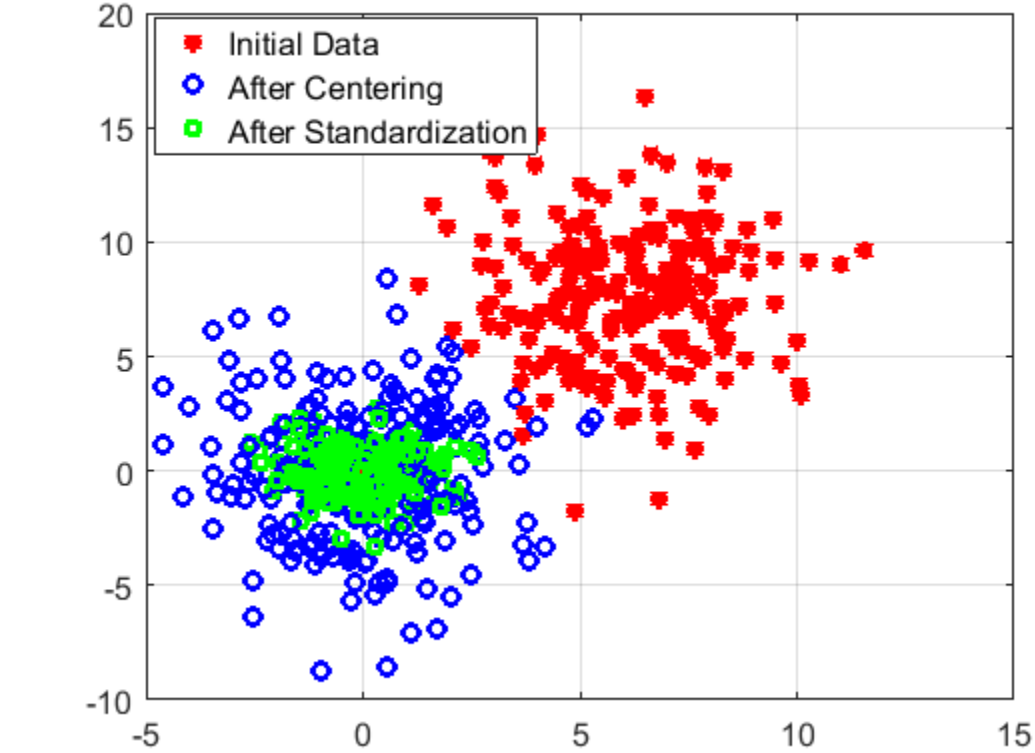
$X=(X_i)$, $\mu=mean(X)$

- Standardization:

  $Y=(X-\mu)/\sigma$,

σ - standard deviation

σ^2 - variance



*Let's try look at file preprocessing.m

Example of preprocessing

Mini-Task

Generate two vectors X and Y of the same volume. Perform data after centering and standardization. Visualize the points (Xi,Yi) on the same axes before and after preprocessing.

Who can, try to do this task yourself in Matlab or open the preprocessing.m script

Measures of Central Tendency in MatLab

Function Name	Description
geomean	Geometric mean
harmmean	Harmonic mean
mean	Arithmetic average
median	50th percentile
mode	Most frequent value

Mean, median [ˈmiːdɪən]

Percentile [pəˈsentaɪl]

arithmetic average – [əˈrɪθmətɪk ˈævərɪʤ]

grpstats – important MatLab function - Summary statistics
See StMatlabHelpStatistics.pdf , §2, p.14 (saved in teams)

Statistical Plots

- boxplot, normplot
- histfit, histogram
- bar

For multivariate data:

- scatter, scatter3,
- plotmatrix
- glyphplot

Qurtiles Q_i

Q_1 – the first quartile , also known as the lower quartile basically separates the lowest 25% of data from the highest 75%.

Q_2 – the second quartile or the middle quartile also the same as the median it divides numbers into 2 equal parts

Q_3 – the third or the upper quartile separates the highest 25% of data

! Quartile divides a set of observations into 4 equal parts!

quartile[ˈkwɔːtaɪl], percent[pəˈsent]

Box & Whisker

MatLab function: boxplot

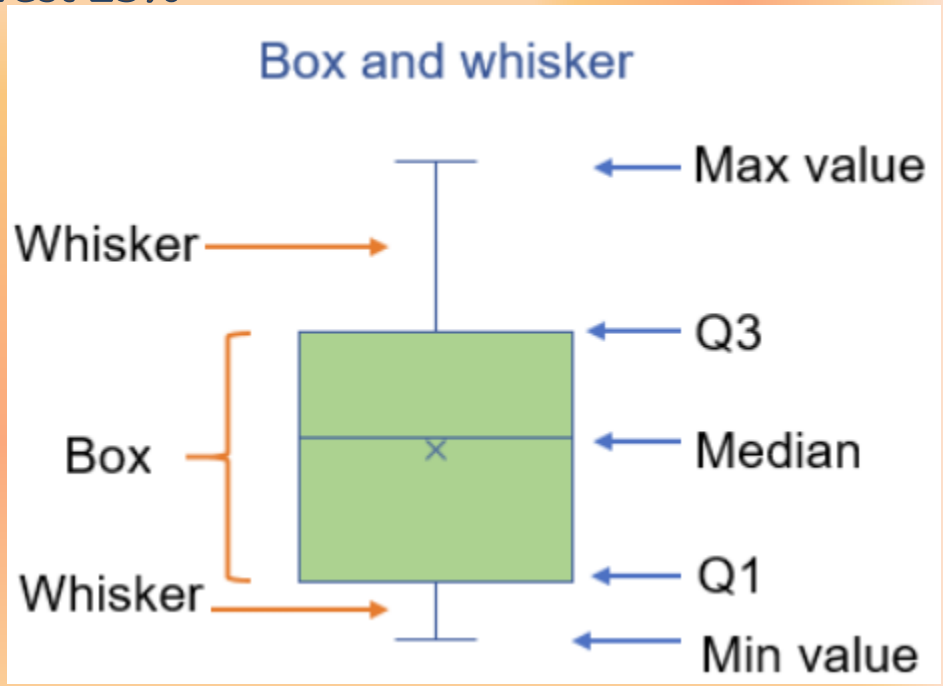
Boxplot makes it possible to estimate:

Q_1 – the first quartile (separates the lowest 25% of the data);

Q_2 – median;

Q_3 – the third quartile (separates the highest 25% of data)

I – Interquartile, $I=Q_3-Q_1$;



and as consequence ⟹

We can draw conclusion:

- on symmetry-the skewness of the data distribution;
- to discover the presence of outliers;
- MaxValue, MinValue
- Range = (MaxValue - MinValue)

Outliers are values outside the range:

MinValue= $Q_1 - (Q_3 - Q_1)1.5$

Error ML: MaxValue= $Q_3 + (Q_3 - Q_1)\underline{1.5}$

Box & Whisker

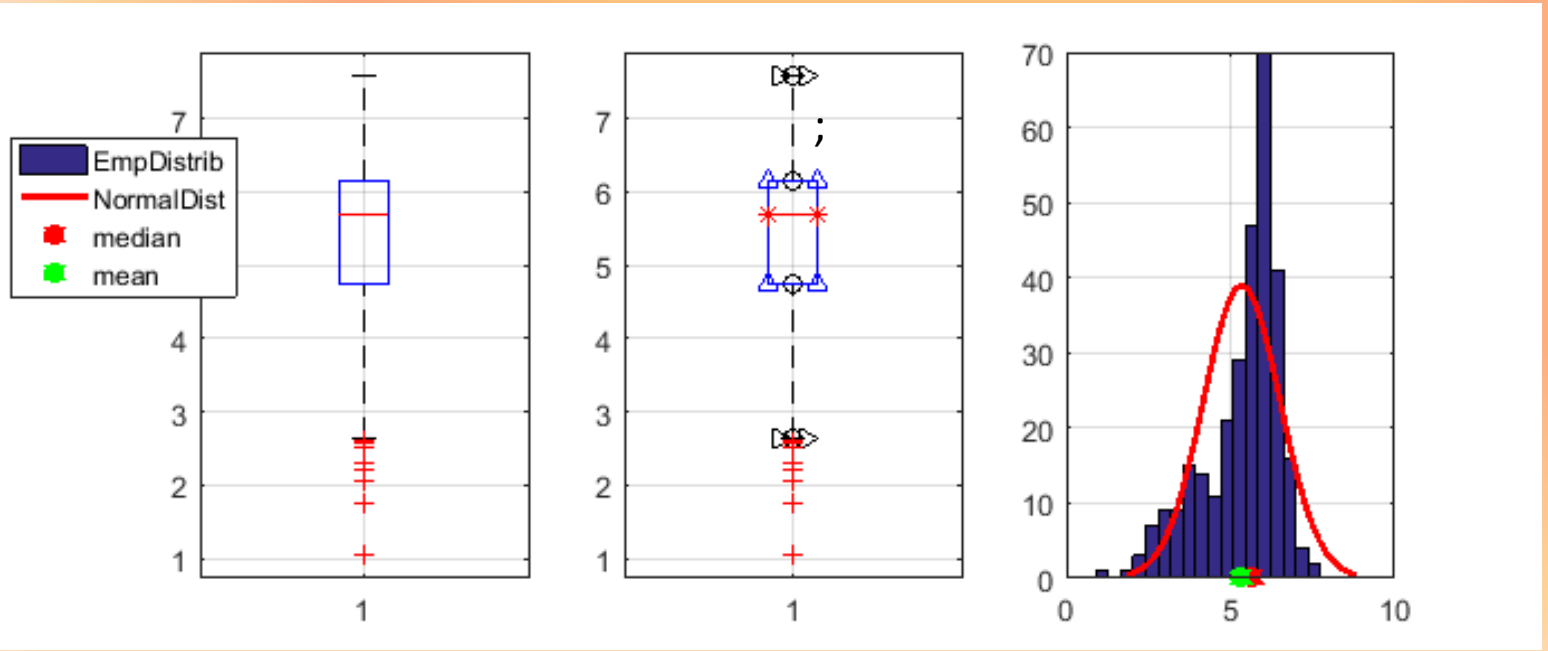
boxplot function syntax

% Compute the quantiles of the sample data x:

p = 0:0.25:1; y = quantile(x,p); z = [p;y]

Q_1 – [p(2),y(2)]; Q_2 – [p(3),y(3)]; median;

Q_3 –[p(4),y(4)]; I – Interquartile, I=Q_3 - Q_1;



For details, see the file StatPlot.m;

Measures of Dispersion in MatLab

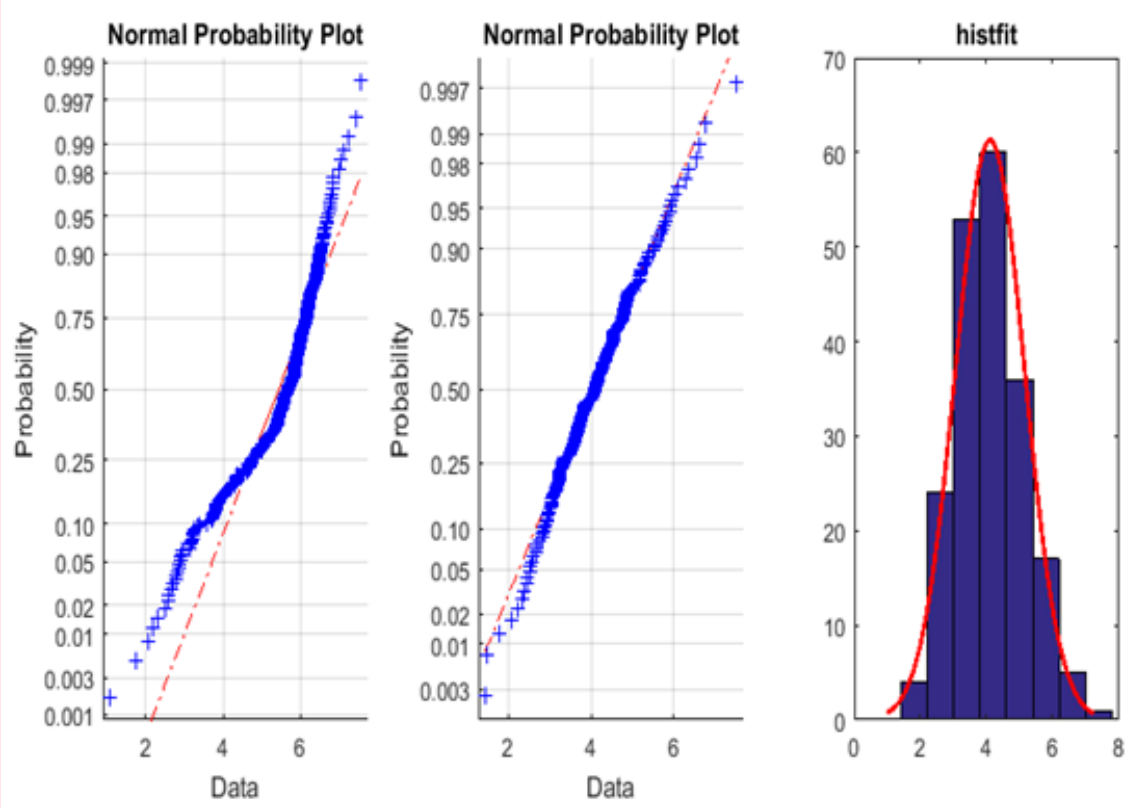
Function Name	Description
`iqr`	Interquartile range $Q_3 - Q_1$
`mad`	Mean absolute deviation
`moment`	Central moment of all orders
`range`	Range
`std`	Standard deviation
`var`	Variance

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Normal Probabilistic Plotter

normplot function




```
figure
x =[normrnd(4,1,1,100),...
      normrnd(6,0.5,1,200)];

subplot(1,3,1), normplot(x)

z=normrnd(4,1,1,200);
subplot(1,3,2), normplot(z)

subplot(1,3,3), histfit(z,8)
title('histfit')
```

Let G – theoretical distribution; $\hat{G}$-empirical distribution;
G^{-1}(P)=x; G^{-1} · $\hat{G}$ (x)≈x if $\hat{G}$ close to G; or G^{-1} · $\hat{G}$ ≈ I (identical mapping)

Empirical distribution

The histogram
as an empirical distribution
is presented in Lecture 2 § 3.

You should disassemble this material yourself!

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Histogram as empirical distribution

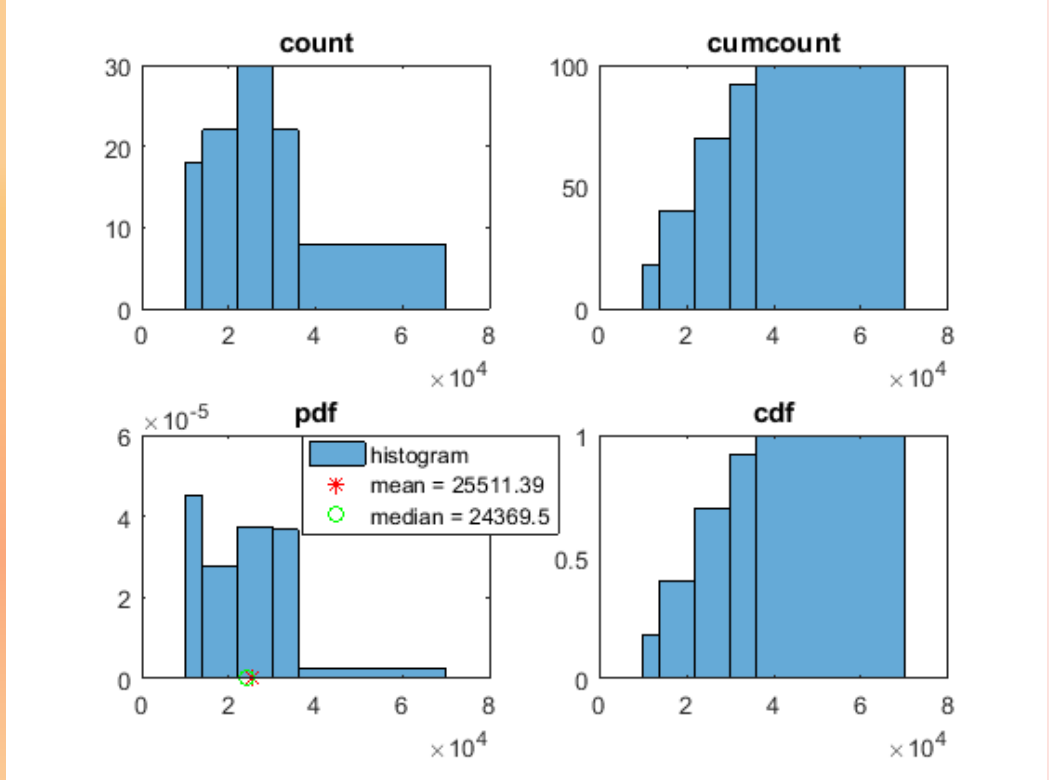
ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

histogram –
important ML function!

Example.
X – sample, 6 bins , the same width:
h=histogram(X,6)

histogram
provides the following features:

1) Option 'Normalization' takes values:
'count', 'countdensity', 'cumcount',
'probability', 'pdf', 'cdf'
histogram(X, 'Normalization', 'pdf')



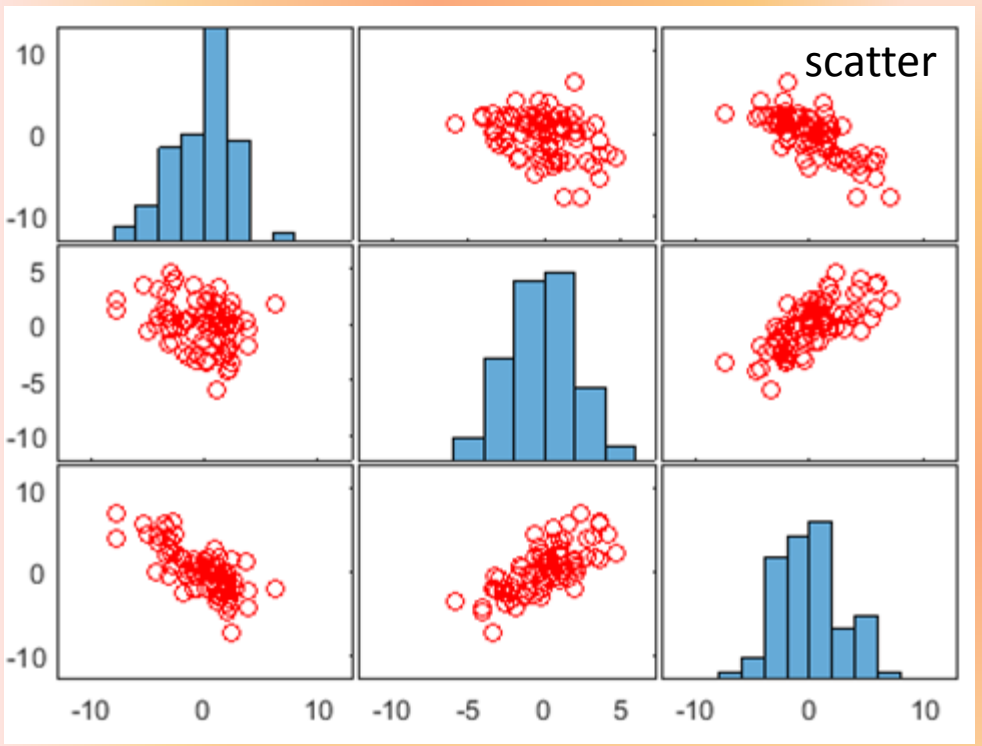
2) Option 'BinEdges' can take given values,
boundary points – vector BinsPoints of their values:
h=histogram(X,6, 'BinEdges', BinsPoints)

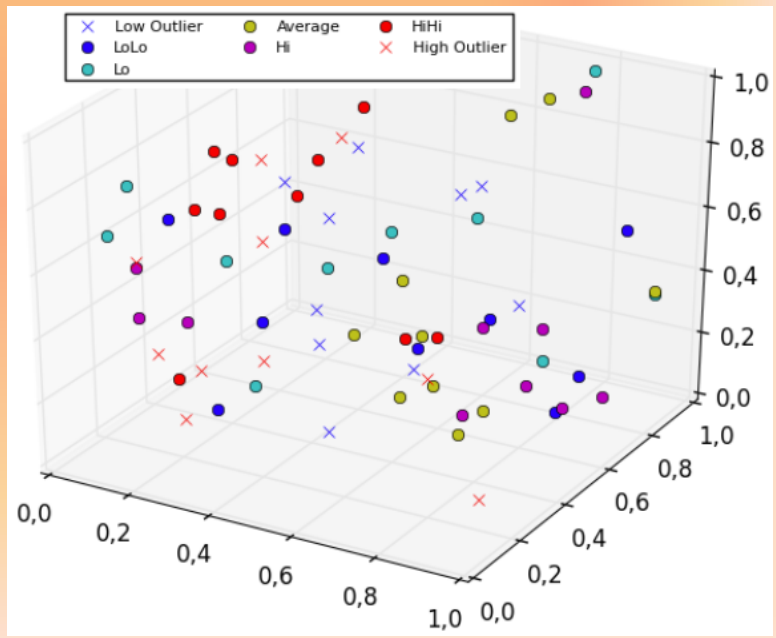
ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Visualization of the source data,

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

plotmatrix





(2,3D) scattering diagram:

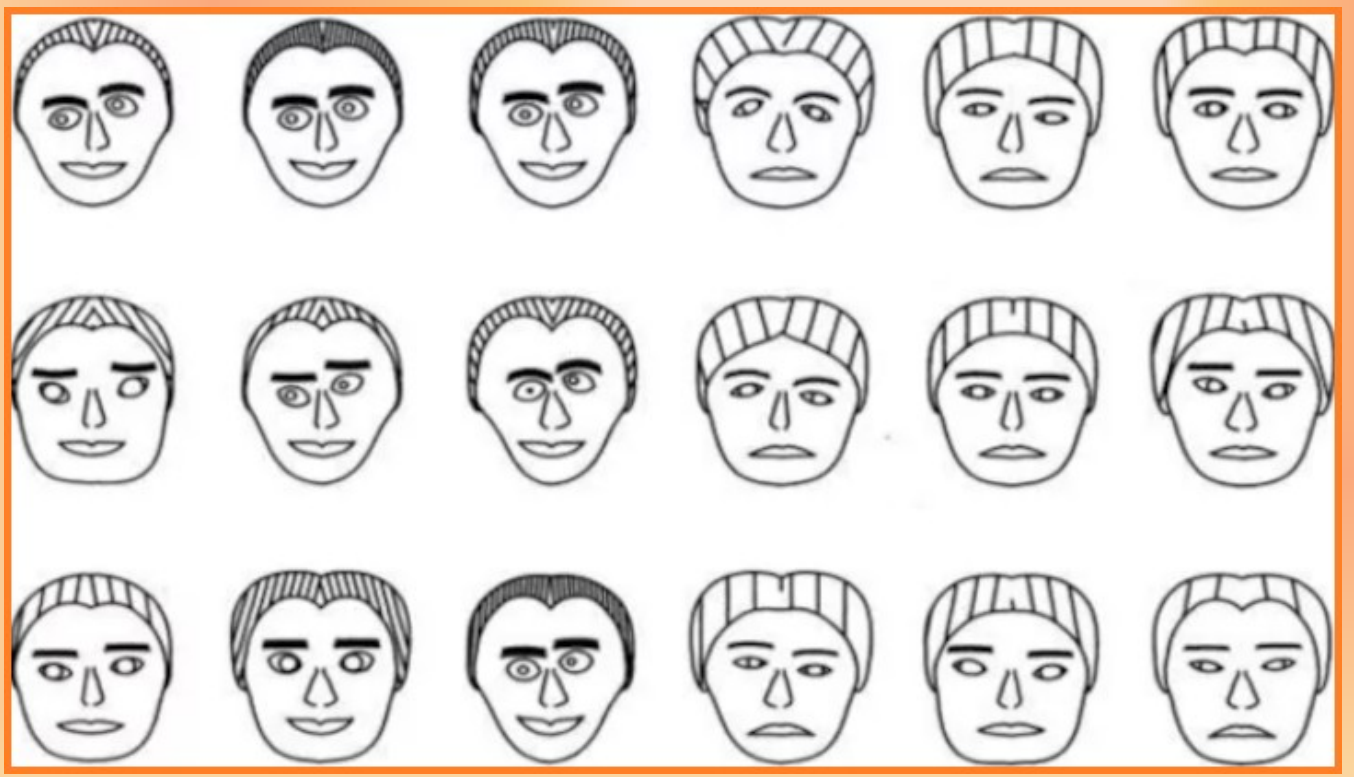
scatter, scatter3, plotmatrix

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Chernov 's Face Diagram
(for multivariate data)

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Goal: Preliminary assessment of the possibility of classification



Estimation* MatLab function **glyphplot** with option **'glyph'** equal to **'face'**

assessment[əˈsesmənt] –оценка, preview[ˈpriːvjuː]

Preliminary statistical analysis

Determination of the main sample characteristics

- Measures of Central Tendency
- Measures of Dispersion
- Shape characteristics (asymmetry, kurtosis)
- Empirical distribution

Purpose: to formulate a hypothesis about the class of a probability space or its individual parameters and properties.

scheme[skiːm]

Statistical criteria

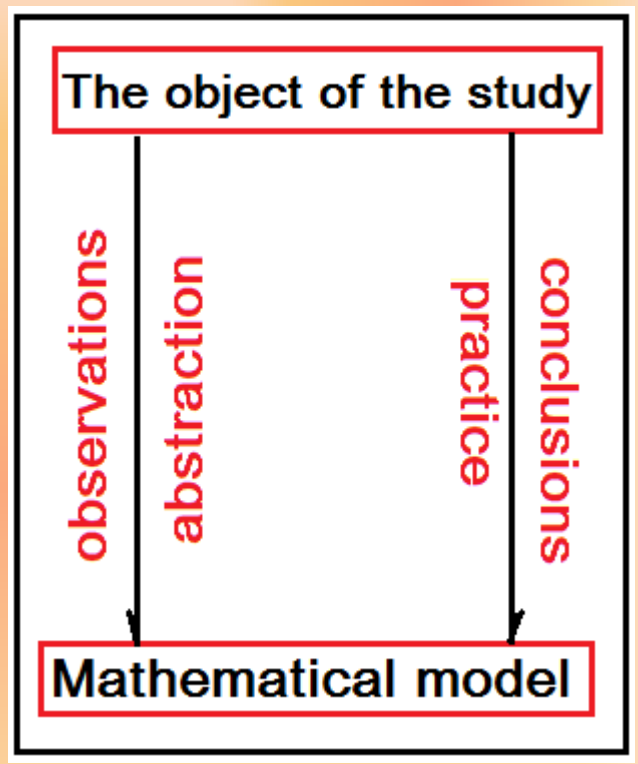
- Estimates of the correspondence of the sample distribution to the theoretical one (Kolmogorov-Smirnov criteria, Chi-squared, etc.)
- Assessment of compliance of the mean, median, variance with the corresponding parameters of the general distribution (Student, Fisher criteria and so on)
- Estimation of the correlation coefficient, parameters of the linear regression model.

Objective: criteria-based substantiation of the statistical significance of the main sample characteristics, followed by the use of theoretical distribution

compliance[kəmˈplaɪəns] соответствие

Constructing a model

- Linear regression models (including multidimensional ones)
- Nonlinear regressions
- Clustering of data (based on Euclidean distance, Mahalanobis*)
- Classification (linear with training, using neural networks – nonlinear case)



* The Mahalanobis distance is a special case of the Euclideandistance, determines the similarity of known and non-known samples, takes into account correlations between variables and is invariant to the scale

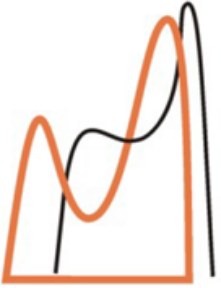
ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

MATLAB (network)

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Instructions for using MATLAB SFU:

- In the address bar in any browser, enter the following https://rdf01.sfedu.ru/RDWeb/
- Log in using the login/password from the account @sfedu.ru
- Go to the "For Training" folder
- Download the file *.rdp connection to Matlab r2021a
- Widows 10



ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК

ЮЖНЫЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

Thanks for your attention!