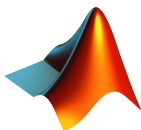


MatLab Programming Fundamentals

guarantor: Maroš Tunák

tel.: 3465

e-mail: maros.tunak@tul.cz



Course objectives

The aim of the course is to acquire basics knowledge and skills of students the MatLab program. At the end of the course students will be able to use MatLab for their own work and will be ready to deepen their programming skills in MatLab.

MatLab Programming Fundamentals

time requirements: 0p+2c

credits: 4

exercises: Monday 10:40-12:15; 12:30-14:05 (B-PC2, Tunák M.)
Tuesday 08:50-10:25; 10:40-12:15 (B-PC2, Tunák M.)

consultation: Wednesday 10:40-12:15 (E-KHT)

Requirements on student/graded credit

- 1 participation in exercises (max. 3 absences)
- 2 elaboration of semester work (after approval of the semester work, you can attend a practical demonstration)
- 3 practical demonstration of acquired skills (there will be 1-2 examples to solve; elaboration time 1 hour; you can use any materials ...)

Content

IS/STAG Syllabus

1. Getting started with Matlab. Working environment, windows, paths, basic commands, variables. Loading, saving and information about variables. Help.
2. Mathematics with vectors and matrices. Creating vectors and matrices. Indexing. Special matrices. Matrix operations. Element by element operations. Relational operations, logical operations, examples and tricks.
3. Control flow. Loops, conditional statements, examples.
4. Script m-files, Function m-files.
5. Visualisation. Two-dimensional graphics. Three-dimensional graphics.
6. Graphical user interface.
- 7.-10. Statistics and Machine Learning Toolbox. Basics of statistical data processing, exploratory data analysis, descriptive statistics, data visualisation, hypothesis testing, confidence intervals, regression analysis, control charts.
- 11.-13. Solution of practical problems in textile and industrial engineering.

Literature

Recommended

MathWorks. *Getting Started with MATLAB*. [Online]. Dostupné z:

<https://www.mathworks.com/help/matlab/getting-started-with-matlab.html>

Study materials

<http://elearning.tul.cz>

Installation

<http://liane.tul.cz/cz/software/MATLAB>

Statistics and Machine Learning Toolbox

Basics of statistical data processing, exploratory data analysis, descriptive statistics, data visualisation, hypothesis testing, confidence intervals, regression analysis, control charts.

Data Visualization

Histogram

Histogram

One of the oldest classical estimates of probability density. This is the outline of a bar graph that visually displays the distribution of data. The range of values of a random variable is divided into so-called class intervals, where we assume the width of the class is constant and we find out the number of values n_i from the given selection in the i -th class interval. We make this finding for each interval and it is valid

$$\sum_{i=1}^n n_i = n.$$

- *Frequency histogram*: bar height - frequency n_i
- *Relative frequency histogram*: bar height - relative frequency $\frac{n_i}{n}$

The disadvantage of these histograms is that they do not correspond to the probability density, because the total area of the rectangles is not equal to one.

Data Visualization

- *Density histogram*: normalized histogram, where the area under the curve (in this case the curve is represented by the height of the bars) is equal to one. The density histogram is given by the following relation

$$\hat{f}(x) = \frac{n_i}{nh} \quad x \in B_i, \quad (1)$$

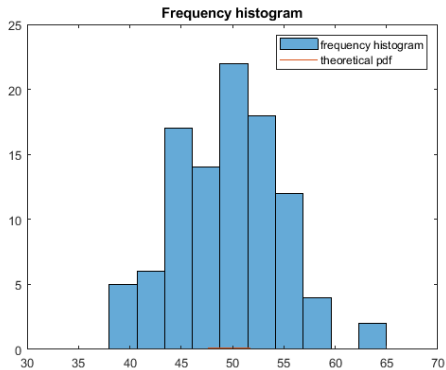
where B_i denotes i -th class, n_i represents the frequency of data belonging to the i -th class, and h is the width of the class.

Data Visualization

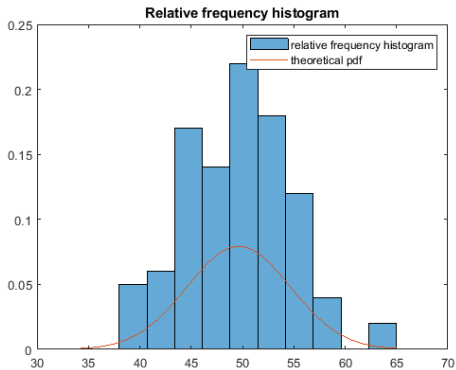
- **Example:** Generate $n = 100$ numbers from the $N(50, 5)$ distribution, create a frequency histogram, relative frequency histogram, and density histogram for 10 classes, and plot a theoretical probability density function in the histogram.

```
1 clear,clc,close all
2
3 data=normrnd(50,5,1,100); % generatinig 100 data with distribution N(50,5)
4 mu=mean(data);          % estimate of the mean value from the data
5 s=std(data);           % estimate of the st.deviation value from the data
6 x=30:0.1:70;          % vector x for which the theoretical pdf will be calculated
7 y=normpdf(x,mu,s);     % theoretical probability density function N(mu,s)
8
9 %% frequency histogram
10 figure,histogram(data,10),title('Frequency histogram')
11 hold on
12 plot(x,y)
13 legend('frequency histogram','theoretical pdf')
14
15 %% relative frequency histogram
16 figure,histogram(data,10,'normalization','probability'),title('Relative . . .
    frequency histogram')
17 hold on
18 plot(x,y)
19 legend('relative frequency histogram','theoretical pdf')
20
21 %% density histogram
22 figure,histogram(data,10,'normalization','pdf'),title('Density histogram')
23 hold on
24 plot(x,y)
25 legend('density histogram','theoretical pdf')
```

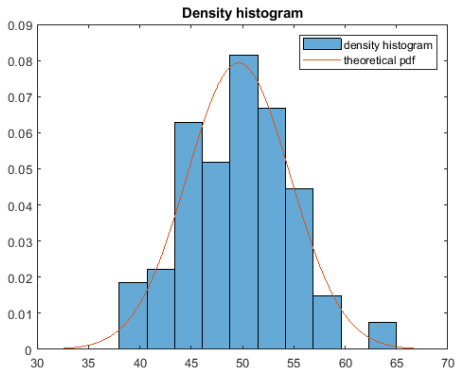

Data Visualization



Data Visualization



Data Visualization



Data Visualization

Boxplot

Boxplot

A (*Boxplot*) is a graphical representation that simultaneously describes several important data characteristics: representation of location estimation, dispersion, symmetry assessment around quantiles, symmetry assessment at distribution ends, identification of outlying data. The box chart represents:

- the tops and bottoms of each box are the 25th and 75th percentiles of the samples, $q(0.25)$ and $q(0.75)$ (first and third quartiles), respectively. The distances between the tops and bottoms are the interquartile ranges *iqr*.
- the line in the middle of each box is the sample median \tilde{x} (50th percentile $q(0.50)$, second quartile), if the median is not centred in the box, it shows sample skewness.
- the (*Whiskers*) are lines extending above and below each box. Whiskers are drawn from the ends of the interquartile ranges to the furthest observations within the whisker length (the *Adjacent Values*)

Data Visualization

- the upper and lower limits are given

$$LL = q_{0.25} - 1.5iqr$$

$$UL = q_{0.75} + 1.5iqr$$

- adjacent values are the most extreme values from the sample that are within limits
- observations that are outside the limits are considered outliers

Boxplot with notches

- notches display the variability of the median between samples
- the confidence interval limits for the median are given

$$IL = \tilde{x} - \frac{1.57iqr}{\sqrt{n}}$$

$$ID = \tilde{x} + \frac{1.57iqr}{\sqrt{n}}$$

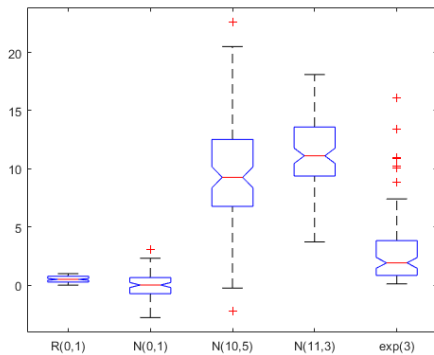
- the width of a notch is computed so that box plots whose notches do not overlap (as above) have different medians at the 5% significance level. The significance level is based on a normal distribution assumption, but comparisons of medians are reasonably robust for other distributions. Comparing box-plot medians is like a visual hypothesis test, analogous to the t-test used for means.

Data Visualization

- **Example:** Generate $n = 100$ numbers from $U(0, 1)$, $N(0, 1)$, $N(10, 5)$, $N(11, 3)$, $exp(3)$ and create notch boxplots.

```
1 clear,clc,close all
2
3 %% data
4 data_1=rand(1,100);           % random data with distr. U(0,1)
5 data_2=normrnd(0,1,1,100);    % random data with distr. N(0,1)
6 data_3=normrnd(10,5,1,100);  % random data with distr. N(10,5)
7 data_4=normrnd(11,3,1,100);  % random data with distr. N(11,3)
8 data_5=exprnd(3,1,100);      % random data with distr. exp(3)
9
10 %% plot
11 boxplot([data_1' data_2' data_3' data_4' data_5'],'notch','on')
12
13 %% labels
14 xticklabels({'U(0,1)', 'N(0,1)', 'N(10,5)', 'N(11,3)', 'exp(3)'})
```

Data Visualization



Data Visualization

Empirical cumulative distribution function

Empirical cumulative distribution function

Theoretical distribution function for a continuous random variable

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$$

expresses the probability that a random variable will take values less than or equal to x .

Empirical distribution function

- $F_n(x)$ is defined as the number of data less than or equal to x ($\#X_{(i)} \leq x$) divided by the number of data n

$$F_n(x) = \begin{cases} 0 & \text{for } x < X_{(1)} \\ \frac{i}{n} & \text{for } X_{(i)} \leq x \leq X_{(i+1)} \\ 1 & \text{for } x \geq X_{(n)} \end{cases}$$

where $X_{(i)}$ are order statistics (sorted data from smallest to largest)

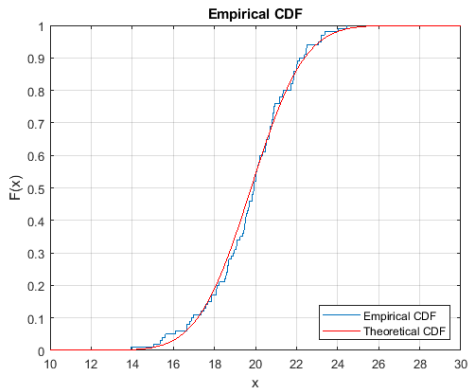
- if all values of $X_{(1)}, \dots, X_{(n)}$ are different from each other, then in each of them $F_n(x)$ has a jump of size $1/n$
- if $X_{(i)}$ occurs in the set of values $X_{(1)}, \dots, X_{(n)}$ occur k times, then $F_n(x)$ has at the point x_i jump size k/n

Data Visualization

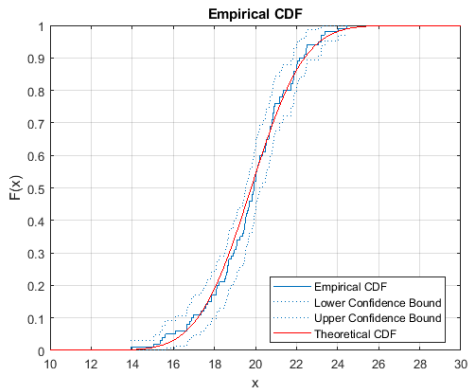
- **Example:** Generate $n = 100$ numbers from the distribution $N(20,2)$ and plot the empirical distribution function together with the theoretical distribution function.

```
1 clear,clc,close all
2
3 %% data
4 y=normrnd(20,2,1,100);% random data with distr. N(20,2)
5 cdfplot(y,hold on % drawing empirical distr. function (cdfplot)
6 x=10:0.1:30; % vector x for which the theoretical pdf will be calculated
7 fx=normcdf(x,mean(y),std(y)); % theoretical cdf N(mean(y),std(y))
8 plot(x,fx,'r')
9 legend('Empirical CDF','Theoretical CDF','Location','SE')
10
11 %%
12 figure
13 ecdf(y,'Bounds','on') % drawing empirical distr. function (ecdf) with 95% . . .
    confidence intervals
14 hold on
15 plot(x,fx,'r')
16 title('Empirical CDF')
17 legend('Empirical CDF','Lower Confidence Bound','Upper Confidence . . .
    Bound','Theoretical CDF','Location','best')
18 grid on
```

Data Visualization



Data Visualization



Data Visualization

Quantile - quantile plot

Q-Q plot

Quantile-quantile graphs (Q-Q graphs) are used to assess the agreement of two sampling distributions. It is possible to compare the agreement of the selection distribution with the theoretical distribution, where the second selection is generated from the theoretical distribution. Order statistics are plotted as an estimate of the Q-Q graph

$$x_{(i)} = x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}, \quad y_{(i)} = y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$$

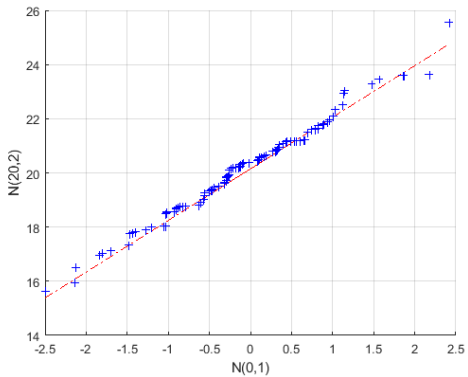
against each other. If the data comes from the same distribution, the plotted points will have an approximately linear dependence. You can use the built-in `qqplot` function to create a Q-Q graph. The function provides a reference line between the lower and upper quartile. The main advantage of quantile graphs is that they do not require the same position and variability parameter for two selections (or selection and theoretical distribution). If they come from the same distribution but differ in position and variability parameters, the points are expected to have an approximately linear relationship. The number of observations and both selections may not be the same.

Data Visualization

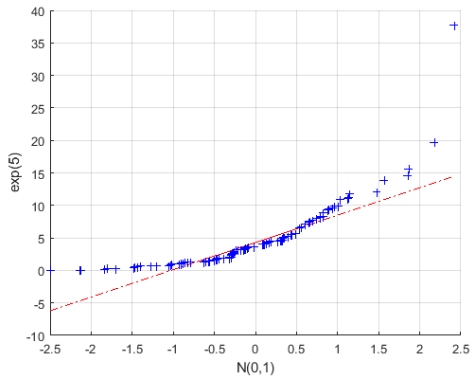
- **Example:** Generate $n = 100$ numbers from the distributions $N(0, 1)$, $N(20, 2)$ and $\exp(5)$ plot Q-Q graphs.

```
1  clear,clc,close all
2
3  %% data
4  X=normrnd(0,1,1,100);           % random data with distr. N(0,1)
5  Y=normrnd(20,2,1,100);         % random data with distr. N(20,2)
6  Z=exprnd(5,1,100);            % random data with distr. exp(5)
7
8  %%
9  figure,qqplot(X,Y)             % Q-Q plot for sample X versus Y (approximately . . .
    linear dependence)
10 grid on
11 xlabel('N(0,1)'),ylabel('N(20,2)')
12 figure,qqplot(X,Z)            % Q-Q plot for sample from X versus Z (points do . . .
    not have a linear dependence, which indicates that they do not come . . .
    from the same distribution)
13 grid on
14 xlabel('N(0,1)'),ylabel('exp(5)')
15 figure,qqplot(Y,Z)            % Q-Q plot for sample from Y versus Z (points do . . .
    not have a linear dependence, which indicates that they do not come . . .
    from the same distribution)
16 grid on
17 xlabel('N(20,2)'),ylabel('exp(5)')
```

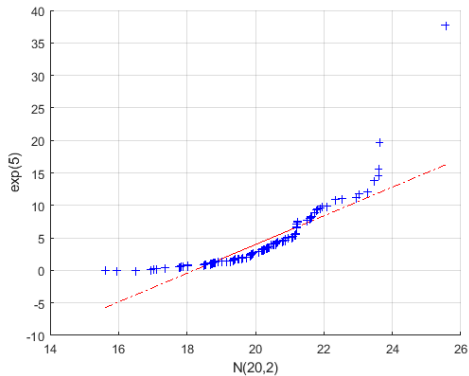
Data Visualization



Data Visualization



Data Visualization



Data Visualization

Probability plots

Probability plots

In probability plots, theoretical quantiles are plotted against sorted data (order statistics). $X_{(i)}$ is plotted on the horizontal axis and on the vertical

$$F^{-1}\left(\frac{i-0.5}{n}\right),$$

where $F^{-1}(\cdot)$ represents the inverse cumulative distribution function for a hypothetical distribution. An example of a quantile plot is a graph comparing a sample distribution with a theoretical normal distribution (*Normal Probability Plot*).

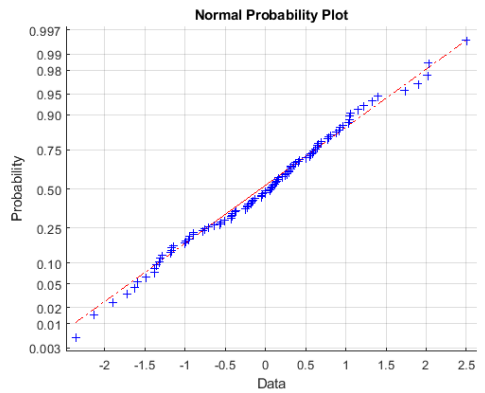
The purpose of a quantile graph for a normal distribution is to graphically verify that the data comes from such a distribution. If the data comes from the normal distribution the points in the graph have an approximately linear dependence. It is possible to use the function `normplot`, where the line between the first and third quartile is also displayed.

Data Visualization

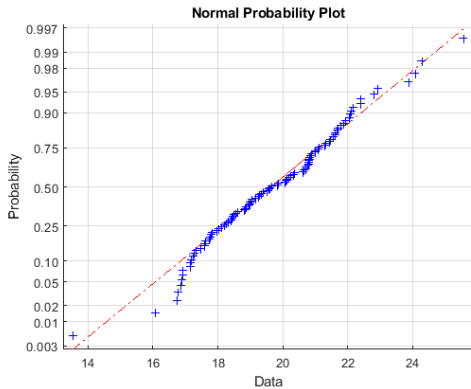
- **Example:** Generate $n = 100$ numbers from the $N(0, 1)$, $N(20, 2)$ and $exp(5)$ distributions and verify that they come from the normal distribution.

```
1 clear,clc,close all
2
3 %% data
4 X=normrnd(0,1,1,100); % random data with distr. N(0,1)
5 Y=normrnd(20,2,1,100); % random data with distr. N(20,2)
6 Z=exprnd(5,1,100); % random data with distr.
7
8 %%
9 figure,normplot(X) % Normplot for sample X (approximately linear dependence)
10 figure,normplot(Y) % Normplot for sample Y (approximately linear dependence)
11 figure,normplot(Z) % Normplot for sample Z (the points have no linear . . .
    dependence)
```

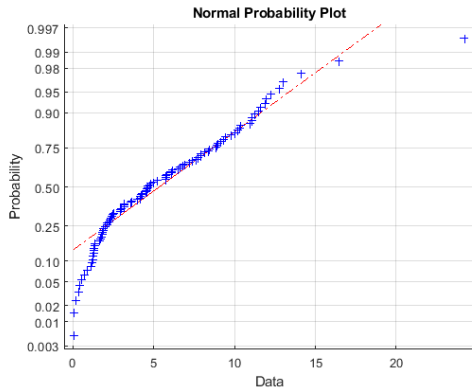
Data Visualization



Data Visualization



Data Visualization



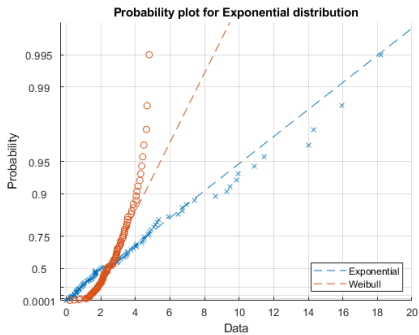
Data Visualization

Function `probplot` creates a probability graph for some other distributions (exponential, extreme values, lognormal, normal, Rayleigh, Weibull).

- **Example:** a probability plot for the exponential distribution for data from the $\text{exp}(3)$ and $\text{Weibull}(3,3)$ distributions.

```
1 clear,clc,close all
2
3 %% data
4 X=exprnd(3,100,1);      % random data with distr. exp(3)
5 Y=wblrnd(3,3,100,1);   % random data with distr. Weibull(3,3)
6
7 %%
8 probplot('exponential',[X Y]),grid on % Probplot for ...
          sample X and Y
9 legend('Exponential','Weibull','Location','SE')
```

Data Visualization



Data Visualization

Command	Description
» histogram	histogram
» boxplot	boxplot
» qqplot	quantile-quantile plot
» normplot	probability plot of normal distribution
» probplot	probability plot of selected distribution
» ecdf	empirical distribution function
» cdfplot	empirical distribution function
» ksdensity	kernel density estimation

Examples for practice

Examples for practice

- 1 Example:** Write a m-file/function (`graphs.m`), where the input parameter will be a vector of values X and the function will return graphs of exploratory data analysis (density histogram + theoretical probability density function $N(\bar{x}, s)$, boxplot and empirical distribution function + theoretical distribution function $N(\bar{x}, s)$), in one figure window. Generate $n = 100$ numbers and try created function.

Solution