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, hypotesis testing, confidence intervals, regression analysis, control charts.

Statistics and Machine Learning Toolbox

The analysis of the statistical set obtained by random sample usually begins with the determination of numerical characteristics that provide a global picture of where and how the data are concentrated and what is the shape of their distribution, i.e. the characteristics useful for further processing. These numerical characteristics are known as the `Descriptive Statistics`. All descriptive statistics listed in the following sections are available through `Statistics and Machine Learning Toolbox` in MatLab.

Descriptive Statistics

Measures of Central Tendency

Measures of central tendency locate a distribution of data along an appropriate scale. The most common location measures are:

- Arithmetic average

  Let's have a random sample of $x_1, x_2, ..., x_n$. The arithmetic average is given

  $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1}$$

  the average is a simple and popular estimate of location. If the data sample comes from a normal distribution, then the sample mean is also optimal. Unfortunately, outliers or data errors exist in almost all real data. The sample mean is sensitive to these problems. One bad data value can move the average away from the center of the rest of the data by an arbitrarily large distance.

  In this case it is possible to use trimmed mean that is resistant to outliers, which is calculated as the arithmetic mean of the sample not containing the highest and lowest (percentage / 2) % observations.

Descriptive Statistics

The geometric mean and harmonic mean, like the average, are not robust to outliers. They are useful when the sample is distributed lognormal or heavily skewed.

- Geometric mean

$$\overline{x}_g = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}, \tag{2}$$

- and harmonic mean

$$\overline{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}. \tag{3}$$

Descriptive Statistics

- The median $\tilde{x}$ divides the sample into two parts, each containing 50 % of the observations. It is a robust estimate, i.e. the median is not sensitive to outliers. If the observation are ordered (order statistics $x_{(i)}$), i.e. $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ is the median

  s for an odd number of observations

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}, \tag{4}$$

and for an even number of observations

$$\tilde{x} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}. \tag{5}$$

Descriptive Statistics

- The quantile of the order p $(0 \leqslant p \leqslant 1)$ is in the ordered data the value $x_{(p)}$ under which lies $100 \times p\%$ observations. The median is then 50% quantile $x_{(0.5)}$. In addition, the following quantiles are often used
  - percentile - $(x_{(0.01)}, x_{(0.02)} \dots x_{(0.99)})$ - divides the ordered sample into hundredths
  - deciles - $(x_{(0.1)}, x_{(0.2)} \dots x_{(0.9)})$ - divides the ordered sample into tenths
  - quintiles - $(x_{(0.2)}, x_{(0.4)} \dots x_{(0.8)})$ - divides the ordered sample into fifths
  - quartiles - $(x_{(0.25)}, x_{(0.5)}, x_{(0.75)})$ - divides the ordered sample into quarters

  Quartiles are used to create box-plots.

Descriptive Statistics

- Mode $\hat{x}$ is defined as the most frequent value for discrete distributions, for continuous distributions as the local maximum on the probability density function.

Descriptive Statistics

Command	Description
» mean	arithmetic average
» trimmean	trimmed mean
» geomean	geometric mean
» harmmean	harmonic mean
» median	median
» quantile	quantile
» prctile	percentile
» mode	mode

Descriptive Statistics

- **Example:** Calculate measures of central tendency for data $\{x_i\} = \{1.3\ 1.5\ 1.7\ 1.3\ 1.2\ 1.1\ 1.8\ 1.0\ 19\ 1.4\}$.

```
>> x=[1.3 1.5 1.7 1.3 1.2 1.1 1.8 1.0 19 1.4]
x =
    1.3000    1.5000    1.7000    1.3000    1.2000    1.1000   ...
     1.8000    1.0000   19.0000    1.4000

>> location=[mean(x) trimmean(x,20) median(x) quantile(x,[0.25 ...
     0.75]) mode(x)]
location =
    3.1300    1.4125    1.3500    1.2000    1.7000    1.3000
```

The arithmetic average (mean) is far from any data value because of the influence of the outlier (19). The median and trimmed mean ignore the outlier value and describe the location of the rest of the data values.

Measures of Dispersion

Measures of Dispersion

The purpose of measures of dispersion is to find out how spread out the data values are on the number line. The most common dispersion measures:

- The range of the data set is the difference between the maximum and the minimum, i.e.

$$R = x_{max} - x_{min}. \tag{6}$$

The disadvantage of using a range as a measure of dispersion is that it is dependent on extreme observations.

- The interquartile range is defined

$$IQR = x_{0.75} - x_{0.25}. \tag{7}$$

IQR is less sensitive than R to outliers.

Descriptive Statistics

- The variance of the (sample variance) data file $x_1, x_2, ..., x_n$ is given by

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \tag{8}$$

is interpreted as the average of the squared differences from the mean.

- Standard deviation - it is more appropriate to use the standard deviation as a measure of variability, because it is in the same units as the measured quantity

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{9}$$

Descriptive Statistics

- Coefficient of variation - when comparing the variability of variables in different units, it is possible to express the relative degree of variability using the coefficient of variation
$$cv = \frac{s}{\overline{x}}. \tag{10}$$
- Mean absolute deviation
$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|. \tag{11}$$

Descriptive Statistics

Command	Description
» range	rozpětí
» iqr	interquartile range
» var	variance
» std	standard deviation
» mad	mean absolute deviation

Descriptive Statistics

- **Example:** Calculate variability measures for data $\{x_i\} = \{1.3\ 1.5\ 1.7\ 1.3\ 1.2\ 1.1\ 1.8\ 1.0\ 19\ 1.4\}$.

```
>> x=[1.3 1.5 1.7 1.3 1.2 1.1 1.8 1.0 19 1.4]
x =
    1.3000    1.5000    1.7000    1.3000    1.2000    1.1000    ...
     1.8000    1.0000   19.0000    1.4000

>> variability=[range(x) iqr(x) var(x) std(x) std(x)/mean(x) ...
mad(x)]
variability =
   18.0000    0.5000   31.1557    5.5817    1.7833    3.1740
```

The interquartile range is the difference between the 75th and 25th percentile of the sample data, and is robust to outliers. The range is the difference between the maximum and minimum values in the data, and is strongly influenced by the presence of an outlier. Variance and the standard deviation are sensitive to outliers.

Descriptive Statistics

Measures of Shape

Of the shape characteristics, skewness and kurtosis are most often used:

- Skewness is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data spreads out more to the left of the mean than to the right. If skewness is positive, the data spreads out more to the right. The skewness of the normal distribution (or any perfectly symmetric distribution) is zero. Sample skewness (for normal distribution equals zero) is given by

$$g_1 = \frac{\sqrt{n}}{(n-1)^1/2\sigma^3/2} \sum_{i=1} n(x_i - \overline{x})^3. \tag{12}$$

- Kurtosis - estimation of kurtosis is given

$$g_2 = \frac{n}{(n-1)^2\sigma^4} \sum_{i=1} n(x_i - \overline{x})^4. \tag{13}$$

Data from the normal distribution have $g_2 = 3$.

Descriptive Statistics

Command	Description
» skewness	skewness
» kurtosis	kurtosis

Descriptive Statistics

- **Example:** Calculate the shape characteristics for the data $\{x_i\} = \{1.3\ 1.5\ 1.7\ 1.3\ 1.2\ 1.1\ 1.8\ 1.0\ 19\ 1.4\}$.

```
>> x=[1.3 1.5 1.7 1.3 1.2 1.1 1.8 1.0 19 1.4]
x =
    1.3000    1.5000    1.7000    1.3000    1.2000    1.1000    ...
     1.8000    1.0000   19.0000    1.4000

>> shape=[skewness(x) kurtosis(x)]
shape =
    2.6567    8.0800
```

Descriptive Statistics

Others

Command	Description
» max	maximum
» min	minimum
» sum	sum
» length	length of vector
» numel	number of elements in array

Descriptive Statistics

Data with Missing Values

Data with Missing Values

Many data sets have one or more missing values. It is convenient to code missing values as NaN (Not a Number) to preserve the structure of data sets across multiple variables and observations.

Příkaz	Popis
» nanmax	maximum ignoring NaN
» nanmin	minimum ignoring NaN
» nanmean	mean ignoring NaN
» nanmedian	median ignoring NaN
» nanvar	variance ignoring NaN
» nanstd	standard deviation ignoring NaN
» nansum	sum ignoring NaN
» nancov	covariance matrix ignoring NaN

Descriptive Statistics

- **Example:** Calculate the mean and variance for data with missing values $\{x_i\} = \{1.3\ 1.5\ NaN\ 1.7\ 1.3\ 1.2\ 1.1\ 1.8\ 1.0\ 19\ NaN\ 1.4\}$.

```
>> x=[1.3 1.5 NaN 1.7 1.3 1.2 1.1 1.8 1.0 19 NaN 1.4]
x =
    1.3000    1.5000       NaN    1.7000    1.3000    1.2000    ...
     1.1000    1.8000    1.0000   19.0000       NaN    1.4000

>> [mean(x) var(x)]
ans =
   NaN   NaN

>> [nanmean(x) nanvar(x)]
ans =
    3.1300   31.1557
```

Examples for practice

Examples for practice

1. **Example:** We have data on the production of two companies A and B, while the observed production characteristic is given in different units:

day	x_i production of company A (thousand pcs)	y_i production of company B (tons)
1	2	12
2	4	12
3	4	10
4	6	16
5	4	18
6	8	8
7	4	8
8	2	12
9	4	10
10	8	14

assess in which company the production is more even (less variable).

Solution