

Tutorial 3^s :

Statistics with MATLAB

Daniela Raicu

draicu@cs.depaul.edu

School of Computer Science, Telecommunications, and Information Systems
DePaul University, Chicago, IL 60604

The purpose of this tutorial is to present several statistics techniques using Matlab Statistics toolbox. For this tutorial, we assume that you know the basics of Matlab (covered in Tutorial 1) and the basics of statistics. The tutorial purpose is to teach you how to use the Matlab built-in functions to calculate the statistics for different data sets in different applications; the tutorial is intended for users running a professional version of MATLAB 6.5, Release 13.

Topics discussed in this tutorial include:

1. Descriptive statistics
2. Linear Models
3. Cluster analysis
4. Principal component analysis

1. Descriptive statistics

A. *Correlation coefficient* for two variables: "corrcoef.m"

```
>> a =
```

```
1 1 3 4
```

```
» b=[1 2 2 3]
```

```
b =
```

```
1 2 2 3
```

```
» corrcoef(a,b)
```

```
ans =
```

```
1.0000 0.8165  
0.8165 1.0000
```

Question: What large the value of 'ans' should be in order to obtain very strongly correlated variables?

B. *Standard deviation* of a variable: "std.m"

```
>> » a=[3 5 7 8 9 11]
```

```
a =
```

```
3 5 7 8 9 11
```

* Event Sponsor: Visual Computing Area Curriculum Quality of Instruction Council (QIC) grant

```
» std(a)
```

```
ans =
```

```
2.8577
```

C. *Variance* of a variable: “var.m”

```
>> » var(a)
```

```
ans =
```

```
8.1667
```

D. *Percentiles* for a data set: “prctile.m”

```
>> » prctile(a,50)
```

```
ans =
```

```
7.5000
```

```
» prctile(a,25)
```

```
ans =
```

```
5
```

```
» prctile(a,75)
```

```
ans =
```

```
9
```

```
» prctile(a,1)
```

```
ans =
```

```
3
```

```
» prctile(a,100)
```

```
ans =
```

```
11
```

E. *Range* for a data set: “range.m”

```
>> range(a)
```

```
ans =
```

```
8
```

2. Linear models

Regression analysis: “regress.m”

```
>> b=[13 15 15 17 18 20]'
```

```
b =
```

```
13  
15  
15  
17  
18  
20
```

```
» a=[3 5 7 8 9 11]'
```

```
a =
```

```
3  
5  
7  
8  
9  
11
```

```
» coef=regress(a,b)
```

```
coef =
```

```
0.4516
```

3. Cluster analysis

Hierarchical clustering on a set of data: "clusterdata.m"

```
» X=[1 1 2 3;7 8 9 7;1 3 2 1;10 9 11 9;2 2 1 1;3 4 1 2]
```

```
X =
```

```
1 1 2 3  
7 8 9 7  
1 3 2 1  
10 9 11 9  
2 2 1 1  
3 4 1 2
```

```
» T = clusterdata(X,0.8)
```

```
T =
```

```
1  
2  
1  
2  
1  
1
```

T = CLUSTERDATA(X,CUTOFF) is the same as

```
Y = pdist(X,'euclid');
Z = linkage(Y, 'single');
T = cluster(Z, CUTOFF);
```

4. Principal Component Analysis (for dimensionality reduction)

Principal components analysis from raw data: "princomp.m"

```
» [pc, score, latent, tsquare]=princomp(X)
```

pc =

```
    0.4961    -0.3697    0.6397   -0.4561
    0.4316   -0.6706   -0.4003    0.4515
    0.6032    0.4179   -0.5296   -0.4254
    0.4514    0.4889    0.3874    0.6381
```

score =

```
 -4.7824    2.0736    0.3947    0.2489
  7.2433    0.0423   -0.7265    0.2473
 -4.8221   -0.2454   -1.1806   -0.1243
 11.2723    0.0763    0.5080   -0.2441
 -5.3608   -0.3624    0.3889   -0.6064
 -3.5502   -1.5843    0.6155    0.4786
```

latent =

```
53.3994
 1.4018
 0.5731
 0.1590
```

tsquare =

```
4.1571
2.2893
3.0077
3.2088
3.2088
4.1284
```

5. Application

Apply correlation analysis, dimensionality reduction, and clustering techniques for the data containing the texture characterization for 193 CT images. The features/attributes consist of 11 texture descriptors calculated from the raw data and used to classify the tissues' textures of five human body organs: heart, backbone, liver, spleen, and kidneys. The data is contained in the file run_length_des.m".