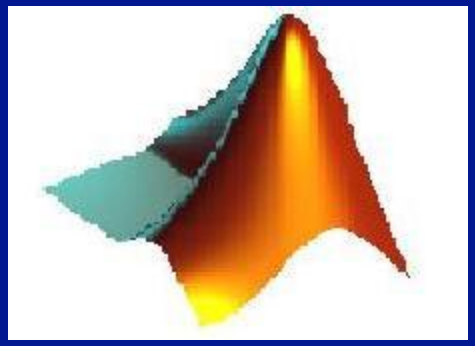
Statistical Analysis in MATLAB



Hot Topic – 18 Jan 2006

Sanjeev Pillai

BARC

MATLAB – Basic Facts

- MATrix LABoratory
- Standard scientific computing software
- Interactive or programmatic
- Wide range of applications
- Bioinformatics and Statistical toolboxes
- Product of MathWorks (Natick, MA)
- Available at WIBR (~20 licenses now)

Basic operations

- Primary data structure is a matrix
- To create a matrix

```
a = [1 2 3 4]          % creates a row vector
b = 1:4                % creates a row vector
c = pi:-0.5:0          % creates a row vector
d = [1 2;4 5;7 8]      % creates a 3x2 matrix
```

- Operations on matrices

```
a+c       % adds 'a' and 'c' to itself if dimensions agree
d'        % transposes d into a 2x3 matrix
size(d)   % gives the dimensions of 'd'
x*y       % multiplies 'x' with 'y' following matrix rules
x .* y    % element by element multiplication
```

Basic operations

- Accessing matrix values
  - d(3,2) % retrieves the 3rd rw, 2nd cl element of d
  - d(3,:) % all elements of the 3rd row
  - d(:,2) % all elements of the 2nd column
  - d(1:2,2) % 1st to 2nd row, 2nd column
- Assigning values to matrix elements
  - d(1,1)=3; % assigns 3 to (r1,c1)
  - d([1 2],:)=d([3 3],:) % change the first 2 rows to the 3rd
  - d=d^2 % squares all values in d

Basic operations

- Strings
  - Row vectors that can be concatenated
  - x = ‘Matlab’
  - y = ‘class’
  - z = [x ‘ ’ y] % z gets ‘Matlab class’
- Useful functions
  - doc, help % for help with various matlab functions
  - whos % Lists all the variables in current workspace
  - clear % clears all variables in the current workspace

Read/Write Data (File I/O)

- Several data formats supported
  - text, xls, csv, jpg, wav, avi etc.
- From the prompt or using ‘Data Import’
- Read into variables in the workspace
  - [V1 V2 V3..] = textread(‘filename’ ,’ format’)
  - eg. [l,o] = textread('energy.txt','%f%f','delimiter',',','headerlines', 1,'emptyvalue',NaN);
- Treated as regular matlab variables
- Write out into files
  - fid=fopen(‘en.txt’ , ‘w’);
  - fprintf(fid, ‘%f\t%f\n’ ,[lean;obese]);
  - fclose(fid);
  - xlswrite('energy.xls',[num2cell([lean obese])]);

Basic Statistics in Matlab

- mean(lean) % calculates the mean
- median(lean)
- std(obese(finite(obese))) % ignores the NaNs
- Visualize data
  - boxplot([lean,obese],'labels',{'Lean','Obese'})

  - Select variables from workspace
  - Use the plotting tool from the interface

Hypothesis testing

- One sample z-test
  - Done to test a sample statistic against an expected value (population parameter)
  - Done when the population sd is known
  - ztest(vector,mean,sd);
  - [h,p,ci,zscore]=ztest(vector,mean,sigma,alpha,tail)
- One sample t-test
  - Done when the population sd is not known.
  - [h,p,ci,tscore]=ttest(vector,mean,alpha,tail)

Two-sample tests

- Paired samples
  - Data points match each other
  - Eg. before/after drug treatment
  - [h,p,ci,stats]=ttest(d1,d2,alpha)
- Independent samples
  - Data points not related
  - Eg. Data from 2 groups of people
  - [h,p,ci,stats]=ttest2(d1,d2,alpha)

Test for assumptions

- Data is normally distributed
  - Paired: Delta is normally distributed
  - Independent: Both data sets are normal
  - normplot(var) or qqplot(var) or qqplot(v1,v2)
- Data is homogenous (equal variances)
  - F-test
  - Tests whether the ratio of the variances is 1.
  - [h,p,ci,stats]=vartest2(g1,g2,0.01)

Non-parametric tests

- Data need not be normal
- Compare ranks instead of values
- By ranking the signs or sums
- Wilcoxon signed rank test (one sample or paired samples)
  - [p,h,stats]=signrank(var1,var2)
- Wilcoxon rank sum test (Independent samples)
  - [p,h,stats]=ranksum(var1,var2)

Multiple hypothesis correction

- Applied when a test is done several times
  - Significance occurs just by chance
  - Eg. Microarray analysis (wild type vs mutant)
- Bonferroni correction
  - Multiply raw p-value with the number of repetitions
  - for i=1:number_of_reps
    - calculate p-value for each
    - correct each p-value
    - store in a data structure
  - end

Comparing proportions

- Analyze proportions instead of values
- Chi-square test
  - No single command in matlab
  - x= [matrix of contingency table];
  - e= sum(x')'*sum(x)/sum(sum(x));
  - X2=(x-e).^2./e
  - X2=sum(sum(X2))
  - df=prod(size(x)-[1,1])
  - P=1-chi2cdf(X2,df)

Some more tests

- Enrichment analysis
  - Is the given data enriched for a category?
  - Used widely in biological data analysis
  - Hypergeometric probability analysis
    - Y = hygecdf(X,M,K,N);
- Correlation
  - Identify correlation between paired values
  - From -1 to +1: perfect +ve and inverse correlations
    - [R,P] = corrcoef(x,y);

Matlab resources

- Online help
  - http://www.mathworks.com/access/helpdesk/help/helpdesk.shtml
- Open source user community
  - Someone may have already done what you need
  - http://www.mathworks.com/matlabcentral/
- Topics not covered
  - Scripts and functions
  - Complex data structures
  - Programming