Introduction to MATLAB

Violeta Ivanova, Ph.D.
Office for Educational Innovation & Technology

violeta@mit.edu
http://web.mit.edu/violeta/www

Topics

- MATLAB Interface and Basics
- Calculus, Linear Algebra, ODEs
- Graphics and Visualization
- Basic Programming
- Programming Practice
- Statistics and Data Analysis

Resources

- Class materials

  http://web.mit.edu/acmath/matlab/IAP2007

  - Previous sessions: InterfaceBasics, Graphics
  - This session: Statistics <.zip, .tar>

- Mathematical Tools at MIT

  http://web.mit.edu/ist/topics/math

MATLAB Help Browser

- MATLAB
  - + Data Analysis
    - + Preparing Data for Analysis
    - + Data Fitting Using Linear Regression
- Curve Fitting Toolbox
  - + Fitting Data
- Statistics Toolbox
  - + Descriptive Statistics
  - + Linear Models
  - + Hypothesis Tests
  - + Statistical Plots

MATLAB Data Analysis

Preparing Data

Correlation

Basic Fitting

Data Input / Output

- Import Wizard for data import

  `File->Import Data …`

- File input with `load`

```
B = load('datain.txt')
```

- File output with `save`

```
save('dataout', 'A', '-ascii')
```

Missing Data

- Removing missing data
  - Removing NaN elements from vectors

```
>> x = x(~isnan(x))
```

  - Removing rows with NaN from matrices

```
>> X(any(isnan(X),2),:) = []
```

- Interpolating missing data

```
YI = interp1(X, Y, XI, 'method')
```

Methods: `'spline'`, `'nearest'`, `'linear'`, …

Correlation

- Definition

  Tendency of two variables to increase or decrease together.

- Measure

  Pearson product-moment coefficient

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Correlation Example

- Import Data: `cancersmoking.dat`
- Correlation coefficient & confidence interval

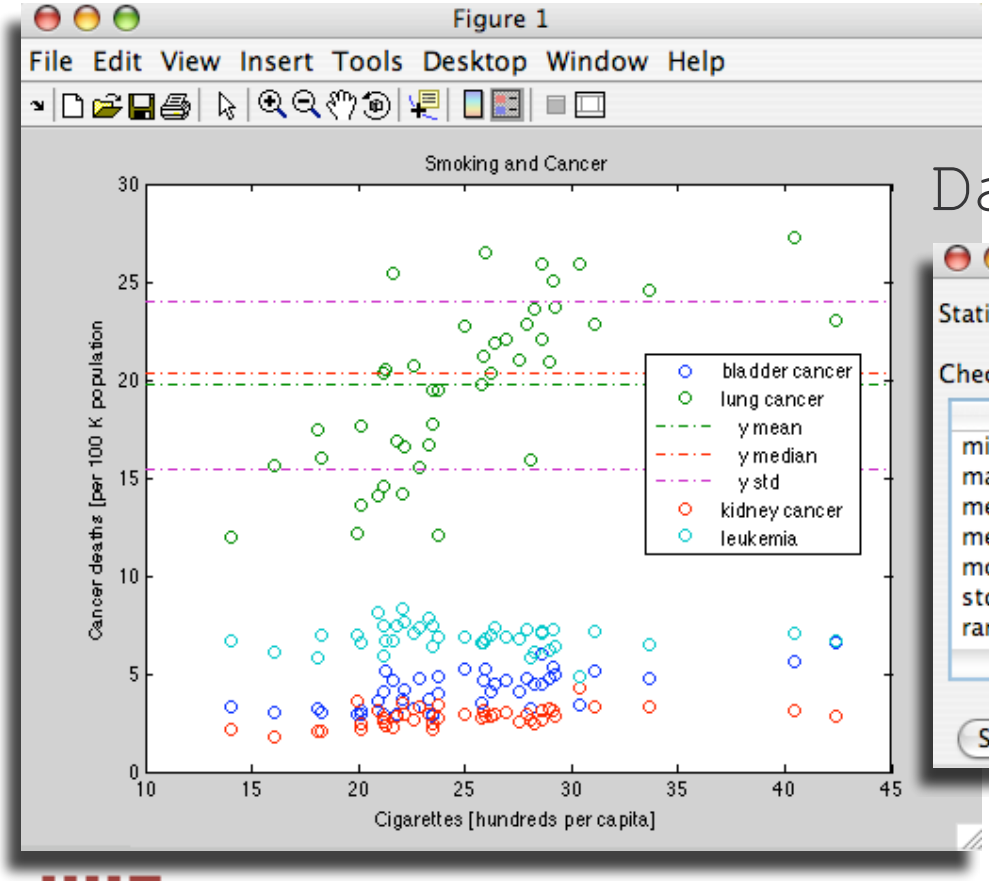
```
>> [R, P] = corrcoef(X);
>> [i, j] = find(P < 0.05);
```

```
X =

   18.2000   17.0500    6.1500
   25.8200   19.8000    6.6100
   18.2400   15.9800    6.9400
   28.6000   22.0700    7.0600
   31.1000   22.8300    7.2000
   33.6000   24.5500    6.4500
   40.4600   27.2700    7.0800
   28.2700   23.5700    6.0700
   20.1000   13.5800    6.6200
   27.9100   22.8000    7.2700
   26.1800   20.3000    7.0000
   22.1200   16.5900    7.6900
   21.8400   16.8400    7.4200
             17.7100    6.4100
             4500       6.7100
                        6.2400
```

```
>> [r,p]=corrcoef(X)

r =

    1.0000    0.6974   -0.0685
    0.6974    1.0000   -0.1516
   -0.0685   -0.1516    1.0000


p =

    1.0000    0.0000    0.6587
    0.0000    1.0000    0.3260
    0.6587    0.3260    1.0000
```

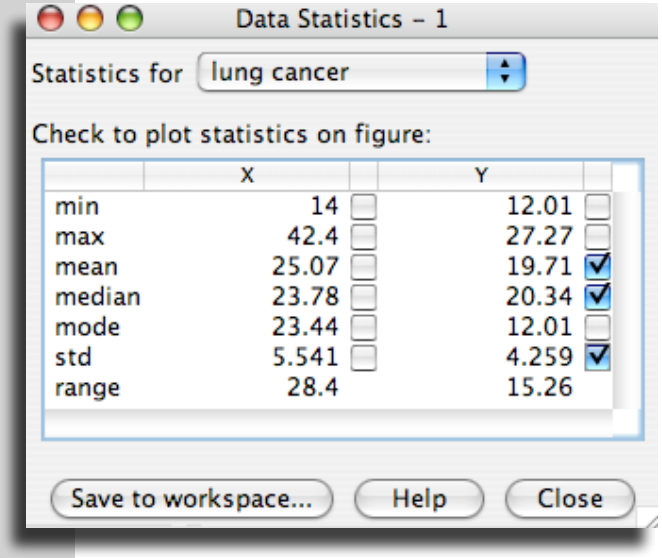
Data Statistics

- Figure Editor: `smokecancer.fig`



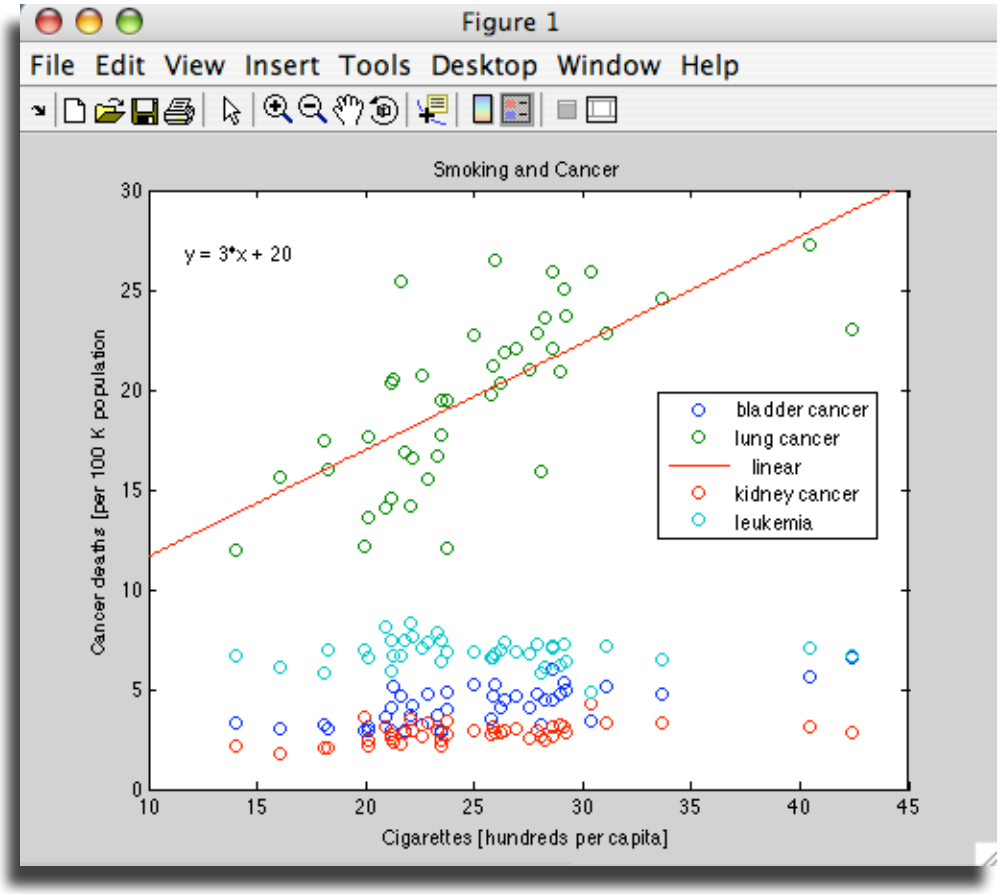
`Tools->`

`Data Statistics`



Basic Fitting

- Figure Editor : `Tools->Basic Fitting …`



Statistics Toolbox

Probability Distributions

Descriptive Statistics

Linear & Nonlinear Models

Hypothesis Tests

Statistical Plots

Descriptive Statistics

- Central tendency

```
>> m    = mean(X)
>> gm   = geomean(X)
>> med  = median(X)
>> mod  = mode(X)
```

- Dispersion

```
>> s = std(X)
>> v = var(X)
```

Probability Distributions

- Probability density functions

```
>> Y = exppdf(X, mu)
>> Y = normpdf(X, mu, sigma)
```

- Cumulative density functions

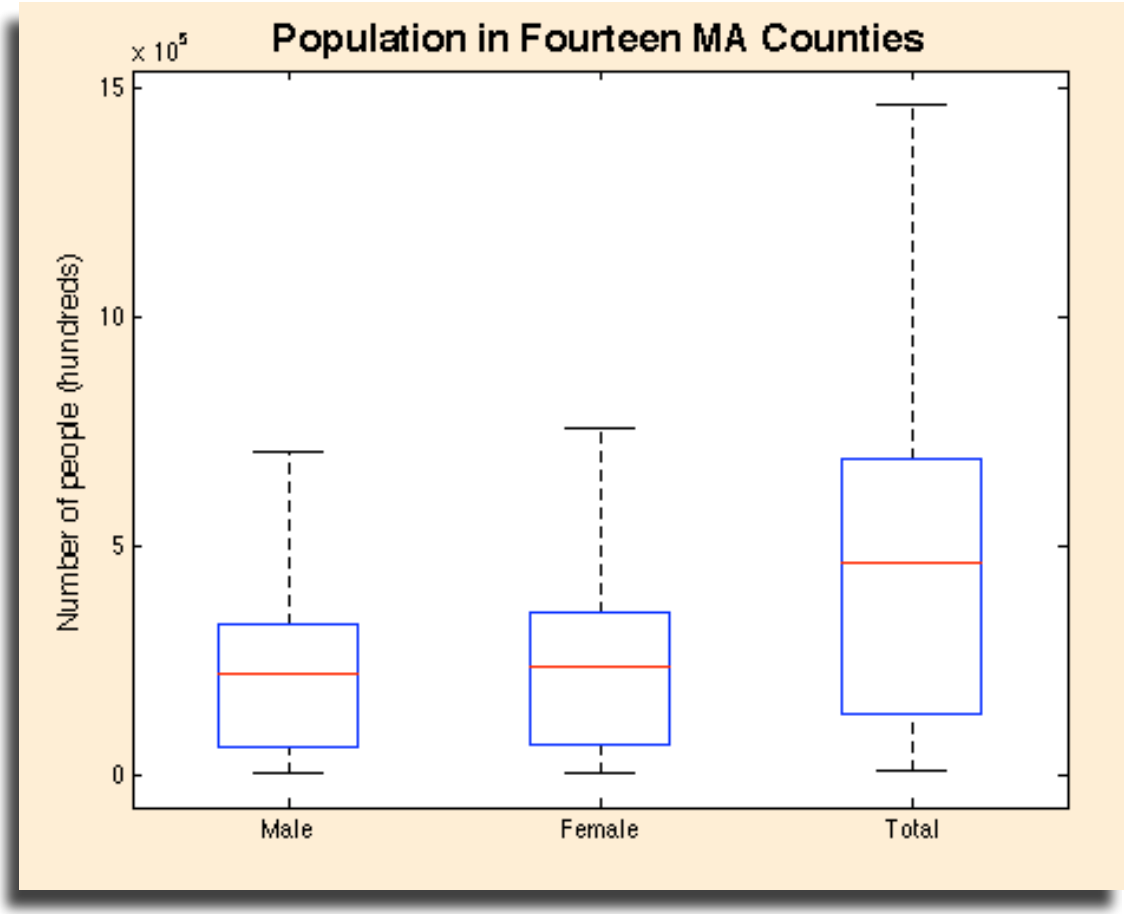
```
>> Y = expcdf(X, mu)
>> Y = normcdf(X, mu, sigma)
```

- Parameter estimation

```
>> m = expfit(data)
>> [m, s] = normfit(data)
```

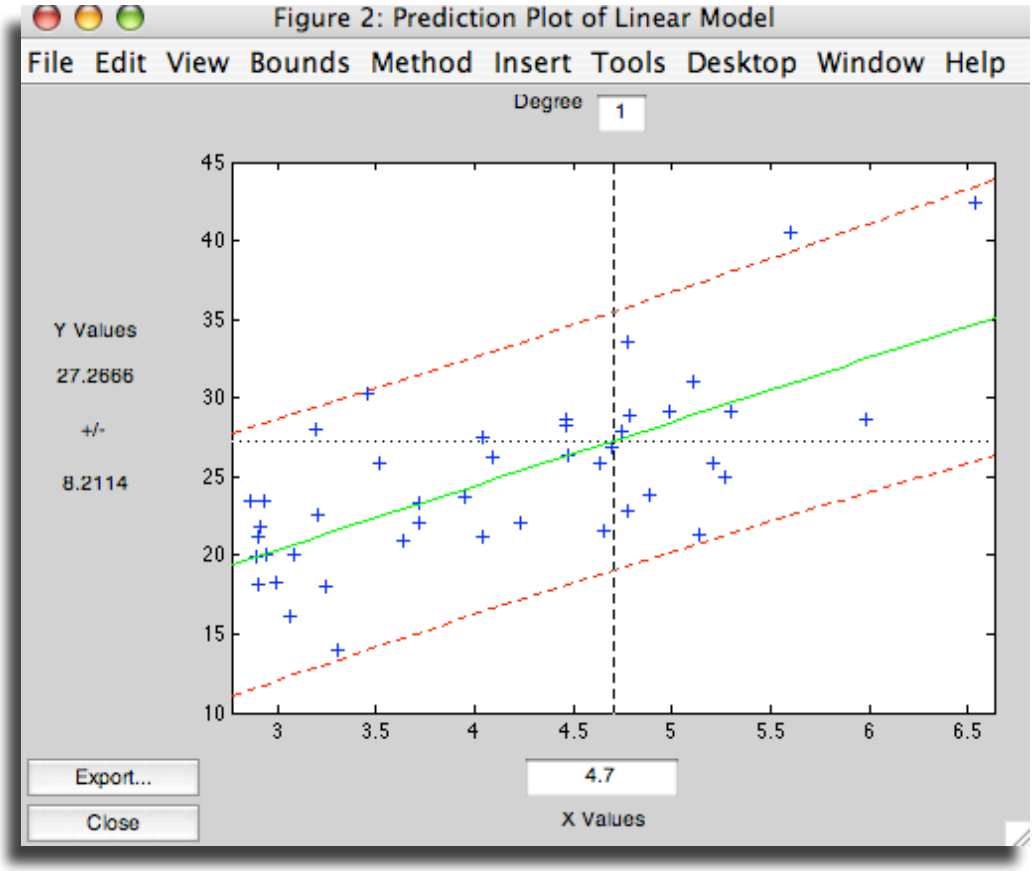
Statistical Plots

```
>> bp = boxplot(X, group)
```



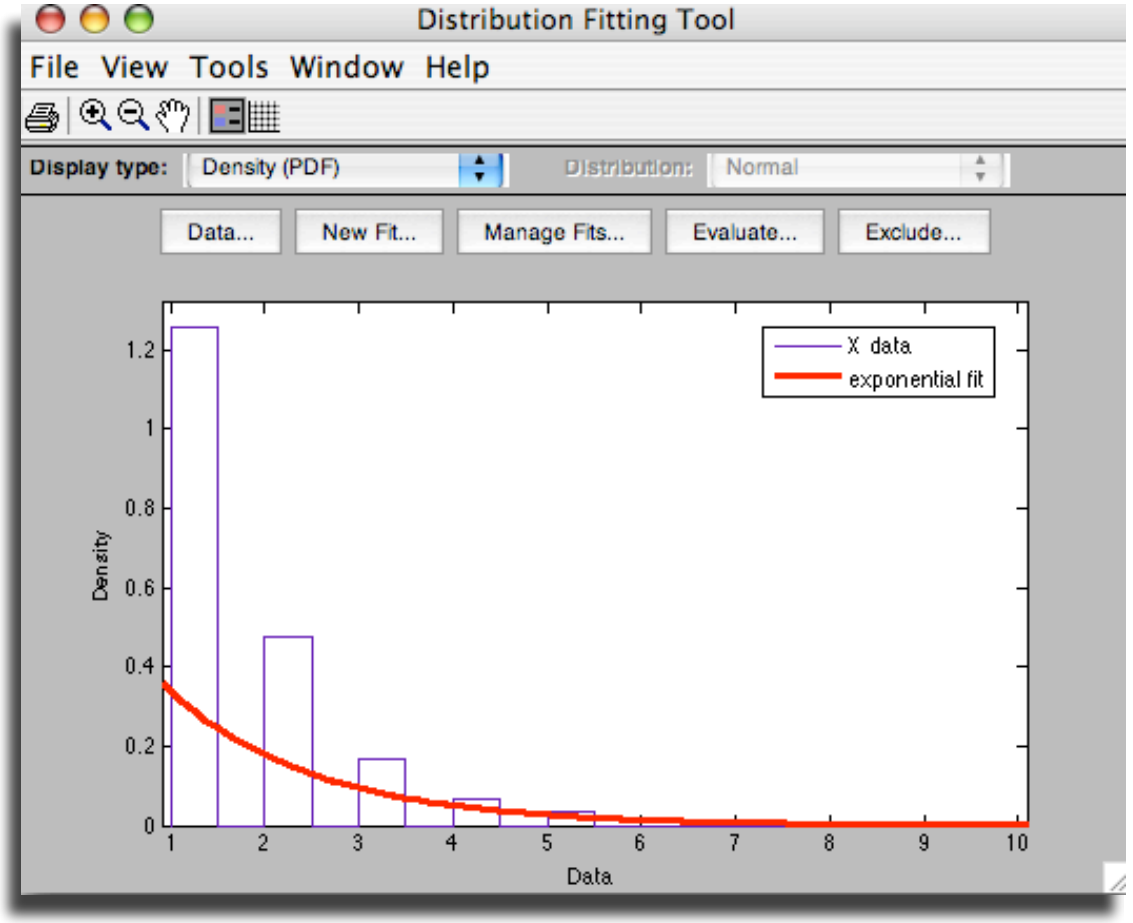
Polynomial Fitting Tool

```
>> polytool(X, Y)
```



Distribution Fitting Tool

```
>> dfittool
```



Linear Models

- Definition:

$$y = X\beta + \varepsilon$$

y: *n x 1* vector of observations

X: *n x p* matrix of predictors

β: *p x 1* vector of parameters

ε: *n x 1* vector of random disturbances

Linear Regression

- Multiple linear regression

```
>> [B, Bint, R, Rint, stats] = regress(y, X)
```

  `B`: vector of regression coefficients

  `Bint`: matrix of 95% confidence intervals for B

  `R`: vector of residuals

  `Rint`: intervals for diagnosing outliners

  `stats`: vector containing R^2 statistic etc.

- Residuals plot

```
>> rcoplot(R, Rint)
```

Hypothesis Testing

- Definition: use of statistics to determine the probability that a given hypothesis is true.
  - Null hypothesis (observations are the result of pure chance) and alternative hypothesis.
  - Test statistic to assess truth of null hypothesis.
  - P-value: probability of test statistic to be that significant if null hypothesis were true.
  - Comparison of P-value to acceptable α-value.

Analysis of Variance (ANOVA)

- One-way ANOVA

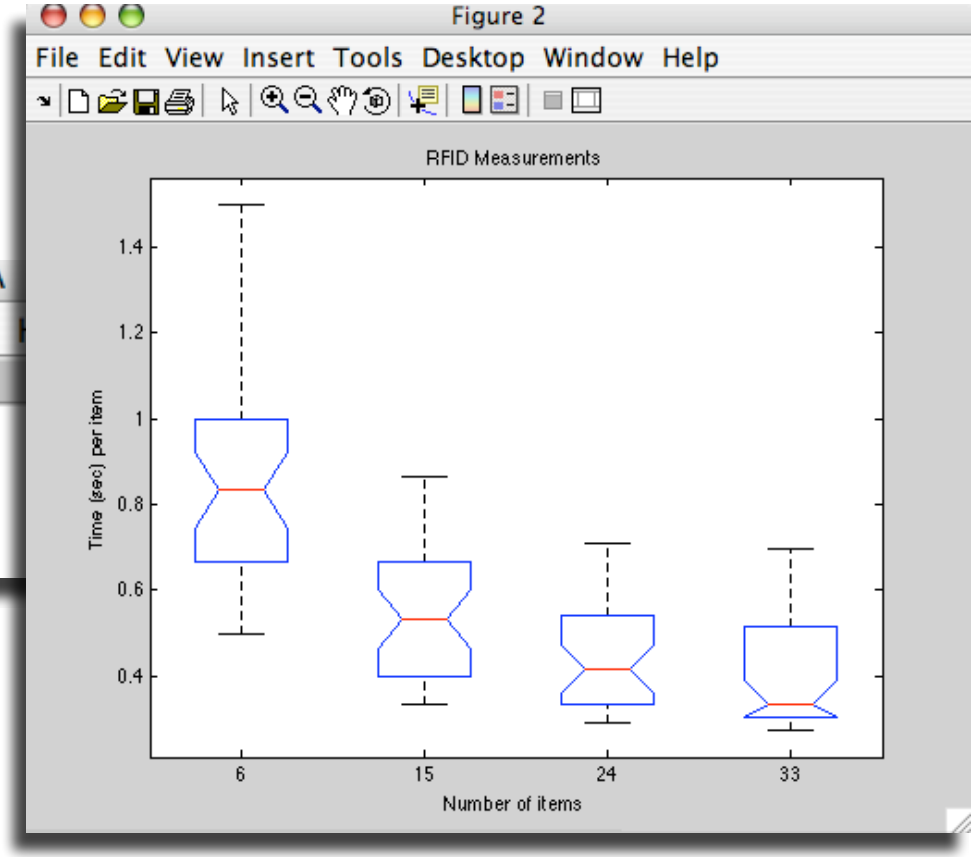
```
>> anova1(X,group)
```

Figure 1: One-way ANOVA

File Edit View Insert Tools Desktop Window

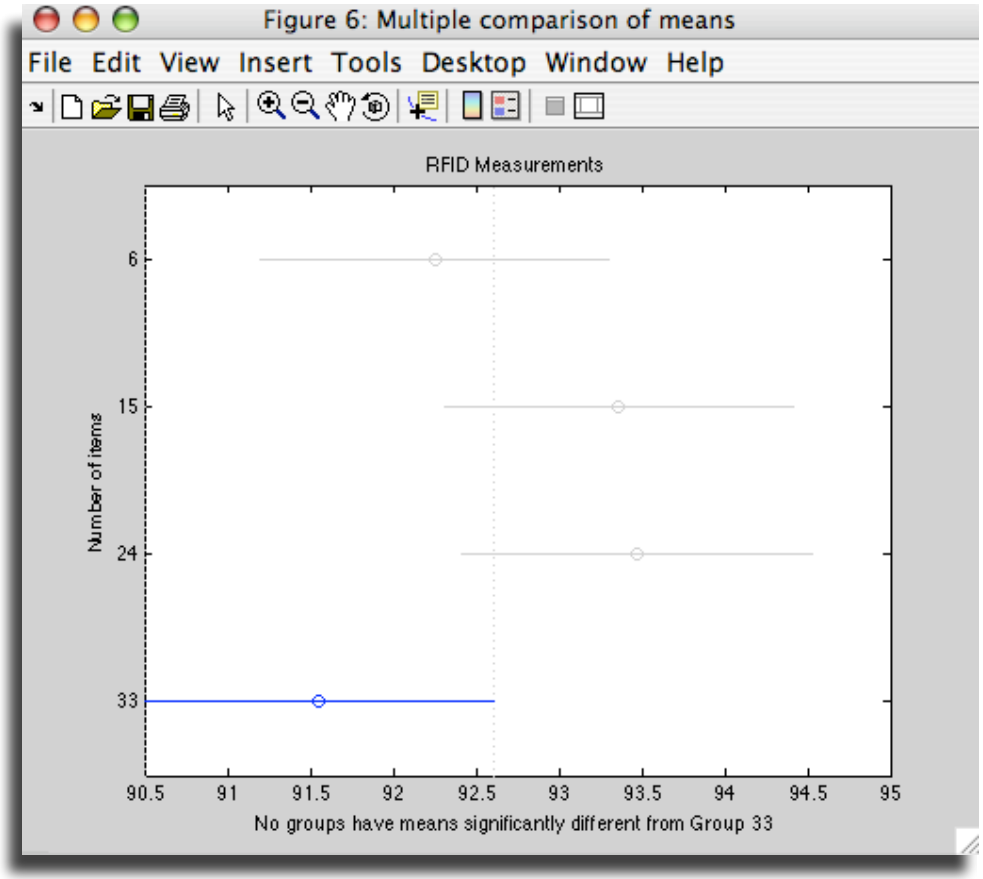
ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	3.95296	3	1.31765	44.37	0
Error	4.15753	140	0.0297		
Total	8.11048	143			



Multiple Comparisons

```
>> [p, tbl, stats]
= anova1(X,group)
>> [c, m] =
multcompare(stats)
```



More Built-In Functions

- Two-way ANOVA

```
>> [P, tbl, stats] = anova2(X, reps)
```

- Other hypothesis tests

```
>> H = ttest(X)
>> H = lillietest(X)
```

Data Analysis Exercises

- Exercise One: `dataanalysis.m, rfid.dat, barcode.dat`
  - Correlation coefficient
  - Hypothesis testing
  - Statistical plots
  - ANOVA

Follow instructions in the m-file …

Curve Fitting Toolbox

Curve Fitting Tool

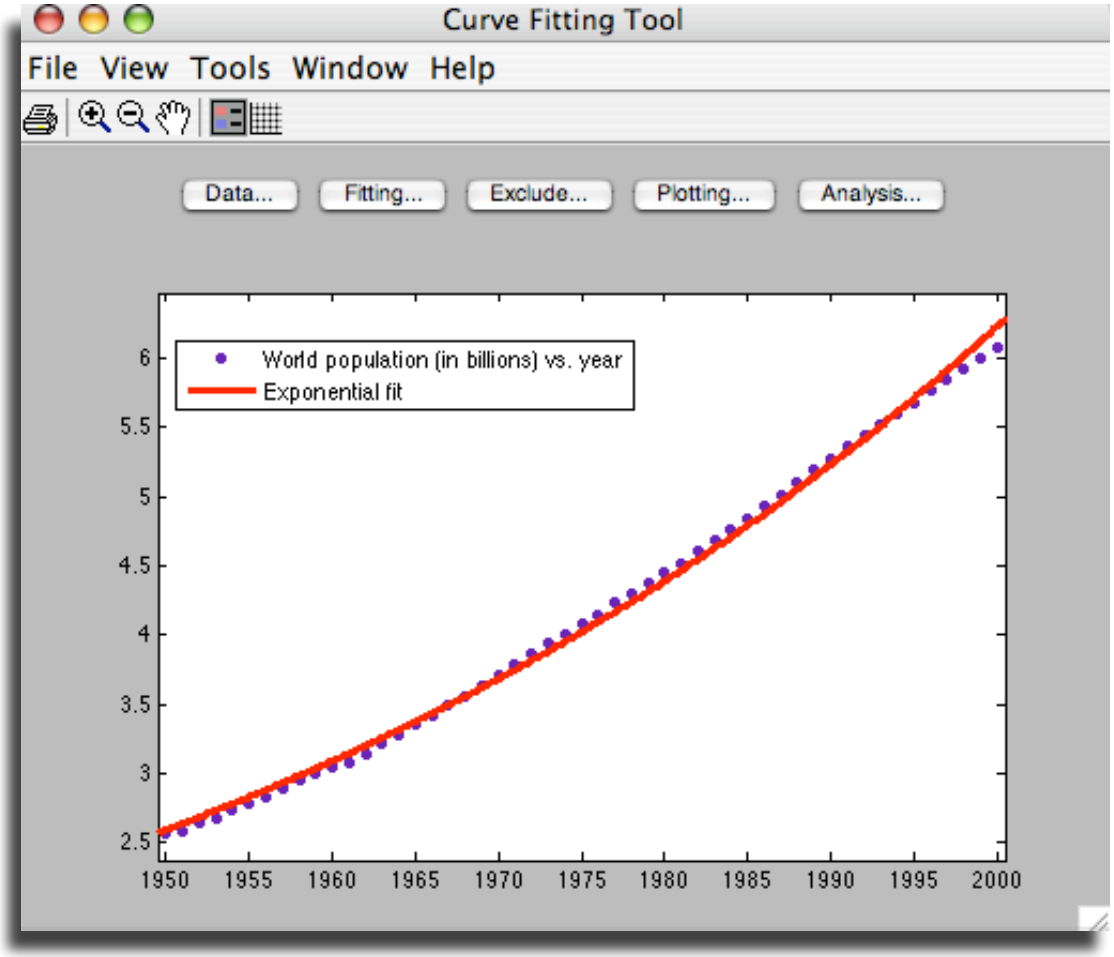
Goodness of Fit

Analyzing a Fit

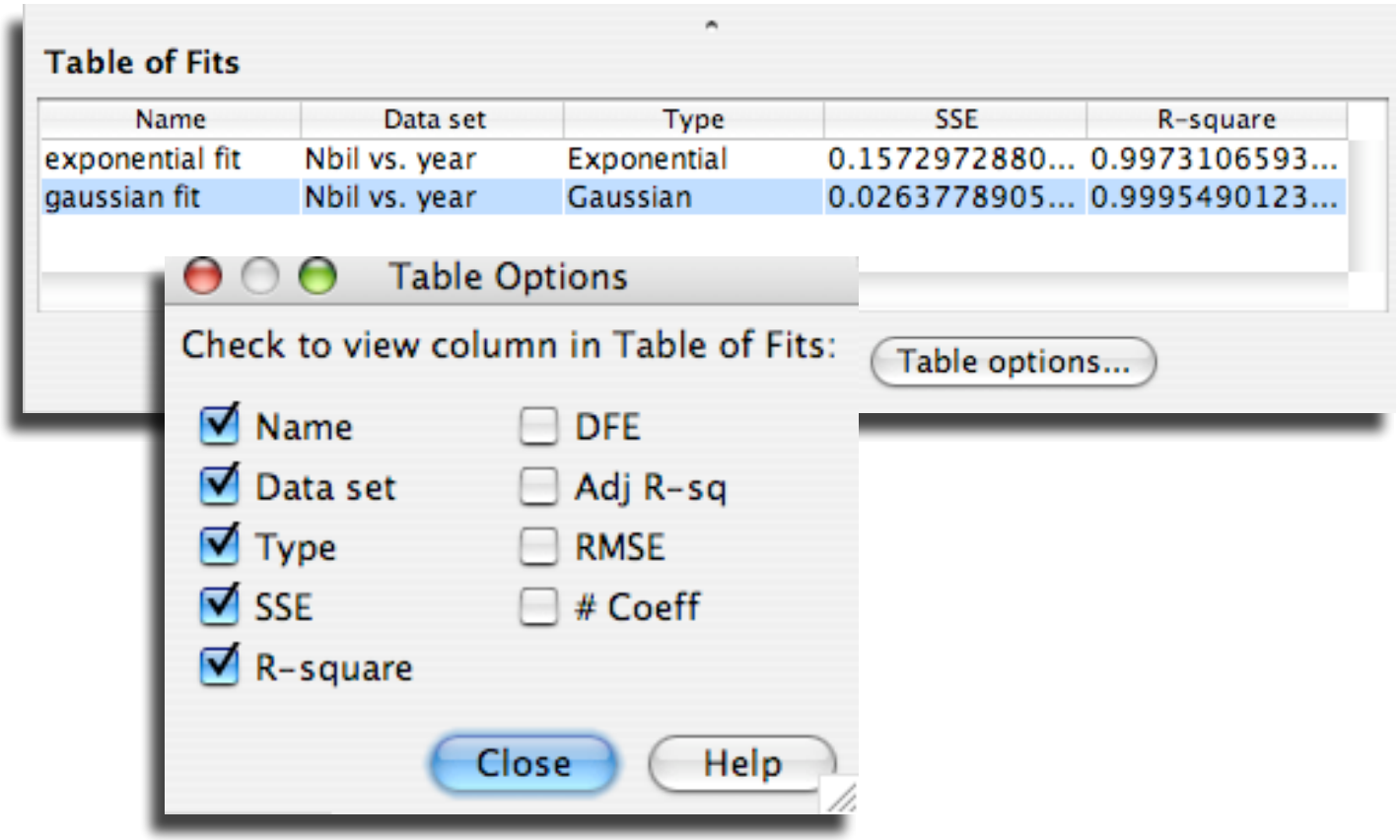
Fourier Series Fit

Curve Fitting Tool

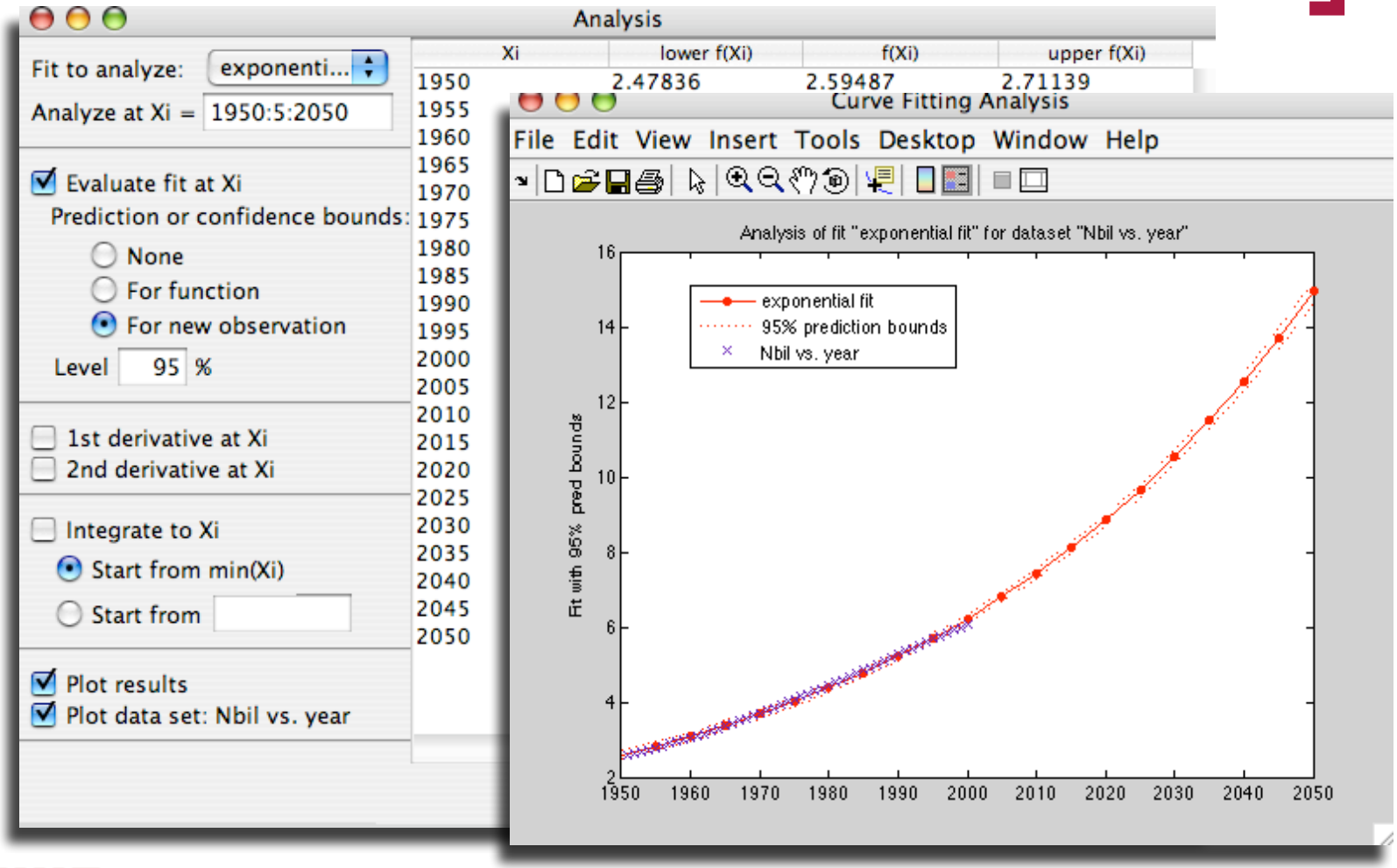
```
>> cftool
```



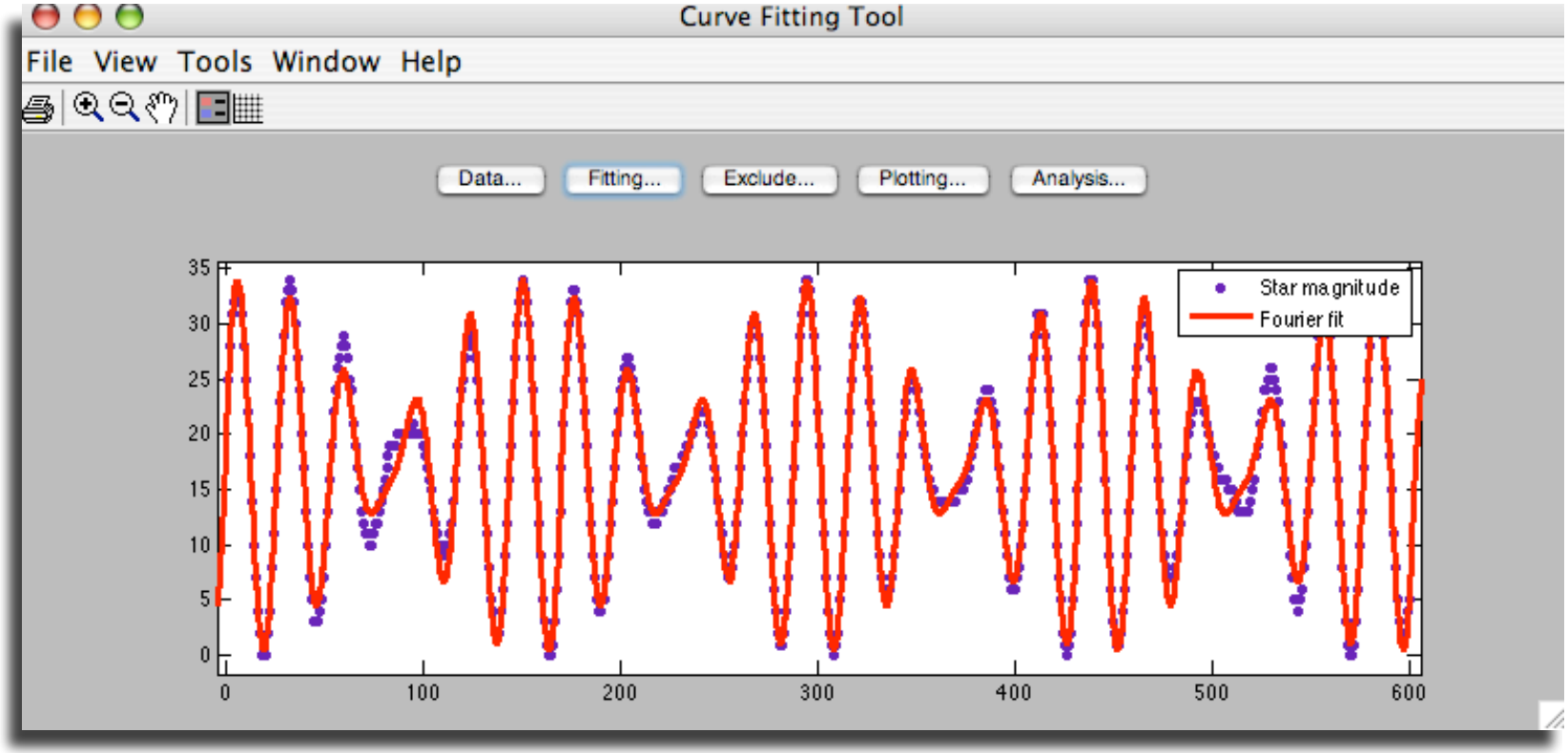
Goodness of Fit Statistics



Analyzing a Fit



Fourier Series Fit



Data Analysis Exercises

- Exercise Two: `regression.m, worlddata.dat, star.txt`
  - Linear regression
  - Polynomial fitting
  - Probability density function fitting
  - Goodness of Fit

Follow instructions in the m-file …