BMI Statistics

Corso LT ‘Meccanica dei Tessuti Biologici’ – Prof. A. Alhuwalia

11/05/2017

ludovica.cacopardo@ing.unipi.it

Aim of the Lesson

- Understand the importance of statics in practical cases
- Understand how to do some literature analysis
- Collect some data
- Do some statistics (on Excel and Matlab)

  - distribution plot

  - scatter plot and correlation

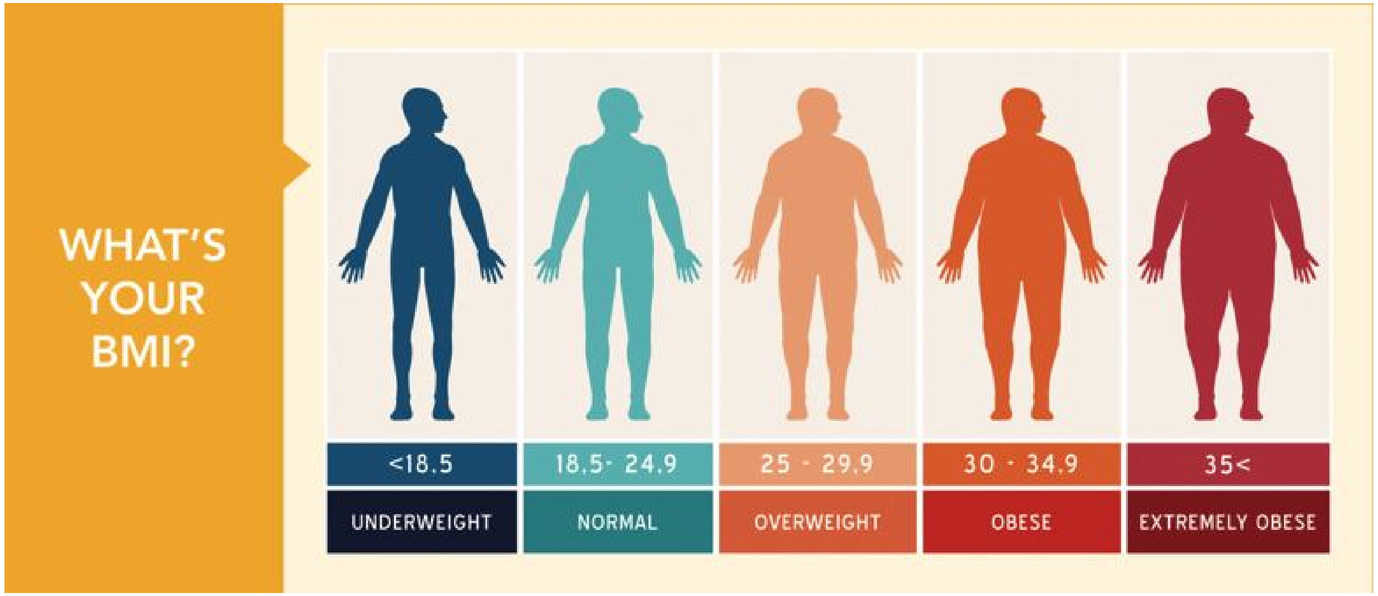
  - t-test & Anova

BMI

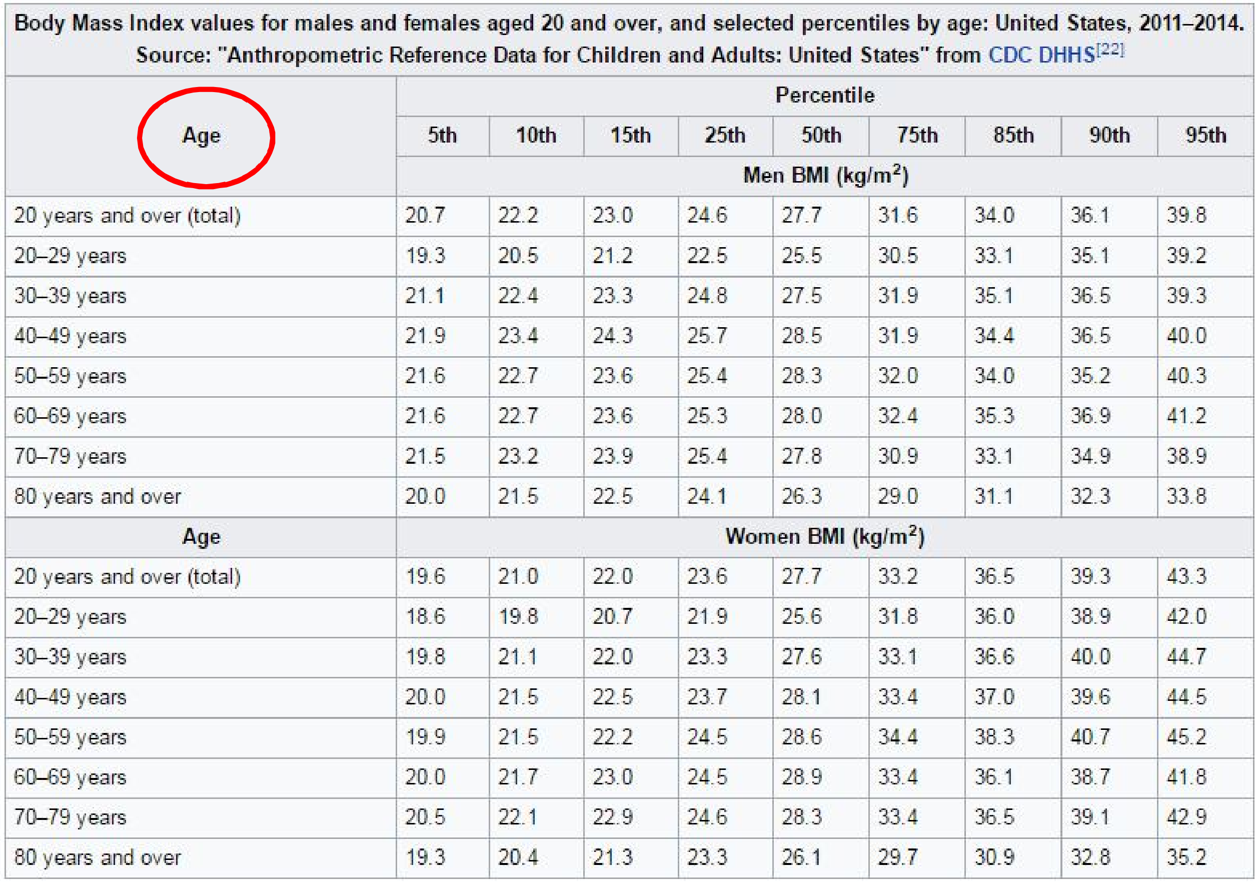
The **Body Mass Index** (BMI) is a biometric parameter defined as the body mass divided by the square of the body height.

$$\mathrm{BMI} = \frac{\mathrm{mass}_{\mathrm{kg}}}{\mathrm{height}_{\mathrm{m}}^2}$$

It is universally expressed in units of kg/m2, resulting from mass in kilograms and height in metres.



BMI and other biometric parameters



Body Mass Index values for males and females aged 20 and over, and selected percentiles by age: United States, 2011–2014.
Source: "Anthropometric Reference Data for Children and Adults: United States" from CDC DHHS[22]

Age	Percentile								
	5th	10th	15th	25th	50th	75th	85th	90th	95th
	Men BMI (kg/m^2)								
20 years and over (total)	20.7	22.2	23.0	24.6	27.7	31.6	34.0	36.1	39.8
20–29 years	19.3	20.5	21.2	22.5	25.5	30.5	33.1	35.1	39.2
30–39 years	21.1	22.4	23.3	24.8	27.5	31.9	35.1	36.5	39.3
40–49 years	21.9	23.4	24.3	25.7	28.5	31.9	34.4	36.5	40.0
50–59 years	21.6	22.7	23.6	25.4	28.3	32.0	34.0	35.2	40.3
60–69 years	21.6	22.7	23.6	25.3	28.0	32.4	35.3	36.9	41.2
70–79 years	21.5	23.2	23.9	25.4	27.8	30.9	33.1	34.9	38.9
80 years and over	20.0	21.5	22.5	24.1	26.3	29.0	31.1	32.3	33.8
Age	Women BMI (kg/m^2)								
20 years and over (total)	19.6	21.0	22.0	23.6	27.7	33.2	36.5	39.3	43.3
20–29 years	18.6	19.8	20.7	21.9	25.6	31.8	36.0	38.9	42.0
30–39 years	19.8	21.1	22.0	23.3	27.6	33.1	36.6	40.0	44.7
40–49 years	20.0	21.5	22.5	23.7	28.1	33.4	37.0	39.6	44.5
50–59 years	19.9	21.5	22.2	24.5	28.6	34.4	38.3	40.7	45.2
60–69 years	20.0	21.7	23.0	24.5	28.9	33.4	36.1	38.7	41.8
70–79 years	20.5	22.1	22.9	24.6	28.3	33.4	36.5	39.1	42.9
80 years and over	19.3	20.4	21.3	23.3	26.1	29.7	30.9	32.8	35.2

Analysis of literature

- Where to start? Which type of analysis I have to do?
- Understand and Define your ‘problem’

PS: if you have to present a document in english, search in English (you cannot use translated references)!

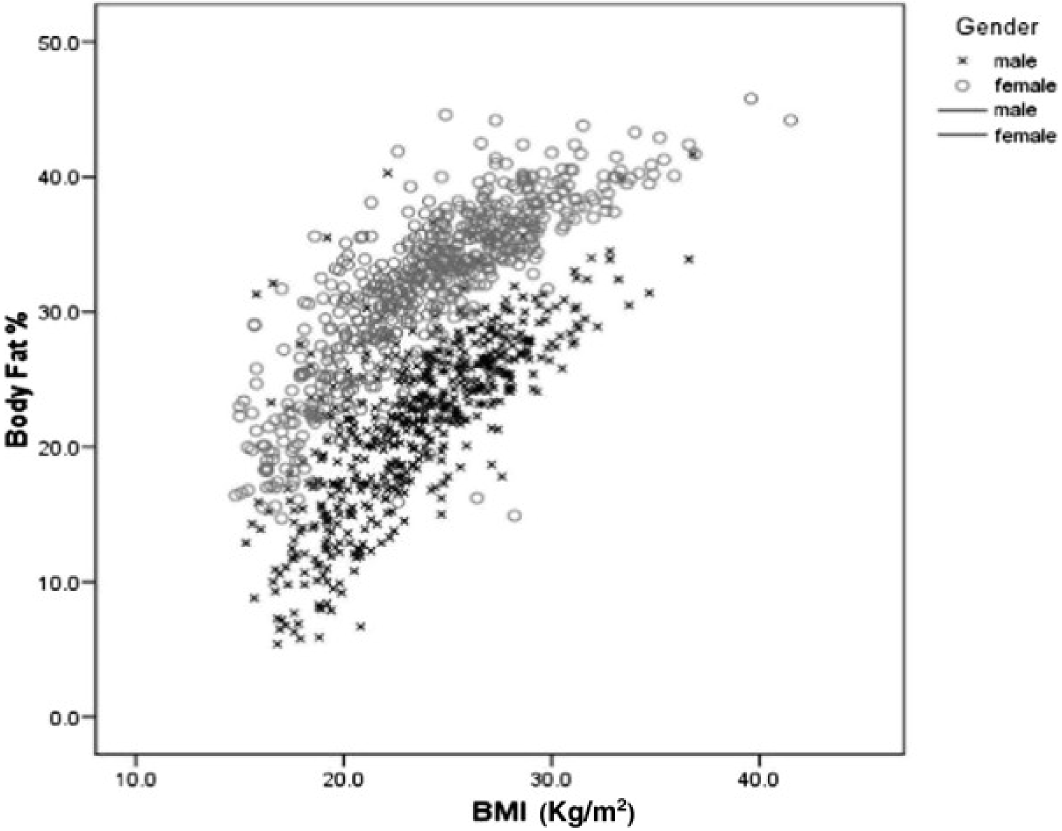
Examples in literature

Ranasinghe et al.

The relationship between **BMI and body fat percentage** (BF %) has been studied in various ethnic groups to estimate the capacity of BMI to predict adiposity.

The work aims to study the BMI–BF% relationship in a group of South Asian adults who have a different body composition compared to presently studied **ethnic groups**. We examined the **influence of age, gender** in this relationship and assessed its linearity or curvilinearity.

- A cross sectional study was conducted, where *adults of 18–83 years were grouped* into **young** (18–39 years) **middle aged** (40–59 years) and **elderly** (>60 years).
- **Pearsons correlation coefficient (r)** was calculated to see the relationship between BMI-BF% in the different age groups. **Multiple regression analysis** was performed to determine the effect of age and gender in the relationship and polynomial regression was carried out to see its' linearity

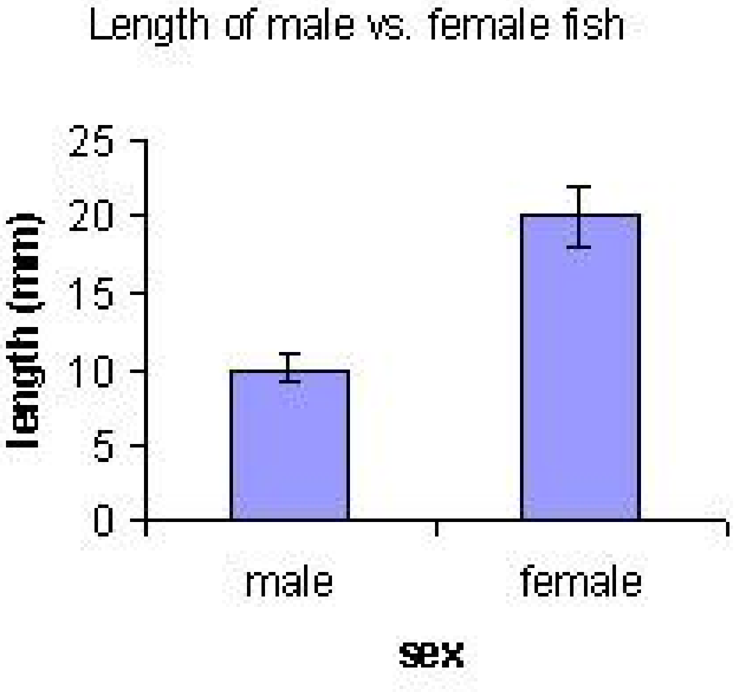


Ranasinghe et al. "Relationship between Body mass index (BMI) and body fat percentage, estimated by bioelectrical impedance, in a group of Sri Lankan adults: a cross sectional study." *BMC Public Health* (2013)

Statistics: Bar Graph

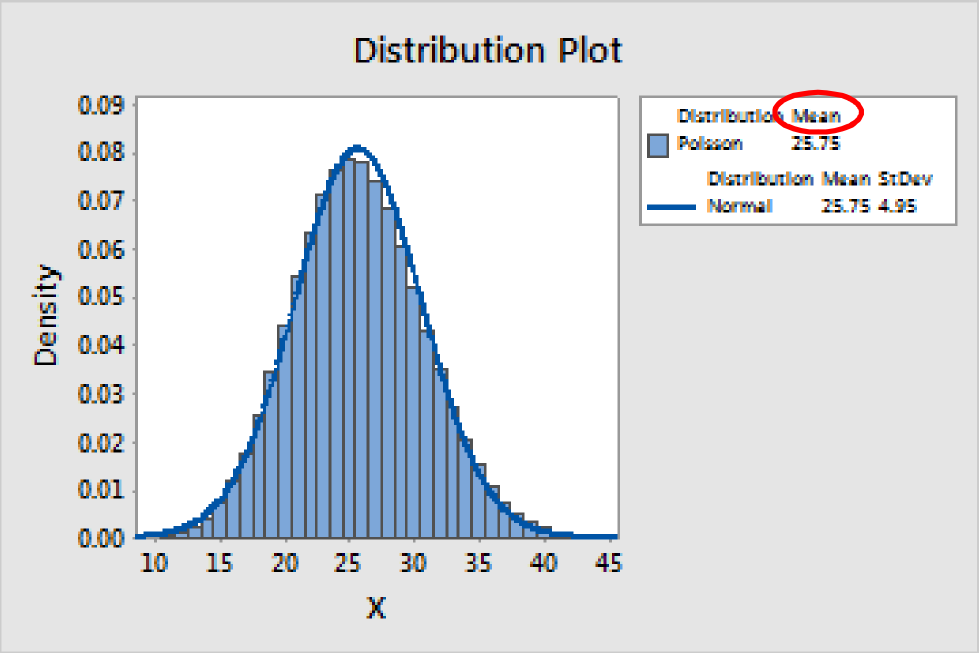
A bar graph is a chart or graph presents **grouped data** with **rectangular bars** with **lengths proportional to the values that they represent**.

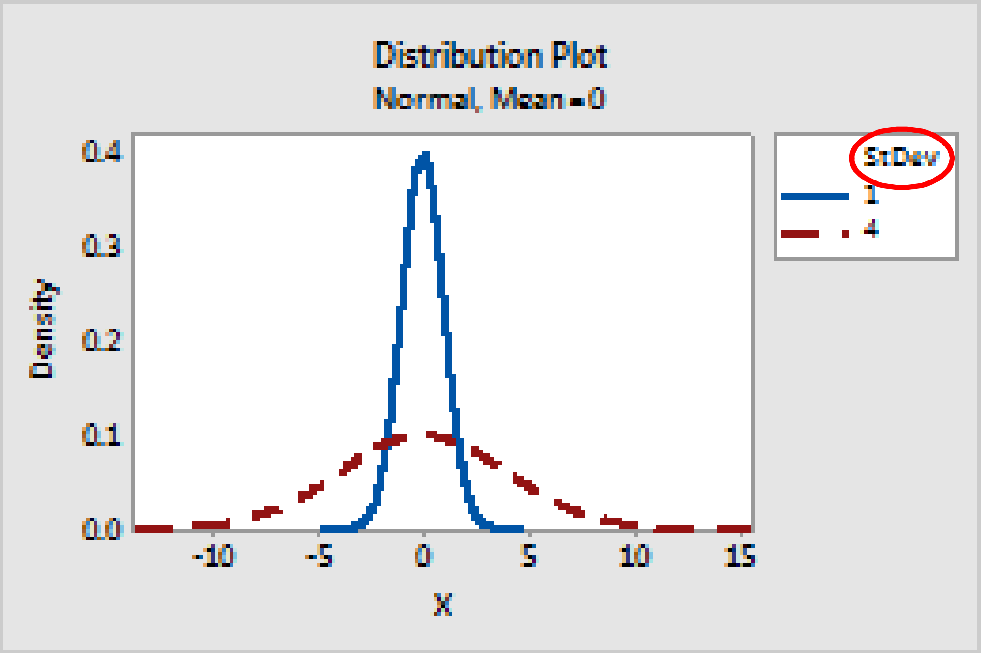
When you put **mean values**, it’s important to put the **error bars (standard deviation)**.



Statistics: Distribution plot

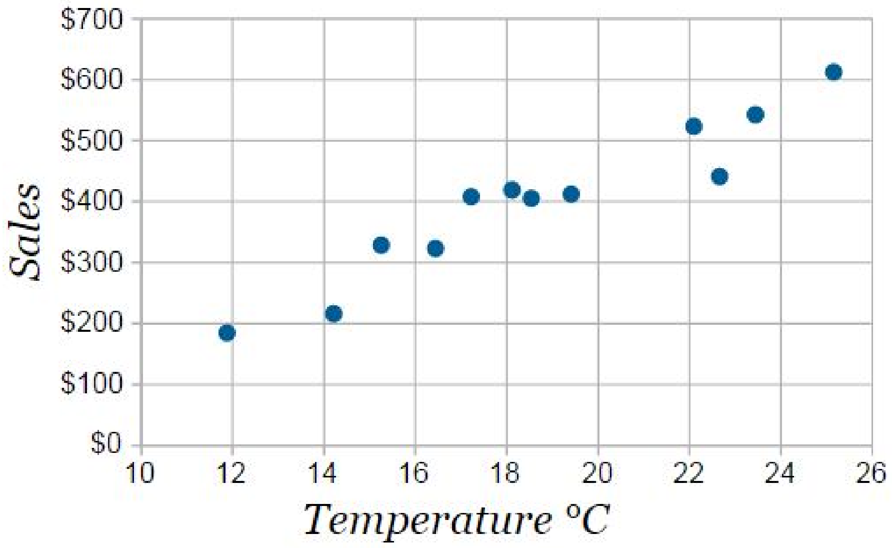
This graph plots probability density functions (PDF) which describe the probability of each data value. Usually, you specify the distribution and parameter values.

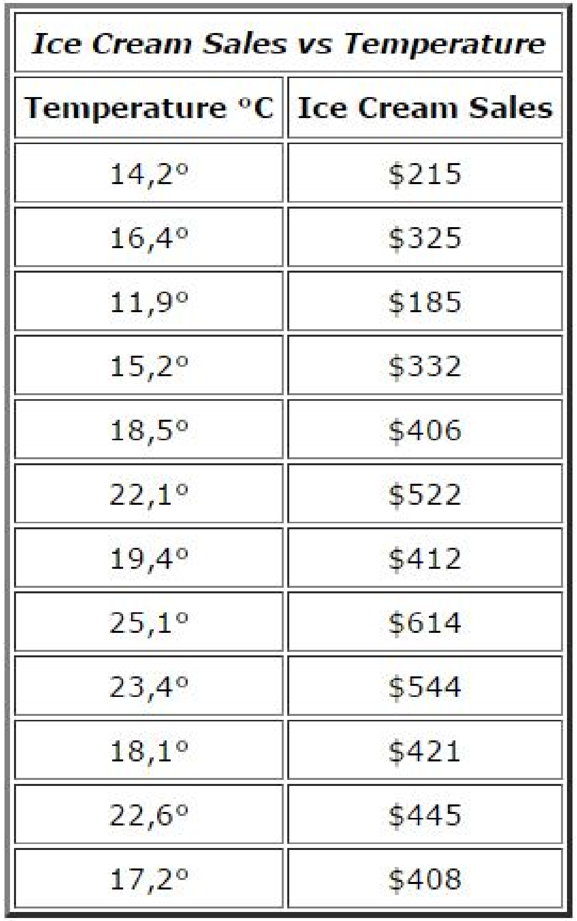




Statistics: Scatter plot

A Scatter (XY) Plot has points that show the **relationship between two sets of data**.



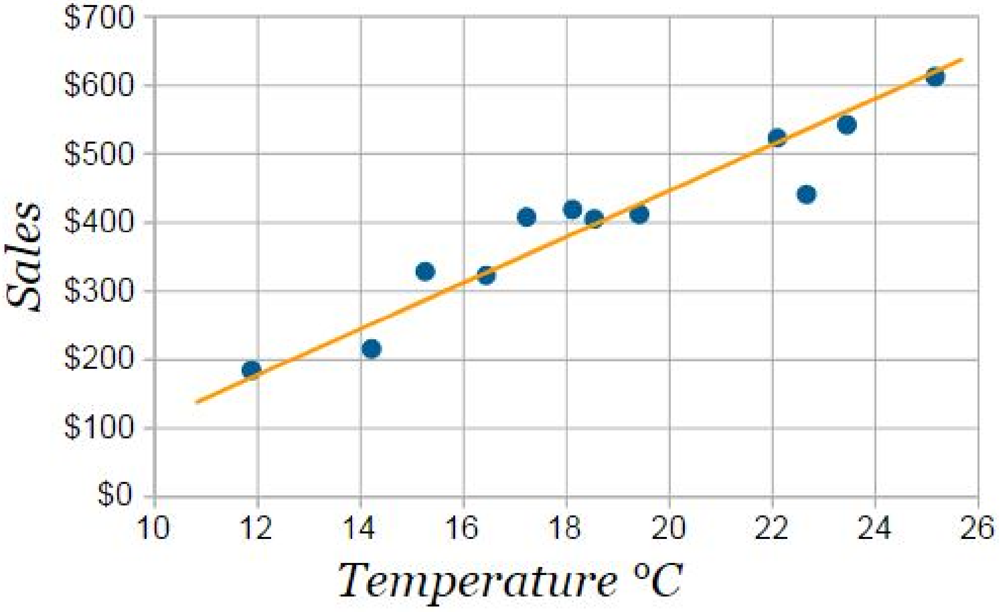


Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14,2°	$215
16,4°	$325
11,9°	$185
15,2°	$332
18,5°	$406
22,1°	$522
19,4°	$412
25,1°	$614
23,4°	$544
18,1°	$421
22,6°	$445
17,2°	$408

Scatter plot - best fit

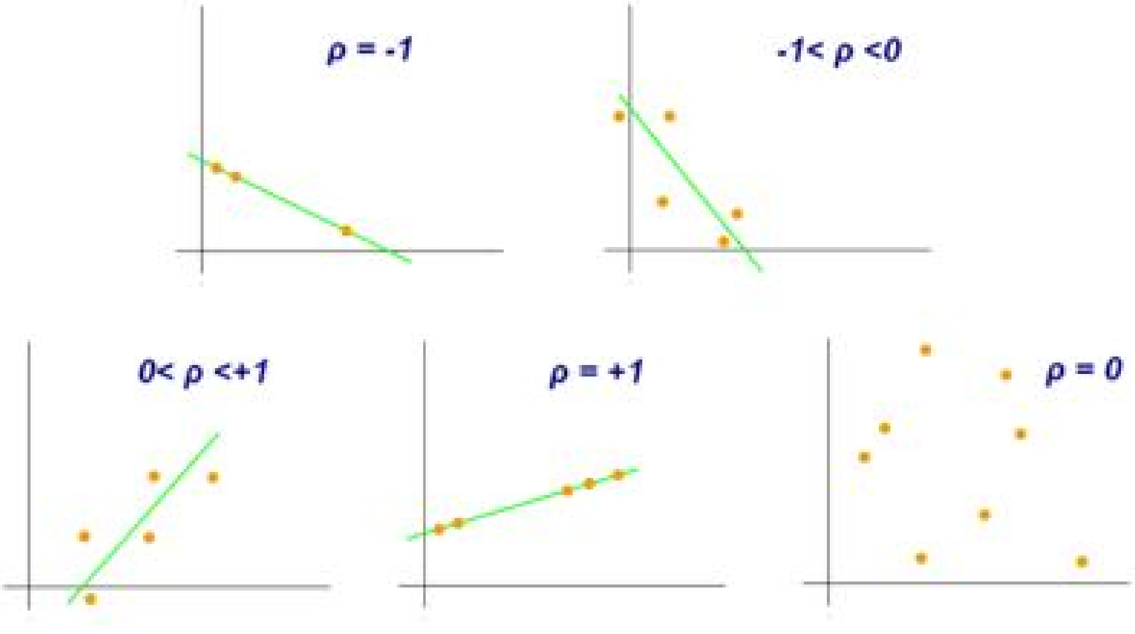
We can also draw a "Line of Best Fit" (also called a "Trend Line") on our scatter plot (**linear fit** in this case). *Try to have the line as close as possible to all points, and as many points above the line as below*.

The ‘goodness of the fitting’ is related with the **coefficient of determination - R^2 (**from 0 to 1)**,** which indicates the *proportion of the variance in the dependent variable that is predictable from the independent variable(s*). It provides a measure of *how well observed outcomes are replicated by the model*.

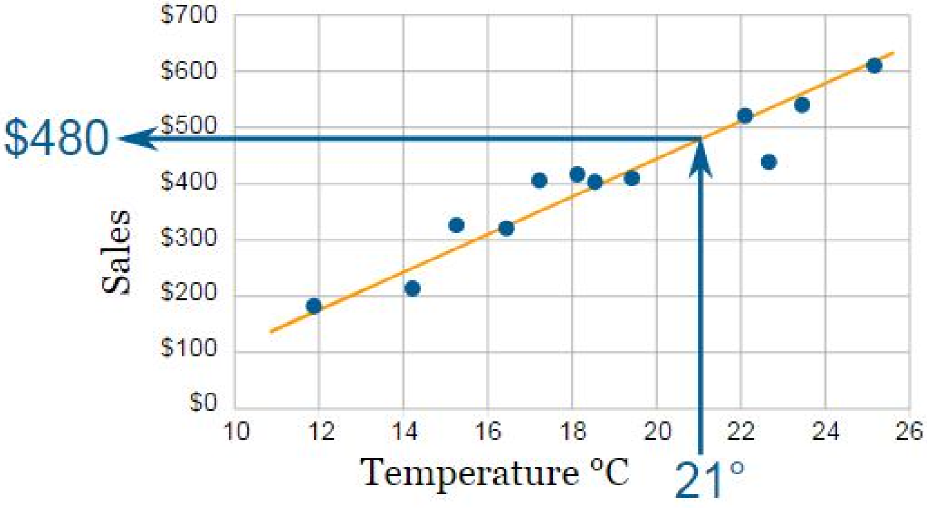


Scatter Plot – Pearson coefficient

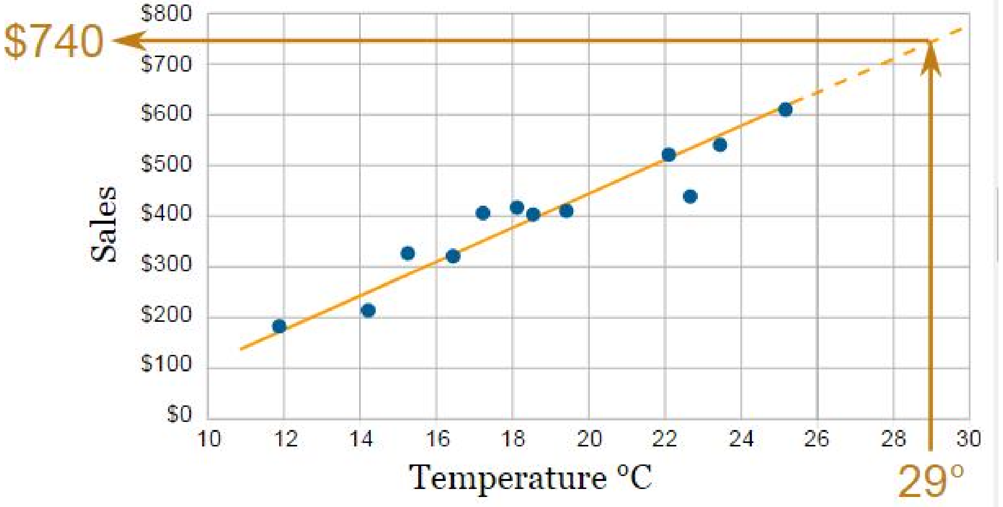
The Pearson correlation coefficient (r) is a measure of the **linear correlation between two variables X and Y**. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.



Scatter plot – Interpolation/Extrapolation



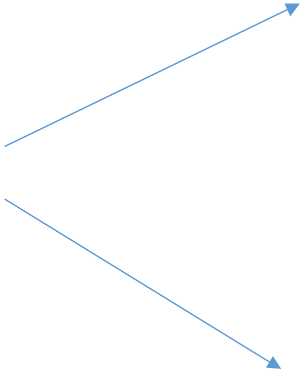
Interpolation is where we find a value *inside our set of data points*.



Extrapolation is where we find a value *outside our set of data points*.

Data Collecting

- Age
- Gender
- Weights
- Heights
- Heart beat



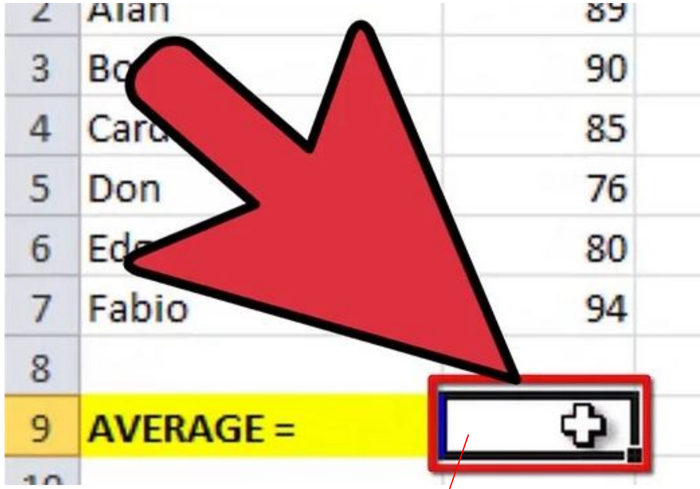
Create a Tab on Excel

Create a Dataset on Matlab:

```
names={'Sesso' 'eta' 'altezz' 'peso' 'nato' 'polso'}
mydataset=mat2dataset(MYdata, 'varnames', names)
```

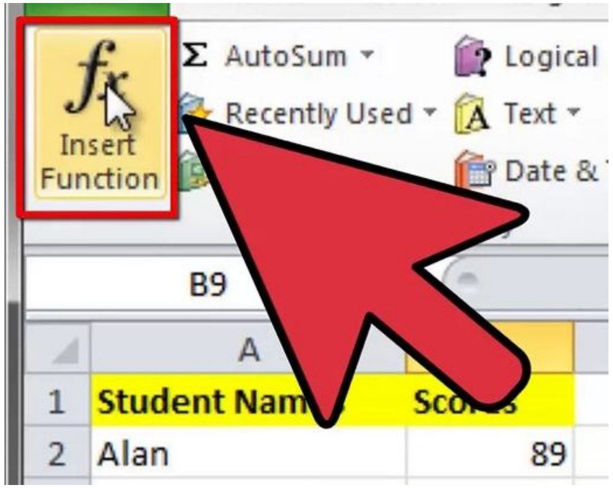
Statistic on excel

- Mean, standard deviation, pearson coeff.

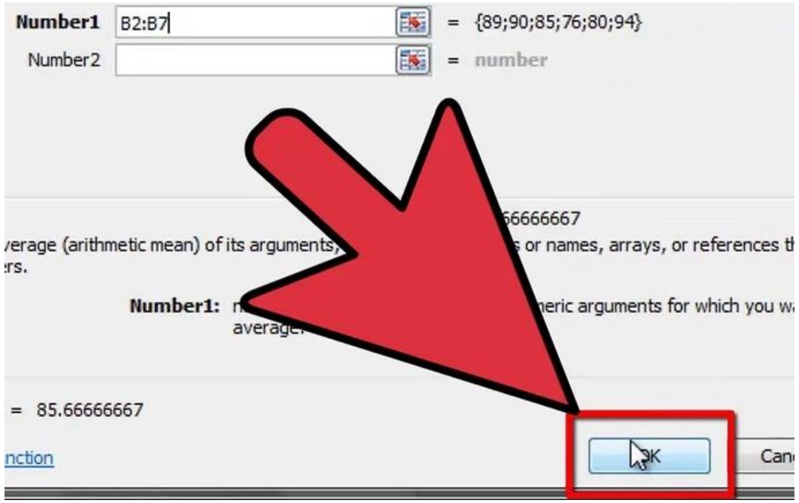


2	Alan	89
3	Bo	90
4	Car	85
5	Don	76
6	Ed	80
7	Fabio	94
8		
9	AVERAGE =	

Write =



	A	
1	Student Nam	Sco s
2	Alan	89



Number1 B2:B7 = {89;90;85;76;80;94}

Number2 = number

66666667

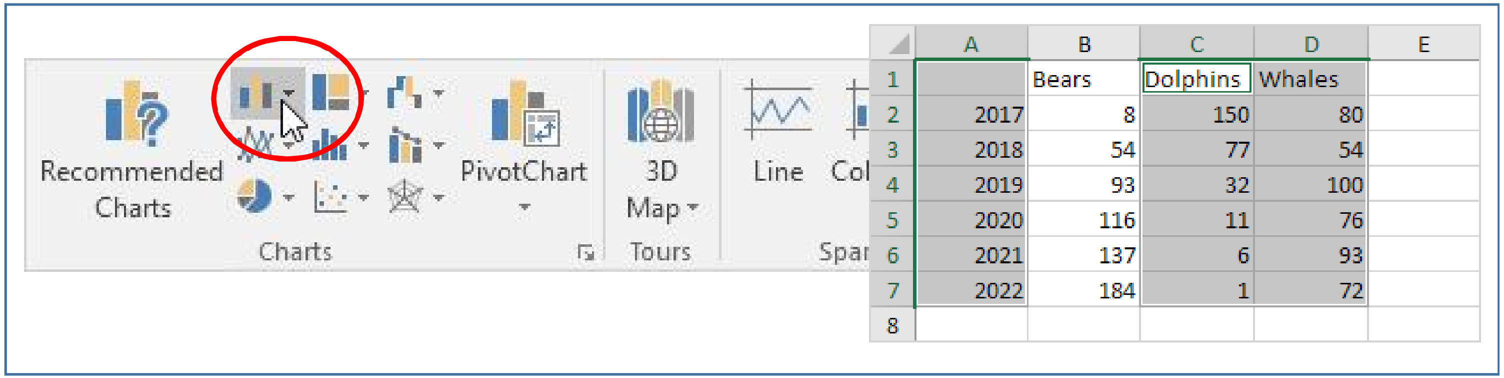
verage (arithmetic mean) of its arguments, s or names, arrays, or references th
rs.

Number1: n eric arguments for which you wa
average.

= 85.66666667

Statistic on excel (2)

- Bar graph with error bars



	A	B	C	D	E
1		Bears	Dolphins	Whales	
2	2017	8	150	80	
3	2018	54	77	54	
4	2019	93	32	100	
5	2020	116	11	76	
6	2021	137	6	93	
7	2022	184	1	72	
8					

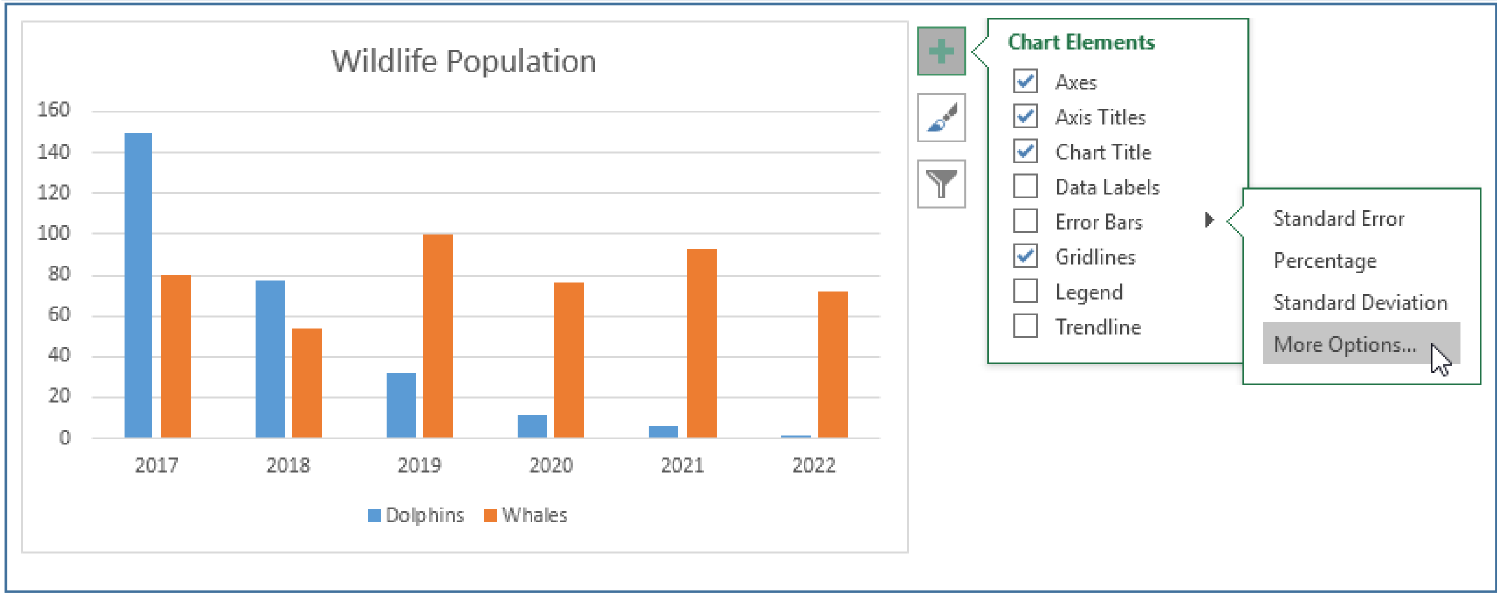
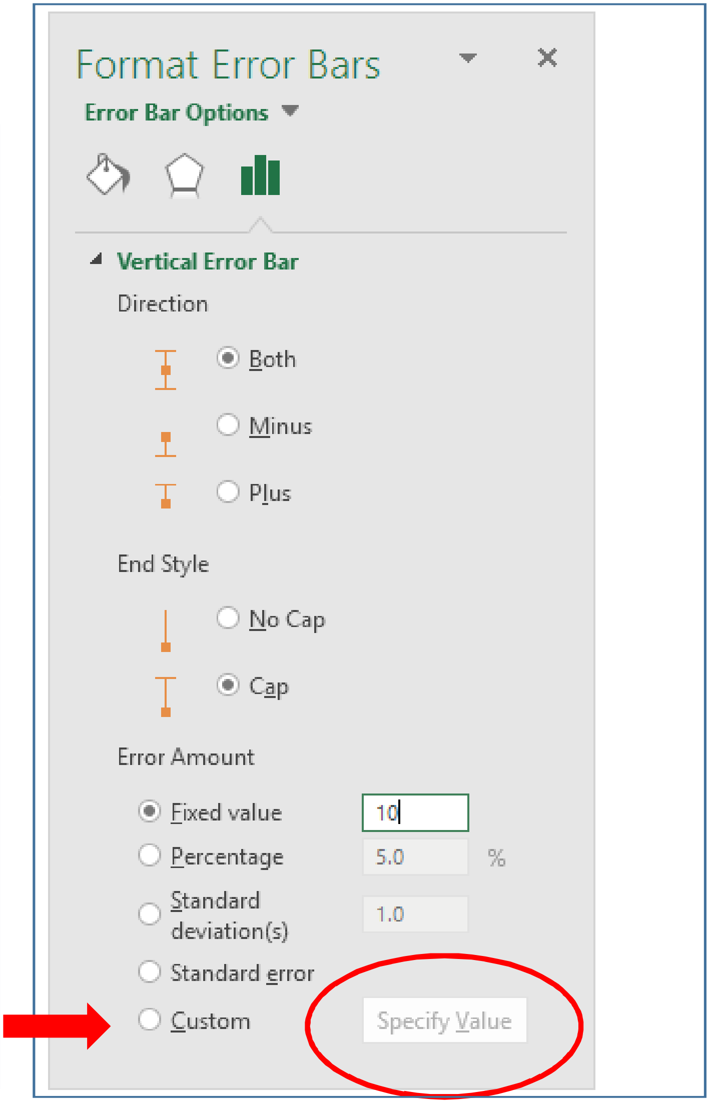


Chart Elements

- ☑ Axes
- ☑ Axis Titles
- ☑ Chart Title
- ☐ Data Labels
- ☐ Error Bars ▸
- ☑ Gridlines
- ☐ Legend
- ☐ Trendline

Standard Error
Percentage
Standard Deviation
More Options...



Format Error Bars

Error Bar Options

Vertical Error Bar

Direction

- ● Both
- ○ Minus
- ○ Plus

End Style

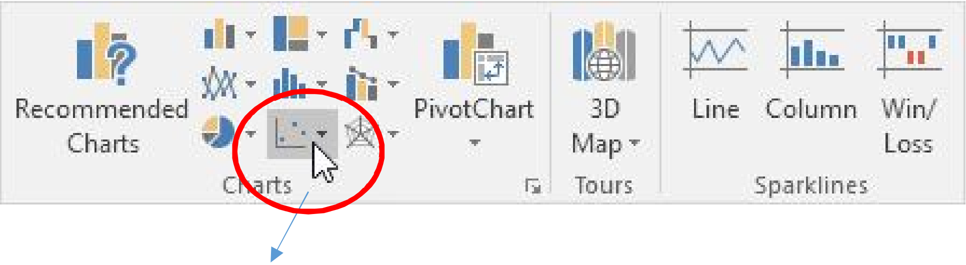
- ○ No Cap
- ● Cap

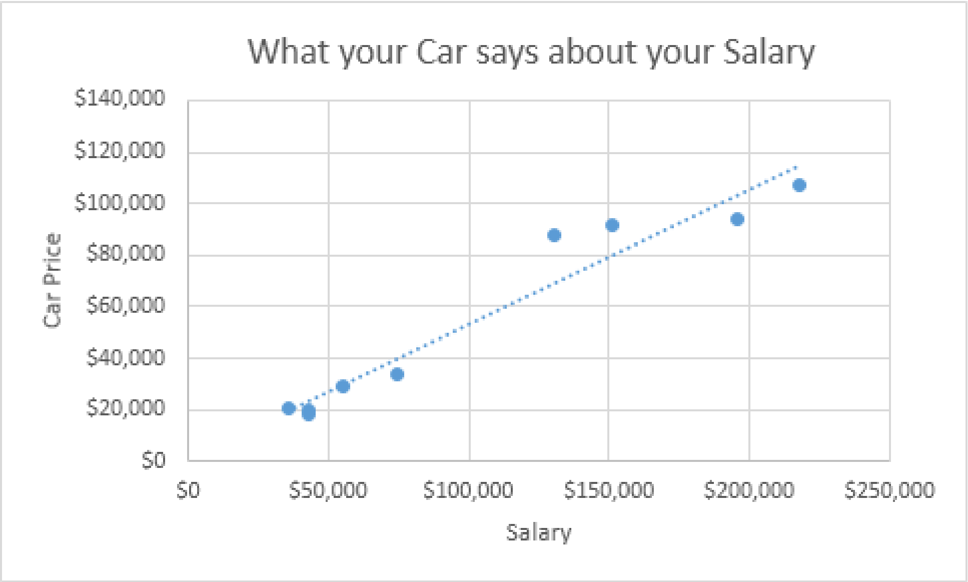
Error Amount

- ● Fixed value 10
- ○ Percentage 5.0 %
- ○ Standard deviation(s) 1.0
- ○ Standard error
- ○ Custom Specify Value

Statistic on excel (3)

- Scatter plot





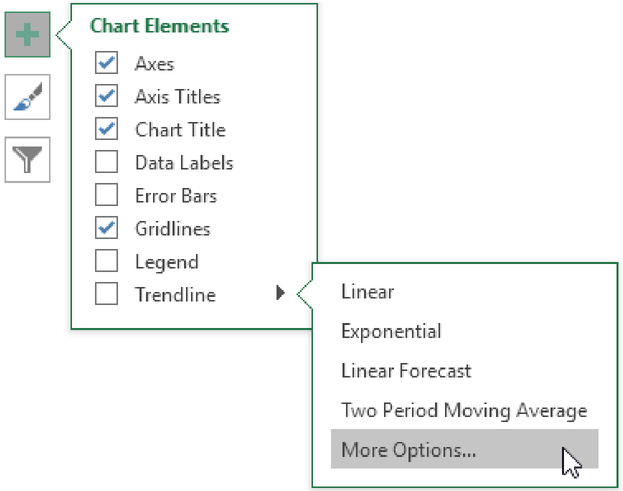
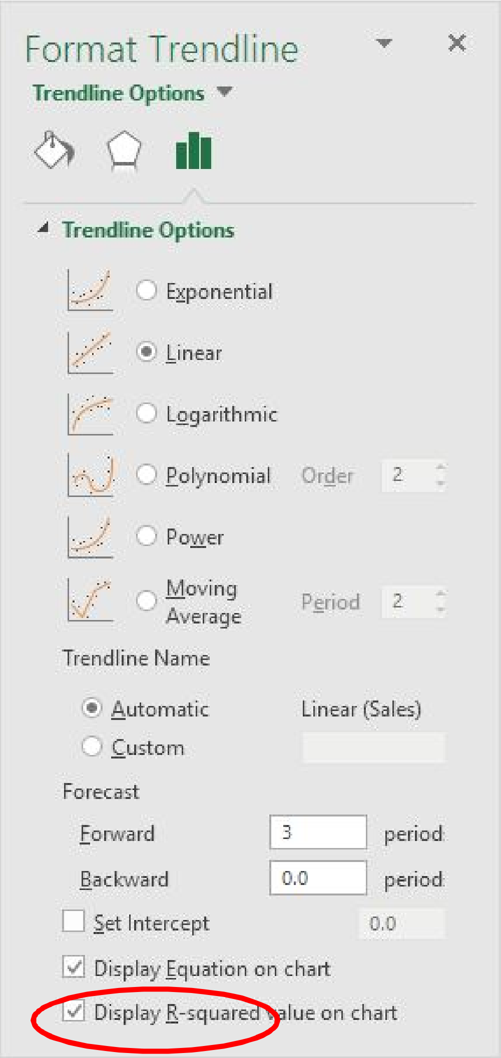


Chart Elements

- ☑ Axes
- ☑ Axis Titles
- ☑ Chart Title
- ☐ Data Labels
- ☐ Error Bars
- ☑ Gridlines
- ☐ Legend
- ☐ Trendline ▸



Format Trendline

Trendline Options

Trendline Options

- ○ Exponential
- ● Linear
- ○ Logarithmic
- ○ Polynomial   Order 2
- ○ Power
- ○ Moving Average   Period 2

Trendline Name

- ● Automatic   Linear (Sales)
- ○ Custom

Forecast

Forward 3 period

Backward 0.0 period

☐ Set Intercept 0.0

☑ Display Equation on chart

☑ Display R-squared value on chart

Statistics on Matlab: bar graph

```
bar(y)
bar(x,y)
bar(__,width)
bar(__,style)
bar(__,color)
bar(__,Name,Value)
```

bar(y) creates a bar graph with **one bar for each element in y**. If y is a matrix, then bar **groups the bars according to the rows in y**.

errorbar(y,err) creates a line plot of the data in y and draws a vertical error bar at each data point. The values in err determine the lengths of each error bar above and below the data points, so the total error bar lengths are double the err values

Statistics on Matlab: Normal Distribution Plot

Normal probability plots are used to assess whether data comes from a **normal distribution (Gaussian distribution).**

Many statistical procedures make the assumption that an underlying distribution is normal, so normal probability plots can provide some assurance that the assumption is justified, or else provide a warning of problems with the assumption.

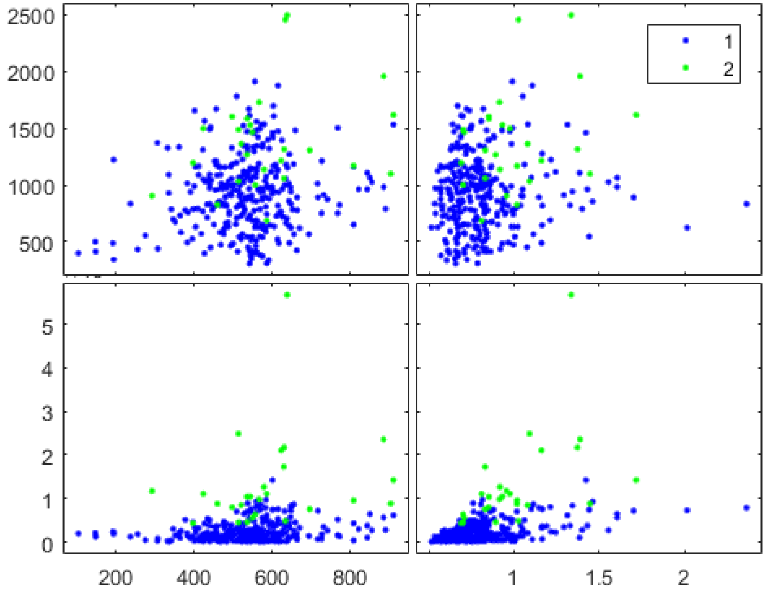
x = normrnd(mu,sigma)

x = normrnd(mu,sigma,m,n,...)

normplot(x) -> *normplot(mydataset.peso)*

Statistics on Matlab: Scatter plot

scatter(x,y)
scatter(x,y,sz)
scatter(x,y,sz,c)
scatter(__,'filled')
scatter(__,mkr)
scatter(__,Name,Value)



scatter(x,y) creates a **scatter plot with circles at the locations specified by the vectors x and y**. This type of graph is also known as a bubble plot.

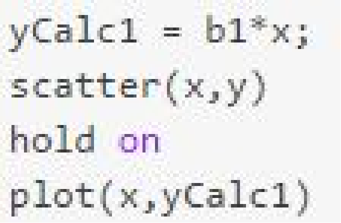
gplotmatrix(x,y,group) creates a matrix of scatter plots. Each individual set of axes in the resulting figure contains a scatter plot of a column of x against a column of y. All plots are grouped by the grouping variable group.

Statistics on Matlab: Linear Regression & correlation

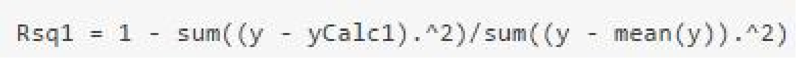
Linear regression models the relation between a **dependent variable (y)** and independent variable x.

y = b0 + b1*x

y-intercept *slope (or regression coefficient)*

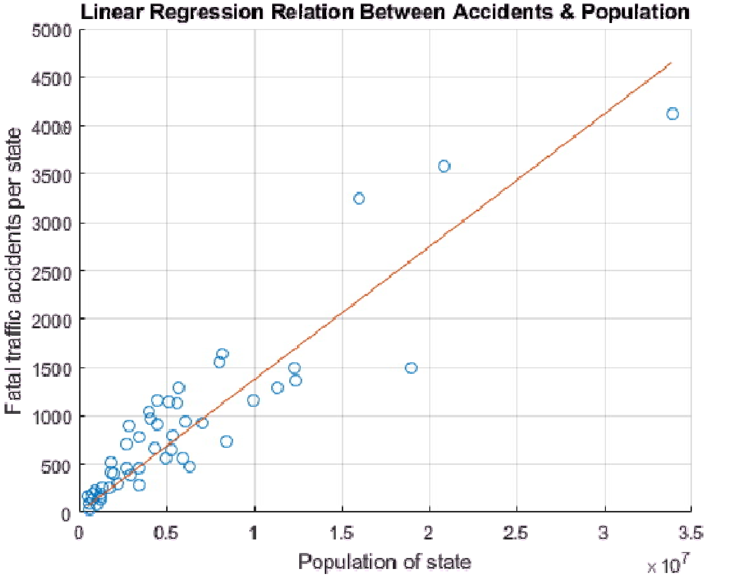


```
yCalc1 = b1*x;
scatter(x,y)
hold on
plot(x,yCalc1)
```



```
Rsq1 = 1 - sum((y - yCalc1).^2)/sum((y - mean(y)).^2)
```

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$



r = corr(x,y) calculates the correlation between columns of X and Y (default: Pearson)

t-test

A t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis.

It can be used to **determine if two sets of data are significantly different from each other.**

A t-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's t distribution.

h= ttest2(x,y)

returns a **test decision for the null hypothesis that the data in vectors x and y comes from independent random samples** from normal distributions with equal means and equal but unknown variances, using the two-sample t-test. The alternative hypothesis is that the data in x and ycomes from populations with unequal means. The result **h is 1 if the test rejects the null hypothesis at the 5% significance level**, and 0 otherwise.

ANOVA

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the **differences among group means** and their associated procedures (such as "variation" among and between groups), developed by statistician and evolutionary biologist Ronald Fisher.

In the ANOVA setting, the observed **variance in a particular variable is partitioned into components attributable to different sources of variation**.

In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups. **ANOVAs are useful for comparing (testing) three or more means (groups or variables) for statistical significance**. It is conceptually similar to multiple two-sample t-tests, but is more conservative (results in less type I error) and is therefore suited to a wide range of practical problems.

p = anova1(y)

returns the p-value for a balanced one-way ANOVA. It also displays the standard ANOVA table (tbl) and a box plot of the columns of y. anova1 tests the hypothesis that the samples in y are drawn from populations with the same mean against the alternative hypothesis that the population means are not all the same.

p= anova2(y,reps) returns the p-values for a balanced two-way ANOVA for comparing the means of two or more columns and two or more rows of the observations in y.