Basic Algorithms

iCB2 – Introduction to Computational Biology and Bioinformatics
November 11, 2015

Emidio Capriotti
http://biofold.org/

Biomolecules
Folding and
Disease

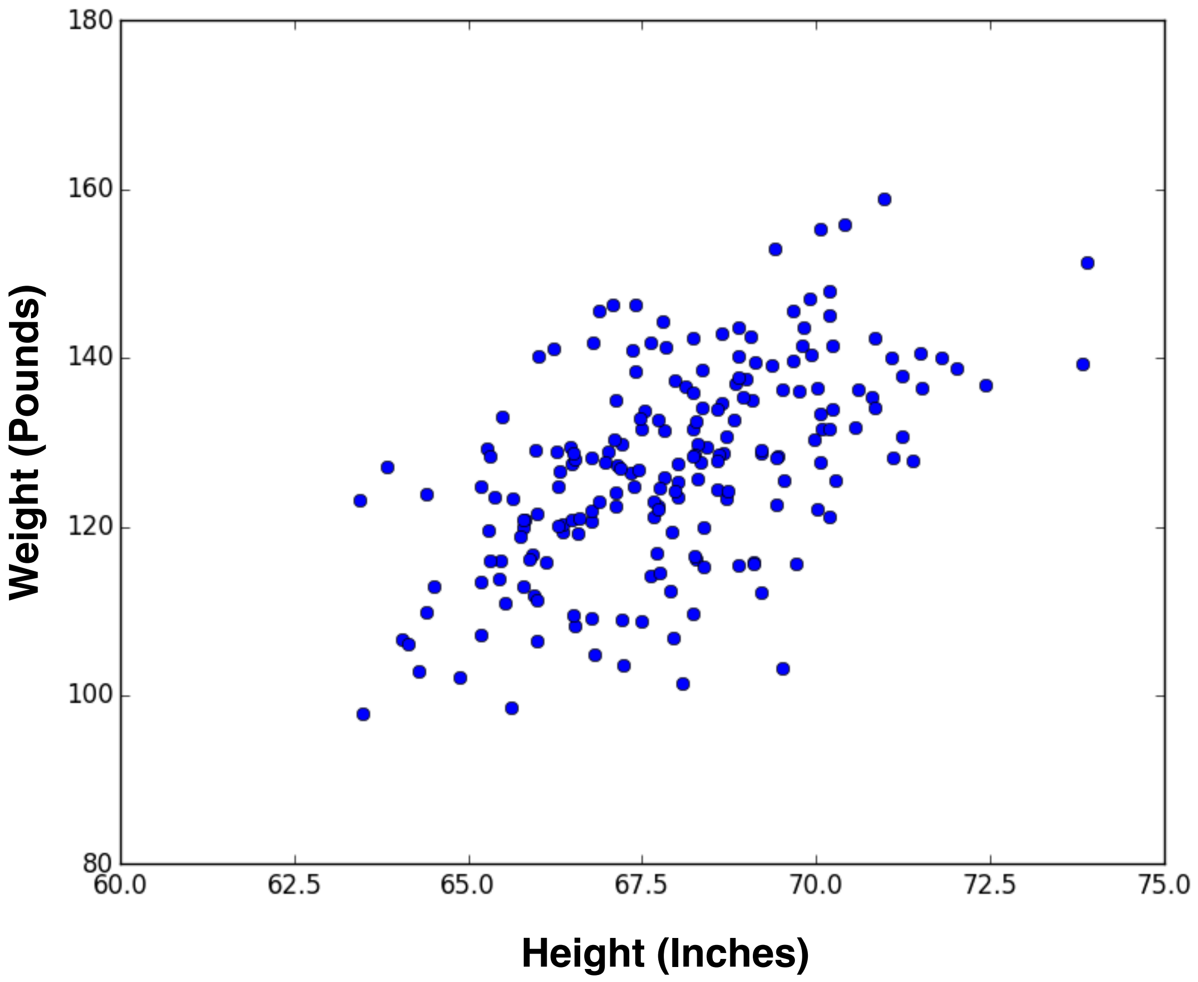
Institute for Mathematical Modeling
of Biological Systems
Department of Biology

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

A basic predictive model

The linear regression model is one of the simplest predictive model that is used in different fields.

For example: it is known that the weight of a person is correlated with its height. Can we build a model to predict the weight of a new person for which the know the height?



Fitting a linear regression

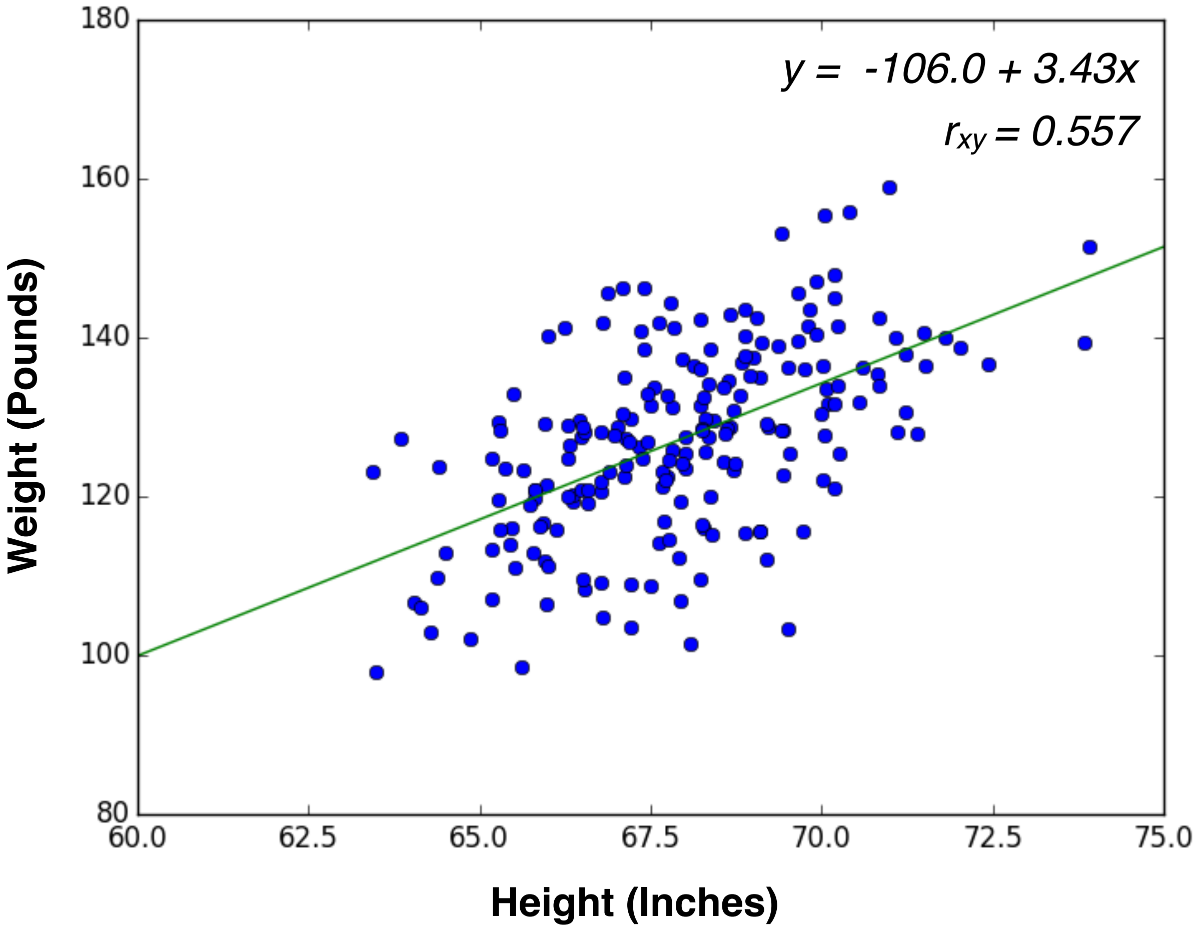
Given a list of points A=[(x_1,y_1), (x_2,y_2) ….. (x_n,y_n)], we calculate the parameters $\tilde{\alpha}$ *and* $\tilde{\beta}$ *of the* line $y = \alpha + \beta x$ that minimizes

$$Q(\alpha,\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

$$\tilde{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\tilde{\alpha} = \bar{y} - \tilde{\beta}\,\bar{x},$$

$$r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

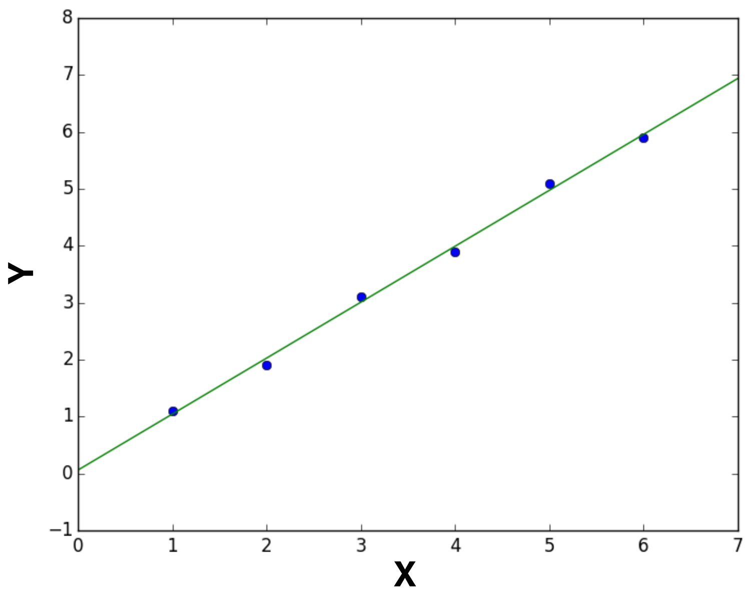


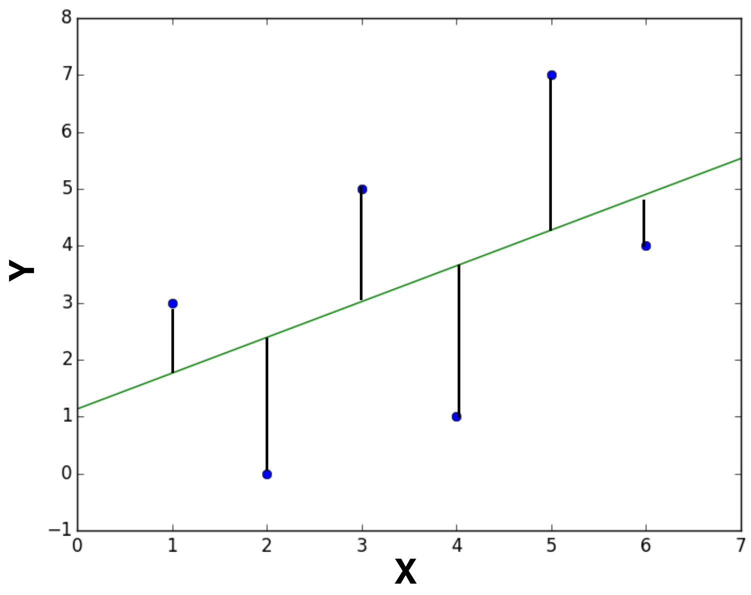
The linregress function

Import linregress from spicy.stats and calculate calculate the fitting curve

```
>>> from scipy.stats import linregress
>>> import numpy as np
>>> x = np.array([1,2,3,4,5,6])
>>> y = np.array([1.1,2.2,2.9,3.98,5.2,6.1])
>>> reg=linregress(x,y)
>>> print reg
LinregressResult(slope=0.98285714285714287,
intercept=0.060000000000000053, rvalue=0.99838143945702995,
pvalue=3.9274872444222332e-06, stderr=0.027994168488950467)
```

what happen if y = [3,0,5,1,7,4]. is this a better fitting? why?





Use matplotlib to plot

Import matplotlib and plot the points and fitting curve.

```
>>> import matplotlib.pyplot as plt
>>> yp=reg[0]*x+reg[1]
>>> plt.plot(x,y,’o’,x,yp,’-’)
>>> plt.show()
```

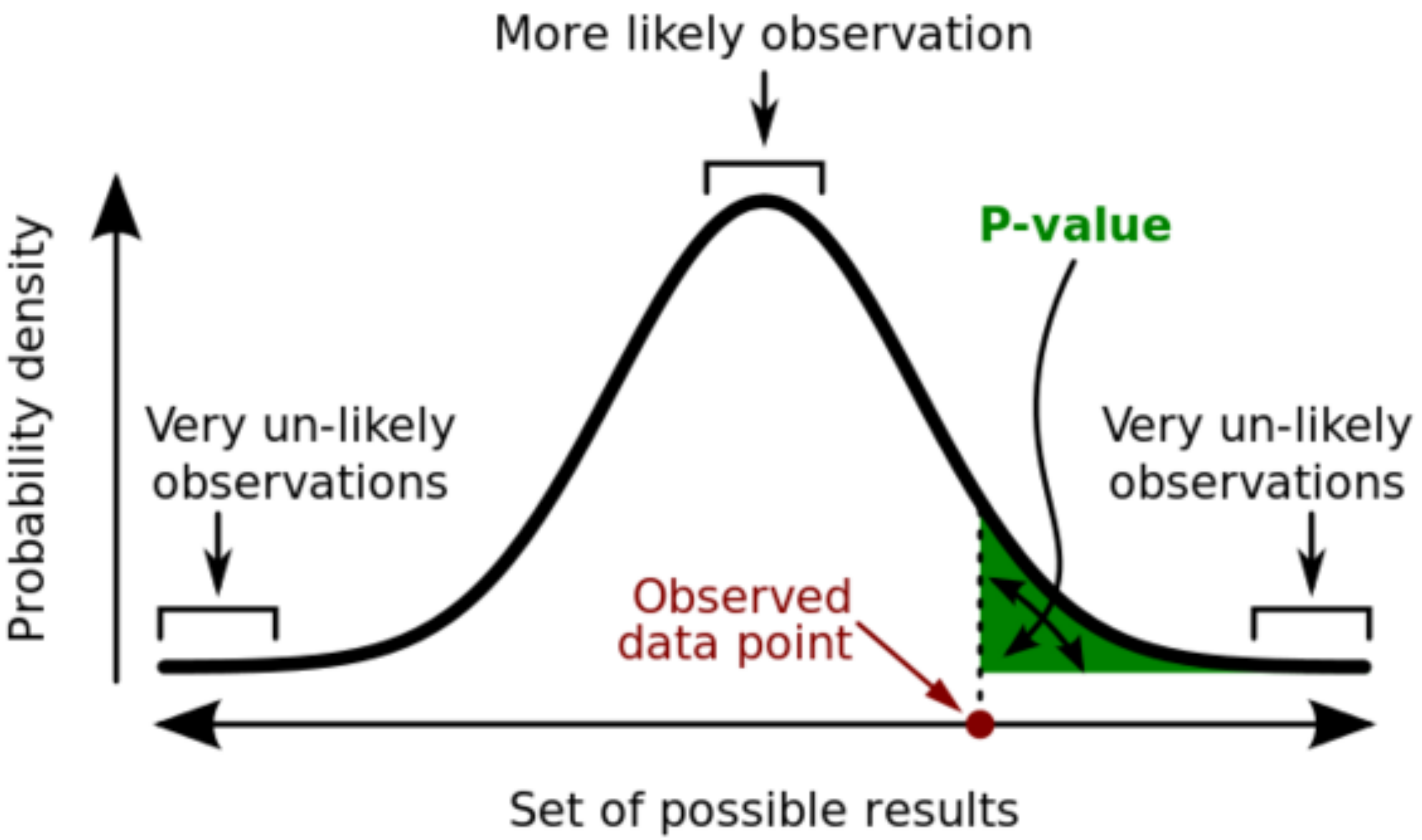
Exercise:

Write a python script that reads a file containing two columns of data (x,y) and calculate the linear regression curve and plots both the points an the regression curve.

For this exercise download the data using the command *wget* from http://biofold.org/courses/docs/data_hw.txt

The p-value

In statistics, the p-value is a function of the observed results that is used for testing a statistical hypothesis. More specifically, the p-value is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed.

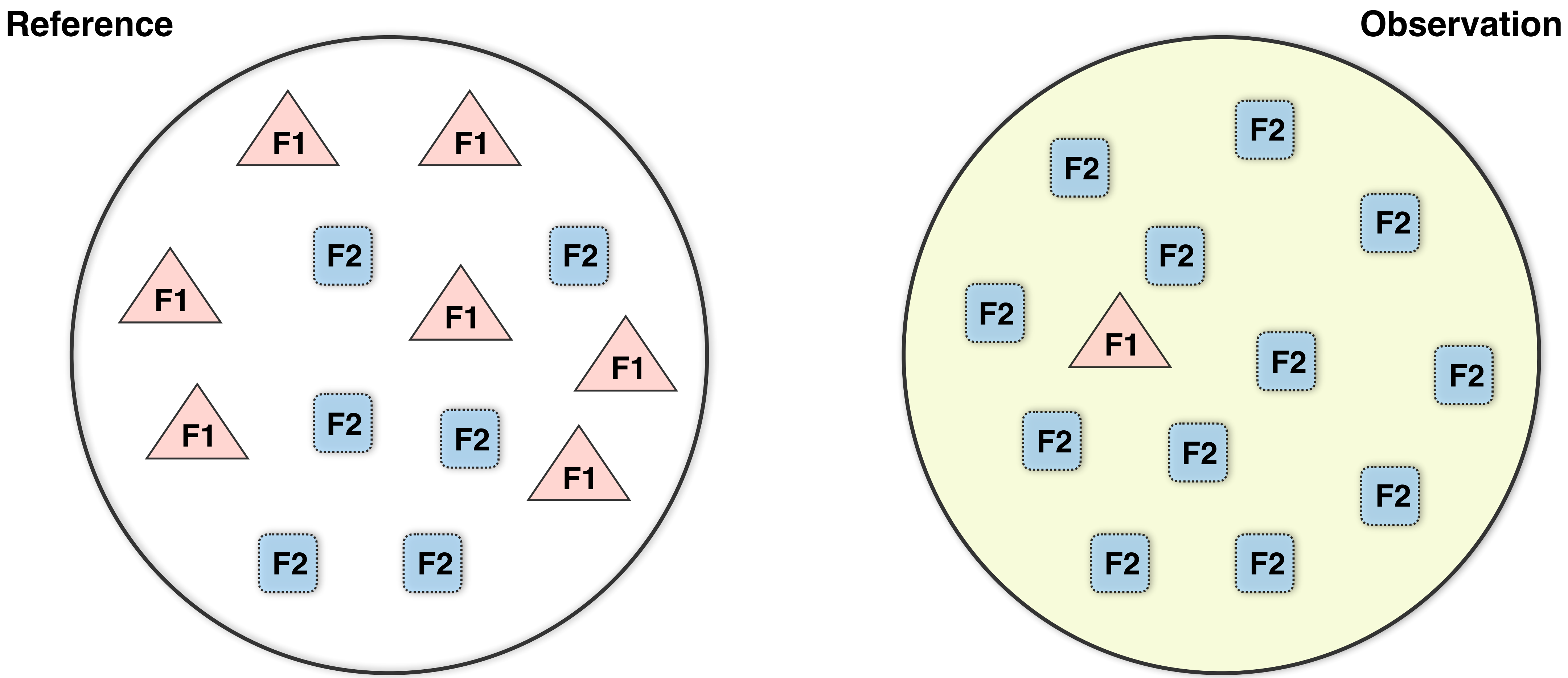


https://en.wikipedia.org/

The enrichment analysis

It is a method to identify classes of genes or proteins or functions that are over-represented in a large set of genes or proteins.

The gene set enrichment analysis is used to understand the functional profile of a set of genes.



Contingency table

In statistics, a contingency table is a type of table in a matrix format that displays the frequency distribution of the variables.

They are heavily used in survey research, business intelligence, engineering and scientific research.

	Function 1	Function 2	Row Total
Reference	7	6	13
Observation	1	12	13
Column Total	8	18	26

Fisher's exact test

Fisher's exact test is a statistical significance test used in the analysis of contingency tables.

It is one of a class of exact tests, so called because the significance of the deviation from a null hypothesis (p-value) can be calculated exactly.

For a generalized contingency tables:

	Function 1	Function 2	Row Total
Reference	a	b	a+b
Observation	c	d	c+d
Column Total	a+c	b+d	a+b+c+d=n

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\;(c+d)!\;(a+c)!\;(b+d)!}{a!\;\;b!\;\;c!\;\;d!\;\;n!}$$

Fisher’s test in python

The fisher_exact function is contained in the spicy.stats module and takes in input a contingency matrix. The function returns *odd ratio* and *p-value.*

```
>>> from scipy.stats import fisher_exact
>>> import numpy as np
>>> cm=np.array([[6,7],[12,1]])
>>> ft=fisher_exact(cm)
>>> print ft
(0.071428571428571425, 0.030205949656750501)
```

Exercise:

Write a python script that reads two file containing a columns with alleles carried by each individual. Use Fisher’s exact test verify if an allele is over represented in one of the populations.

For this exercise download the data using the command *wget* from
http://biofold.org/courses/docs/pop1_allele.txt
http://biofold.org/courses/docs/pop2_allele.txt