**Probability distributions and Bayesian Networks**

*Accessing data from excel:*

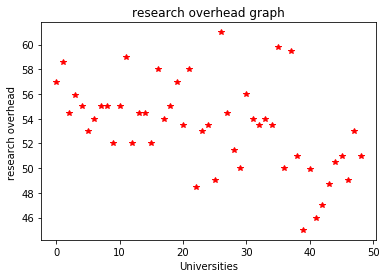
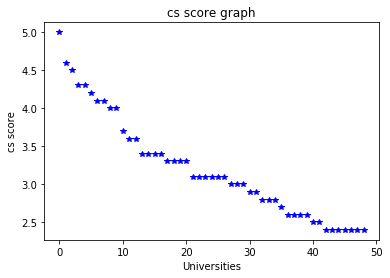
Xlrd: It is used to read excel data. Using inbuilt functions in xlrd, it is possible to extract data columnwise forming a list. Convert this list into array using numpy’s inbuild command “asarray’.

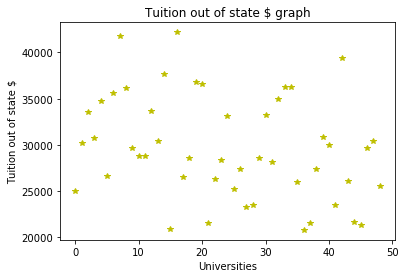
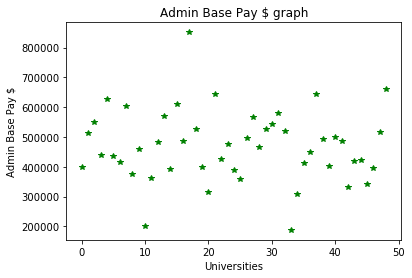
So as this point I have 4 variables CS\_score, Ad\_base, Res\_Over and Tuition.

1. *Calculating Mean, Variance and Standard-deviation:*

Numpy have set of inbuild functions i.e. numpy.mean, numpy.var, numpy.std to calculate mean, variance and standard deviation respectively. Print this values at output.

Mean is sample average of numbers in a variable. So, as we can observe range of CS\_score is 2.4 to 5.0 and mean is about 3.214 which tells average value of this variable. Standard deviation gives variation of data values from the mean. Therefore, 3\*σ = 3\*0.669 = 2, gives information that 99% of data lies in (μ-3\*σ, μ+3\*σ) i.e. (2, 5.214).



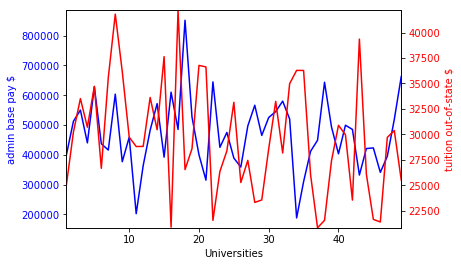
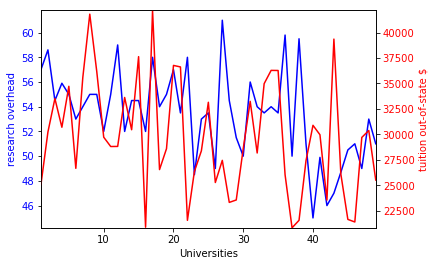
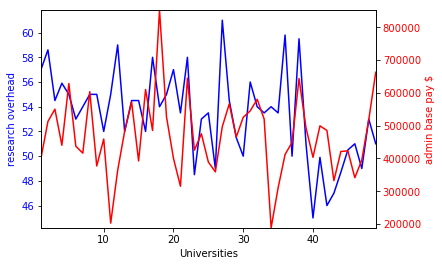
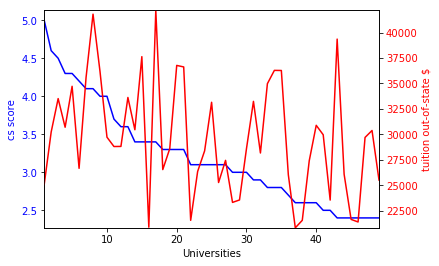
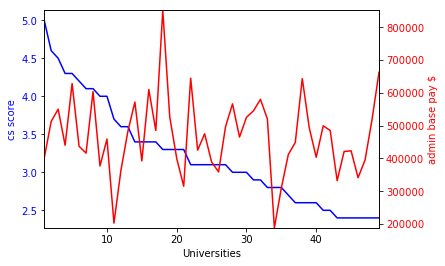
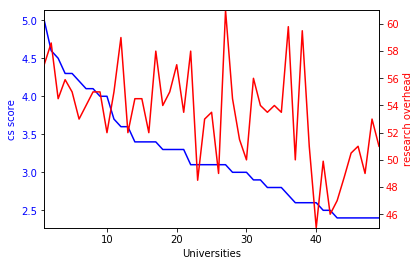


1. *Calculating Co-variance and Cor-relation matrix:*

Used inbuild function in numpy called numpy.cov, numpy.corref for covariance and correlation coefficient calculation.

This will give us 4\*4 matrix for given 4 variables. Here diagonal elements are covariance of variable with itself and hence diagonal values should match with variance of that variable calculated. Correlation of any element with itself will give value 1 (this is clear from formula itself). Therefore, diagonal elements in correlation matrix is 1.

We can also observe if variables are correlated using pairwise plot of all variables as shown below:



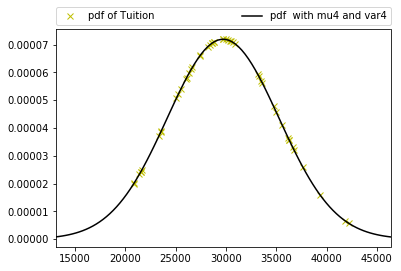
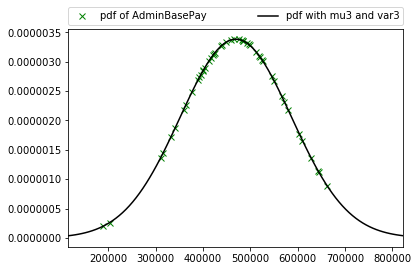
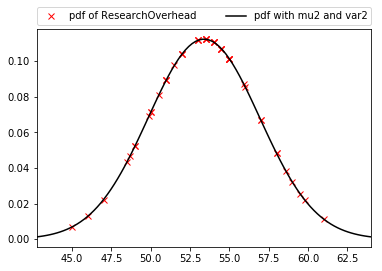
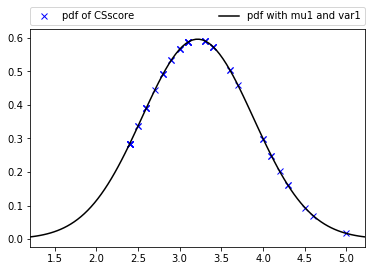
Using correlation matrix and the plot of pairwise data following is the conclusion:

Most correlated variables: CS Score and Research Overhead.

Least correlated variables: CS score and Admin base pay ($).

1. *Log likelihood of data:*

Continuous random variables with gaussian distribution for CS score, Research Overhead, Admin base pay, Tuition are as follows:



Formula for multivariate gaussian distribution:

If variables are independent, joint pdf distribution is product of each pdf distribution.

Thus: . Calculate individual pdf for all variables (using scipy.stats.norms.pdf) followed by logarithm and summation. Here we obtain loglikelihood of ***X***. Loglikelihood of given variables assuming they are independent is: **-1315.099**

Similarly, loglikelihood of multivariate distribution using multivariate\_normal.pdf in python we get **-1262.327**

**Minimizing the negative log-likelihood of our data with respect to**our parameters**is equivalent to minimizing the mean squared error between the observed**y**and our prediction thereof.**

1. *Bayesian Network:*

In **probability** theory and statistics, **Bayes**' **theorem** describes the **probability** of an event, based on prior knowledge of conditions that might be related to the event.

Bayesian network shows probabilistic dependence of variables with each other. Bayes theorem considers the problem of overfitting.

For this given problem, I modelled a Directed acyclic Graph (DAG) as follows:

Using this graph, joint pdf becomes; p(**X**)=p(X3|X4)\*p(X4|X1)\*p(X1|X2)\*p(X2). This is forming Markov chain of variables as explained in conditional probabilities using DAG (1) below.

Therefore, Bayesian loglikelihood becomes=**-2267.347**

For this example, BNgraph becomes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CS score | Research Overhead | Admin base pay | Tuition out-of-state |
| CS score | 0 | 0 | 0 | 1 |
| Research Overhead | 1 | 0 | 0 | 0 |
| Admin base pay | 0 | 0 | 0 | 0 |
| Tuition out-of-state | 0 | 0 | 1 | 0 |

***Conditional probabilities based on 3 types of adjacent triplets in DAG*:**

First is example of **Markov chain**, i.e. parameter Y depends on X and similarly Z depends on Y. *f(*X, Z|Y*)=f(*X|Y*)f(Z|Y)*, i.e. X and Z are conditionally independent given Y.

Second is based similar to first, i.e. X and Z both depends on Y to happen and so X and Z are conditionally independent given Y.

Third example, however, shows how X and Z are marginally independent of Y. i.e. *p*(X,Y,Z) = *p*(Y|X,Z)*p*(X)*p*(Z).

There are other types of network as well. If none of variables depend on other to occur, then these are marginal independent variables as we solved for question 3.

***Results:***

UBit name= jthakur

person no= 50206922

mu1=3.214

mu2=53.386

mu3=469178.816

mu4=29711.959

var1=0.448

var2=12.588

var3=13900134681.701

var4=30727538.733

sigma1=0.669

sigma2=3.548

sigma3=117898.832

sigma4=5543.243

CovarianceMat= [[ 0.457 1.106 3879.782 1058.480]

[ 1.106 12.850 70279.376 2805.789]

[ 3879.782 70279.376 14189720820.903 -163685641.258]

[ 1058.480 2805.789 -163685641.258 31367695.790]]

CorrelationMat= [[ 1.000 0.456 0.048 0.279]

[ 0.456 1.000 0.165 0.140]

[ 0.048 0.165 1.000 -0.245]

[ 0.279 0.140 -0.245 1.000]]

logLikelihood=-1315.099

BNgraph= [[0 0 0 1]

[1 0 0 0]

[0 0 0 0]

[0 0 1 0]]

BNlogLikelihood=-2267.347

logLikelihood using gaussian multivariate normal distribution=-1262.327

**References:**

1. <https://www.ics.uci.edu/~rickl/courses/cs-171/cs171-lecture-slides/cs-171-17-BayesianNetworks.pdf>
2. <https://en.wikipedia.org/wiki/Bayesian_network>
3. <https://en.wikipedia.org/wiki/Bayes%27_theorem>