Assignment - 1

TITLE:
Main Statistical Measures

PROBLEM STATEMENT:
Compute estimators of main statistical measures lik
Mean, Variance, standard deviation, Covariance, Correlation and standard error with respect to any example display graphically.

OBJECTIVE:
To ~~undou~~ understand modern computational methods used in statistics.

OUTCOMES:
Identify suitable method of statistics on the given data to solve problem of any heuristic approach of prediction.

PREREQUISITE:
1) Basics of statistics
2) Any programming Language (Ex. Python)

THEORY:
1) Mean:
The most commonly used measure of central tendency of a set of observation is the mean of observations. Mean is their average. It is equal to sum divided by no. of observation in a set. The sample mean is denoted by:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Where Σ is summation notation. The summation extends over all data points. When our observation set constitutes an entire population instead of denoting mean $\bar{x}$ we use symbol μ (mu). For population, we use N as no. of elements instead of n. It is defined as follows:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Population vs Sample

A population refers to summation of all elements of interest to researcher.

→ Examples: The no. of people in country, the no. of hedge fund in US or even total no. of CFA candidates.

2) Variance

The variance of set of observations is average squared deviation of the data points from their means.

When our data constitutes sample, variance is denoted by s^2 and averaging is done by dividing the sum of the squared deviations from mean by $n-1$. When our observations constitute an entire population, the variance is denoted by σ^2 and averaging done by dividing by N.

Sample variance :

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i = \mu)^2}{N}$$

3) Standard Deviation

The standard deviation of a set of observation is the square root of the variance of the set.

The standard deviation of a sample is the square root of sample variance, and the standard deviation of population is the square root of variance of population.

Sample : $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

Population : $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{n-1}}$

4) Covariance

Covariance is a measure of how closely two assets move together. In this, we focus on relationship between deviations of some two variables rather than the deviation from mean of one variable.

If the mean of random variables x and y are known, then covariance between the two random variables are as follows:

$$\hat{\sigma}_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)$$

If we don't know mean, then eq. is:

$$\hat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{u}_x)(y_i - \hat{u}_y)$$

5) Correlation

Correlation is a concept that is closely to covariance in the following way:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

It ranges between 1 to -1 and is therefore much easier to interpret. If it is +1 then variables are perfectly correlated, if 0 then uncorelated and if -1 then move in perfectly opposite direction.

6) Standard Error

The mean square error MSE is unbiased estimator of variance of population errors E which is σ^2.

$$MSE = \frac{SSE}{n-(k+1)} = \frac{\sum_{i=1}^{n} (y_i - \bar{y}_j)^2}{n-(k+1)}$$

The standard error of estimate is:

$$S = \sqrt{MSE}$$

7) Python Numpy Package

Numpy stands for numerical python. It supports N-dimensional array objects that can be used for processing multi-dimensional data. Support different data-types. Using numpy we can perform:

i) Mathematical and logical operations on array
ii) Fourier Transforms
iii) Linear algebra operations
iv) Random number generations

Syntax → numpy.array(object)

```
import numpy as np
In [2]: x = np.array([2,3,4,5])
In [3]: print(type(x))
<class 'numpy.ndarray'>
In [4]: print(x)
[2, 3, 4, 5]
```

8) Python Pandas Package

Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool built on top of Python. Series are one-dimensional labeled python Pandas arrays that can contain any type of data, which is used to specify missing data.

9) Python Matplotlib Package

Matplotlib is arguably most popular graphing and data visualisation library for python.

Following steps were followed.

→ Define x- axis and corresponding y-axis values as lists.

→ Plot them on canvas using plot ( ) function.

→ Give a name to x-axis and y-axis using .xlabel ( ) and y. label( ) function.

→ Give a title to your plot using .title ( ) function.

→ Finally, to view your plot we use .show ( ) function.

```
# importing the required module
importmatplotlib. pyplot as plt
# x axis values
  x = [1,2,3]
 # corresponding y-axis values
   y = [2,4,1]
 # plotting the points
   plt. plot (x,y)
```

CONCLUSION:

Hence we are able to study basic concepts of statistics and display the distribution of samples graphically.

Computational Statistics Laborat
Text To Handwriting Converter -
AIML Assignment1.ipynb - Colab
colab.research.google.com/drive/1P5hECoIdEhlUuKBjrCxBE_qs2UhSuJqN
Apps
HackerRank
W3Schools
Elearn
Codeforces
Django
ERDPlus
LeetCode
Tinkercad
Coursera
My Captain
Google Tech Dev G...
Reading list
AIML Assignment1.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Last edited on January 10
Comment
Share
Table of contents
Section
+ Code
+ Text
Connect
Waiting for colab.research.google.com...
Type here to search
17°C AQI 137
ENG
20:48
12-01-2022

Compute Estimators of the main statistical measures like Mean, Variance, Standard Deviation, Covariance, Correlation and Standard error with respect to any example. Display graphically the distribution of samples.

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    np.random.seed(5)
    x = np.random.randint(10,70,10)
    y = np.random.randint(20,40,10)
    x.sort()
    y.sort()
    print(x)
    print(y)
```

```
[18 19 24 26 45 46 48 49 57 64]
[27 27 32 32 33 35 36 36 36 37]
```

```
[ ]
    def calc_mean(dataset):
        '''
        Def : Mean is defined as the arithmetic average of
        a population.

        Formula : (sum of obesrvations)/(No. of observations)
        '''
        return dataset.sum()/len(dataset)

    def calc_variance(dataset,mean):
```

Computational Statistics Laborat
Text To Handwriting Converter -
AIML Assignment1.ipynb - Colab
colab.research.google.com/drive/1P5hECoIdEhlUuKBjrCxBE_qs2UhSuJqN
Apps
HackerRank
W3Schools
Elearn
Codeforces
Django
ERDPlus
LeetCode
Tinkercad
Coursera
My Captain
Google Tech Dev G...
Reading list
AIML Assignment1.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Last edited on January 10
Comment
Share
Table of contents
Section
+ Code
+ Text
Connect
Editing
[ ]
Waiting for colab.research.google.com...
Type here to search
17°C AQI 137
ENG
20:48
12-01-2022

```
def calc_mean(dataset):
    '''
    Def : Mean is defined as the arithmetic average of
    a population.

    Formula : (sum of obesrvations)/(No. of observations)
    '''
    return dataset.sum()/len(dataset)

def calc_variance(dataset,mean):
    '''
    Def : Variance is the degree of variation/spread
    in the dataset.
    Formula : 1) Σ((X - X_mean)^2) / n
    '''
    squared_diff = np.square(dataset-mean)
    return calc_mean(squared_diff)

def calc_SD(variance):
    '''
  Def : 1) Standard deviation is the amount of deviation
  of points around the mean.
  2) Variation but in terms of the actual dataset.
  Formula : √(variance)
    '''


    return np.sqrt(variance)

def calc_covariance(dataset1,dataset2):
    '''
  Def : Covariance measures the relationship trend
```

Computational Statistics Laborat
Text To Handwriting Converter -
AIML Assignment1.ipynb - Colab
colab.research.google.com/drive/1P5hEColdEhlUuKBjrCxBE_qs2UhSuJqN
Apps
HackerRank
W3Schools
Elearn
Codeforces
Django
ERDPlus
LeetCode
Tinkercad
Coursera
My Captain
Google Tech Dev G...
Reading list
AIML Assignment1.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Last edited on January 10
Comment
Share
Table of contents
Section
+ Code
+ Text
Connect
Editing
Waiting for colab.research.google.com...
Type here to search
17°C AQI 137
ENG
20:48
12-01-2022

```
[ ]     return np.sqrt(variance)

    def calc_covariance(dataset1,dataset2):
        '''
     Def : Covariance measures the relationship trend
     between two sets of data.
     Formula : 1) Σ((X - X_mean)*(Y - Y_mean)) / n
        '''

        mean1 = calc_mean(dataset1)
        mean2 = calc_mean(dataset2)
        return np.sum(np.multiply(dataset1-mean1,dataset2-mean2))/len(dataset1)

    def calc_correlation(dataset1,dataset2):
        '''
     Def : Covariance measures the relationship trend
     between two sets of data.
     Formula : 1) Σ((X - X_mean)*(Y - Y_mean)) / √(Σ(X - X_mean)^2*Σ(Y - Y_mean)^2)
     '''
        mean1 = calc_mean(dataset1)
        mean2 = calc_mean(dataset2)
        num = np.sum(np.multiply(dataset1-mean1,dataset2-mean2))
        de = np.multiply(np.sum(np.square(dataset1-mean1)),np.sum(np.square(dataset2-mean2)))
        return num/np.sqrt(de)

    def calc_SE(dataset,sd):
        '''
        Def : The standard error is a statistical term that
        easures the accuracy with which a sample
        distribution represents a population by using
        standard deviation.
```

Computational Statistics Laborat
Text To Handwriting Converter -
AIML Assignment1.ipynb - Colab
colab.research.google.com/drive/1P5hECoIdEhlUuKBjrCxBE_qs2UhSuJqN
Apps
HackerRank
W3Schools
Elearn
Codeforces
Django
ERDPlus
LeetCode
Tinkercad
Coursera
My Captain
Google Tech Dev G...
Reading list
AIML Assignment1.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Last edited on January 10
Comment
Share
Table of contents
Section
+ Code
+ Text
Connect
Editing
Waiting for colab.research.google.com...
Type here to search
17°C Cloudy
ENG
20:48
12-01-2022

```
[ ] def calc_SE(dataset,sd):
        '''
        Def : The standard error is a statistical term that
        easures the accuracy with which a sample
        distribution represents a population by using
        standard deviation.
        Formula : Standard_deviation / √(n)
        '''
        return sd/np.sqrt(len(dataset))
```

```
[ ] mean = calc_mean(x)
    mean2 = calc_mean(y)
    variance = calc_variance(x,mean)
    S_D = calc_SD(variance)
    covariance = calc_covariance(x,y)
    correlation = calc_correlation(x,y)
    S_E = calc_SE(x,S_D)

    print(mean,mean2,variance,S_D,covariance,correlation,S_E)
```

```
39.6 33.1 244.64000000000001 15.640971836813723 49.84 0.9164339069491503 4.946109582287882
```

Dataset

```
[ ] plt.plot(x,"mo:",label="dataset 1")
    plt.plot(y,"go:",label = "dataset 2")
    plt.legend(loc="upper left")
```

```
<matplotlib.legend.Legend at 0x7f920d74f610>
```

Computational Statistics Laborat
Text To Handwriting Converter -
AIML Assignment1.ipynb - Colab
colab.research.google.com/drive/1P5hECoIdEhlUuKBjrCxBE_qs2UhSuJqN
Apps
HackerRank
W3Schools
Elearn
Codeforces
Django
ERDPlus
LeetCode
Tinkercad
Coursera
My Captain
Google Tech Dev G...
Reading list
AIML Assignment1.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Last edited on January 10
Comment
Share
Table of contents
Section
+ Code
+ Text
Connect
Editing
Waiting for colab.research.google.com...
Type here to search
17°C Cloudy
ENG
20:48
12-01-2022

Dataset

```
[ ] plt.plot(x,"mo:",label="dataset 1")
    plt.plot(y,"go:",label = "dataset 2")
    plt.legend(loc="upper left")
```

```
<matplotlib.legend.Legend at 0x7f920d74f610>
```

dataset 1
dataset 2
60
50
40
30
20
0
2
4
6
8

Mean

```
[ ] plt.plot(x,"mo:",label="Dataset 1")
    plt.axhline(mean,color='b',marker= 'o', linestyle=':',label="Mean")
    plt.legend(loc="upper left")
```

```
<matplotlib.legend.Legend at 0x7f920d249cd0>
```

Computational Statistics Laborat
Text To Handwriting Converter -
AIML Assignment1.ipynb - Colab
colab.research.google.com/drive/1P5hEColdEhlUuKBjrCxBE_qs2UhSuJqN
Apps
HackerRank
W3Schools
Elearn
Codeforces
Django
ERDPlus
LeetCode
Tinkercad
Coursera
My Captain
Google Tech Dev G...
Reading list
AIML Assignment1.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Last edited on January 10
Comment
Share
Table of contents
Section
+ Code
+ Text
Connect
Editing
Dataset 1
Mean
60
50
40
30
20
0
2
4
6
8
Dataset 2
Mean
36
34
32
30
Waiting for colab.research.google.com...
Type here to search
17°C Cloudy
ENG
20:48
12-01-2022

```
<matplotlib.legend.Legend at 0x7f920d249cd0>
```

```
[ ] plt.plot(y,"ro:",label="Dataset 2")
    plt.axhline(mean2,color='g',marker= 'o', linestyle=':',label="Mean")
    plt.legend(loc="upper left")
```

```
<matplotlib.legend.Legend at 0x7f92051dee90>
```

Computational Statistics Laborat
Text To Handwriting Converter -
AIML Assignment1.ipynb - Colab
colab.research.google.com/drive/1P5hECoIdEhlUuKBjrCxBE_qs2UhSuJqN
Apps
HackerRank
W3Schools
Elearn
Codeforces
Django
ERDPlus
LeetCode
Tinkercad
Coursera
My Captain
Google Tech Dev G...
Reading list
AIML Assignment1.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Last [redacted] on [redacted]uary 10
Comment
Share
Table of contents
Section
+ Code
+ Text
Connect
Editing
14000
12000
10000
8000
6000
0
2
4
6
8
Waiting for colab.research.google.com...
Type here to search
17°C Cloudy
ENG
20:48
12-01-2022

Correlation

```
[ ] corr = np.correlate(x, y, "same")
    plt.plot(list(corr),"go:",label = "Correlation")
```

```
[<matplotlib.lines.Line2D at 0x7f9205158550>]
```

Variance

```
[ ] plt.plot(x,"mo:",label="Dataset 1")
    plt.axhline(mean,color='g',marker= 'o', linestyle=':',label="Mean")
    plt.axhline(variance,color='b',marker= 'o', linestyle=':',label="Variance")
    plt.legend(loc="upper left")
```

```
<matplotlib.legend.Legend at 0x7f92050bfe50>
```

Computational Statistics Laborat
Text To Handwriting Converter -
AIML Assignment1.ipynb - Colab
colab.research.google.com/drive/1P5hECoIdEhlUuKBjrCxBE_qs2UhSuJqN
Apps
HackerRank
W3Schools
Elearn
Codeforces
Django
ERDPlus
LeetCode
Tinkercad
Coursera
My Captain
Google Tech Dev G...
Reading list
AIML Assignment1.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Last edited on January 10
Comment
Share
Table of contents
Section
+ Code
+ Text
Connect
Editing
[ ]
100
50
0
2
4
6
8

Covariance#

```
plt.plot(y,"ro:",label="Dataset 2")
plt.plot(x,"go:",label="Daraset 1")
plt.axhline(covariance,color='b',marker= 'o', linestyle=':',label="covariance")
```

```
<matplotlib.lines.Line2D at 0x7f920502dd10>
```

60
50
40
30
20
0
2
4
6
8
Waiting for colab.research.google.com...
Type here to search
17°C Cloudy
ENG
20:48
12-01-2022