CENTRAL UNIVERSITY OF SOUTH BIHAR

DATA SCIENCE AND APPLIED STATISTICS

(SADS ASSIGNMENT)

SEMESTER - II

[SESSION: 2023-2025]

SUBMITTED TO

Dr. KAMLESH KUMAR

ASSISTANT PROFESSOR

SUBMITTED BY

SURESH Kr. PRAJAPATI

Enroll. no. :CUSB2302222008

DEPARTMENT OF STATISTICS

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE CENTRAL UNIVERSITY OF SOUTH BIHAR

SADS-II problem-2 **Date: 31-01-2024**

The data of weight, height and chest circumference of 15 UG students selected from the population of 50 UG students are given below.

Weight(kg)	Height(cm)	Chest circumference(cm)
60	155	36
61	158	45
54	155	39
49	146	34
63	165	46
69	170	49
67	170	46
50	150	39
57	153	45
55	151	40
59	155	41
67	171	47
63	161	41
49	150	39
50	152	40

Find the sample mean vector $\bar{X}$, sample variance covariance matrix S and sample correlation matrix r of the above data.

Result:-

The sample mean vector:-

```
  # to make the matrix
> x=c(60,61,54,49,63,69,67,50,57,55,59,67,63,49,50)
> y=c(155,158,155,146,165,170,170,150,153,151,155,171,161,150,152)
> z=c(36,45,39,34,46,49,46,39,45,40,41,47,41,39,40)
> # TO make the column matrix
> Mtr=cbind(x,y,z)
> Mtr
         x   y  z
 [1,] 60 155 36
 [2,] 61 158 45
 [3,] 54 155 39
 [4,] 49 146 34
 [5,] 63 165 46
 [6,] 69 170 49
 [7,] 67 170 46
 [8,] 50 150 39
 [9,] 57 153 45
[10,] 55 151 40
[11,] 59 155 41
[12,] 67 171 47
[13,] 63 161 41
[14,] 49 150 39
[15,] 50 152 40
> # For finding the sample mean
> sample_mean=colMeans(Mtr)
```

```
> sample_mean
        x        y        z
 58.2000 157.4667  41.8000
> # For finding the covarience matrix
> covarience=cov(Mtr)
> covarience
           x        y        z
x 47.31429 52.11429 23.11429
y 52.11429 65.69524 28.95714
z 23.11429 28.95714 18.60000
> # To finding the correlation matrix
> correlation=cor(Mtr)
> correlation
             x         y         z
x 1.0000000 0.9347462 0.7791622
y 0.9347462 1.0000000 0.8283851
z 0.7791622 0.8283851 1.0000000
> # For sample mean
> sample_mean=colMeans(Mtr)
> sample_mean
        x        y        z
 58.2000 157.4667  41.8000
> # For variance covariance matrix
> covarience=cov(Mtr)
> covarience
           x        y        z
x 47.31429 52.11429 23.11429
y 52.11429 65.69524 28.95714
z 23.11429 28.95714 18.60000
> # for finding the correlation
> correlation=cor(Mtr)
> correlation
             x         y         z
x 1.0000000 0.9347462 0.7791622
y 0.9347462 1.0000000 0.8283851
z 0.7791622 0.8283851 1.0000000
  #for finding the diagonal and finding diagonal matrix
> Mat=diag(covarience)
> Mat
        x        y        z
47.31429 65.69524 18.60000
> v=diag(Mat,nrow=3,ncol=3,names=true)
> v
           [,1]     [,2] [,3]
[1,] 47.31429  0.00000  0.0
[2,]  0.00000 65.69524  0.0
[3,]  0.00000  0.00000 18.6
> # For finding the square root and inverse of matrix
> squa_root=sqrt(v)
> squa_root
           [,1]    [,2]     [,3]
[1,] 6.878538 0.00000 0.000000
[2,] 0.000000 8.10526 0.000000
[3,] 0.000000 0.00000 4.312772
> Inverse_matrix=solve(squa_root)
> Inverse_matrix
             [,1]      [,2]      [,3]
[1,] 0.1453797 0.0000000 0.0000000
[2,] 0.0000000 0.1233767 0.0000000
[3,] 0.0000000 0.0000000 0.2318694



> # To cheack for the correlation
> Rn=Inverse_matrix %*%covarience%*%Inverse_matrix
> Rn
```

```
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.9347462 0.7791622
[2,] 0.9347462 1.0000000 0.8283851
[3,] 0.7791622 0.8283851 1.0000000
>
```

SADS-4 Problem-3 **21-02-2024**

SURESH KUMAR PRAJAPATI

Generate the data for vector $X' = [X_1, X_2, X_3]$ using multivariate normal distribution with mean vector $\mu = [160, 60, 30]$ and covariance matrix $\Sigma = \begin{bmatrix} 4 & 3 & 2 \\ 3 & 6 & 5 \\ 2 & 5 & 9 \end{bmatrix}$. Also calculate the sample mean vector, covariance matrix and correlation matrix of vector *X*. Here X_1, X_2, X_3 are taken as height, weight and chest circumference of students.

SOLUTION:-

```
> #the mean vector given matrix
> Mean_vec=matrix(c(160,60,30),nrow=3,ncol=1)
> Mean_vec
     [,1]
[1,]  160
[2,]   60
[3,]   30
> #the covariance matrix is given
> Co_var_mat=matrix(c(4,3,2,3,6,5,2,5,9),nrow=3,ncol=3)
> Co_var_mat
     [,1] [,2] [,3]
[1,]    4    3    2
[2,]    3    6    5
[3,]    2    5    9
> #the diagonal matrix
> dia=diag(Co_var_mat)
> dia
[1] 4 6 9
> #the random sample to genrate the using mass function
> library(MASS)
> ran_genrate=mvrnorm(10,Mean_vec,Co_var_mat)
> ran_genrate
          [,1]     [,2]     [,3]
 [1,] 158.7836 58.74057 28.47533
 [2,] 162.6535 61.95354 29.79249
 [3,] 157.9313 57.70417 30.02561
 [4,] 161.7552 60.11172 29.51152
 [5,] 158.5643 61.58477 33.45692
 [6,] 160.0259 60.76642 33.90260
 [7,] 158.2184 59.25295 27.07845
 [8,] 157.6580 60.62638 27.28491
 [9,] 159.7754 64.59677 32.35139
[10,] 158.9140 61.13182 30.12351
> # to find the sample mean vector
```

```
> sam_mean_vec=matrix(c(mean(ran_genrate[,1]),
+ mean(ran_genrate[,2]),mean(ran_genrate[,3])))
> sam_mean_vec
          [,1]
[1,] 159.42795
[2,]  60.64691
[3,]  30.20027
> # with the help of random genrate for finding covariance and correlation
> Co_var_mat1=cov(ran_genrate)
> Co_var_mat1
         [,1]     [,2]     [,3]
[1,] 2.728836 1.189848 0.842079
[2,] 1.189848 3.662393 2.222868
[3,] 0.842079 2.222868 5.648654
> co_rr=cor(ran_genrate)
> co_rr
          [,1]      [,2]
[1,] 1.0000000 0.3763747
[2,] 0.3763747 1.0000000
[3,] 0.2144824 0.4887185
          [,3]
[1,] 0.2144824
[2,] 0.4887185
[3,] 1.0000000
>
```

CONCLUSION:-We have generate random data of sample size 10 using mvrnorm function and calculated sample mean vector, sample cov matrix, sample cor matrix from it. The values of sample mean vector and cov matrix are obtain nearest to the values of population mean vector and population cov matrix.

SADS PROBLEM-1 **24**-01-2024

SURESH KUMAR PRAJAPATI

The matrices A and B are given below:

$$A=\begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix} \text{ and } B=\begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Find the eigen values and the associated normalized eigen vectors of the above matrices using R software.

SOLUTION:-

Eigen value of given matrix is,

```
> #The eigen value of a matrix and eigen vector
> x=c(1,-5,-5,1)
> x
[1]  1 -5 -5  1
> y=matrix(x,nrow = 2,ncol = 2)
> y
      [,1] [,2]
[1,]    1   -5
[2,]   -5    1
> z=eigen(y)
> z
eigen() decomposition
$values
[1]  6 -4

$vectors
              [,1]       [,2]
[1,] -0.7071068 -0.7071068
[2,]  0.7071068 -0.7071068

> extr=z$vectors
> extr
              [,1]       [,2]
[1,] -0.7071068 -0.7071068
[2,]  0.7071068 -0.7071068
> # To extract the data column and also cheack the normality.
> extr=z$vectors
> extr
              [,1]       [,2]
[1,] -0.7071068 -0.7071068
[2,]  0.7071068 -0.7071068
> col1=extr[,1]
> col1
[1] -0.7071068  0.7071068
> sq=col1*col1
> sq
[1] 0.5 0.5
> sm=sum(sq)
> sm
[1] 1
> col2=extr[,2]
> col2
```

```
[1] -0.7071068 -0.7071068
> sq1=col2*col2
> sq1
[1] 0.5 0.5
> sm1=sum(sq1)
> sm1
[1] 1
```

[2]:→ For matrix B solve the eigen value and eigen vector and cheack the normality.

```
 > # second matrix eigen value and normal eigen vector

> T=c(1,-2,0,-2,5,0,0,0,2)
> T
[1]  1 -2  0 -2  5  0  0  0  2
> U=matrix(T,nrow=3,ncol=3)
> U
      [,1] [,2] [,3]
[1,]    1   -2    0
[2,]   -2    5    0
[3,]    0    0    2
> V=eigen(U)
> V
eigen() decomposition
$values
[1] 5.8284271 2.0000000 0.1715729

$vectors
              [,1] [,2]      [,3]
[1,] -0.3826834    0 0.9238795
[2,]  0.9238795    0 0.3826834
[3,]  0.0000000    1 0.0000000

> W=V$vectors
> W
              [,1] [,2]      [,3]
[1,] -0.3826834    0 0.9238795
[2,]  0.9238795    0 0.3826834
[3,]  0.0000000    1 0.0000000
 > # To extract the data in matrix by column and cheack normality

> X=W[,1]
> X
[1] -0.3826834  0.9238795  0.0000000
> SQ=X*X
> SQ
[1] 0.1464466 0.8535534 0.0000000
> SM=sum(SQ)
> SM
[1] 1
> X1=W[,2]
> X1
[1] 0 0 1
> SQ1=X1*X1
> SQ1
[1] 0 0 1
> SM1=sum(SQ1)
> SM1
[1] 1
> X2=W[,3]
> X2
[1] 0.9238795 0.3826834 0.0000000
> SQ2=X2*X2
> SQ2
[1] 0.8535534 0.1464466 0.0000000
```

```
> SM3=sum(SQ1)
> SM3
[1] 1
```

Conclusion:-Above these two matrices have eigen value and associated eigen vector are normalized.

SADS Problem 5

Suresh kumar prajapati

2024-02-23

OPERATION

The data of weight (kg), height (cm) and chest circumference (cm) of 15 UG students selected from the multivariate normal distribution with mean vector sigma and variance- covariance matrix data, are given below.

```
sigma=matrix(c(44,49,21,49,61,27,21,27,17),nrow=3,ncol=3)
sigma
```

```
##      [,1] [,2] [,3]
## [1,]   44   49   21
## [2,]   49   61   27
## [3,]   21   27   17
```

Make the matrix in data frame

```
WeightX1=c(59,63,57,47,66,68,70,52,56,53,56,64,62,48,51)
HeightX2=c(152,157,156,147,166,168,171,151,154,152,156,169,163,151,153)
ChestX3=c(35,46,40,33,47,42,45,38,44,41,42,48,42,39,41)
data=data.frame(WeightX1,HeightX2,ChestX3)
data
```

WeightX1 <dbl>	HeightX2 <dbl>	ChestX3 <dbl>
59	152	35
63	157	46
57	156	40
47	147	33
66	166	47
68	168	42
70	171	45
52	151	38
56	154	44
53	152	41

1-10 of 15 rows Previous **1** 2 Next

```
x=colMeans(data)
x
```

```
##  WeightX1  HeightX2   ChestX3
##  58.13333 157.73333  41.53333
```

Inverse of sigma

```
inv=solve(sigma)
 inv
```

```
##       [,1]       [,2]       [,3]
## [1,]  0.22 -0.1900000  0.0300000
## [2,] -0.19  0.2192857 -0.1135714
## [3,]  0.03 -0.1135714  0.2021429
```

to find the Xbar -mu

```
mu=c(58,157,42)
mu
```

```
## [1]  58 157  42
```

```
X= x-mu
 X
```

```
##   WeightX1   HeightX2    ChestX3
##  0.1333333  0.7333333 -0.4666667
```

chi square test

```
res=15*(t(X)%*%inv%*%X)
res
```

```
##          [,1]
## [1,] 3.040571
```

Conclusion:- At the given level of significane chi square is less the tabulated value of chi square test.i.e.The data of weight,height,chest circumference of 15 ug with sample is not varies,i.e. accepted

SADS Problem-6

Suresh kumar prajapati

2024-03-06

Perspiration from a sample of 20 healthy females was analyzed. Three components, X_1 = Sweet rate, X_2= Sodium content and X_3= Potassium content, were measured and the results, which we call the sweat data are given in the below table.

Sweat Data			
Individual	X_1 (Sweat rate)	X_2 (Sodium)	X_3 (Potassium)
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14.0
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Population mean vector for Sweat rate, Sodium content and Potassium content is given as $\mu_0^{'}=[4,50,10]$. Test the hypothesis $H_0:\mu^{'}=\mu_0^{'}$ against $H_1:\mu^{'}\neq\mu_0^{'}$ at given level of significance $\alpha=0.05$ $[F_{(3,17)}(0.05)=3.197]$.

H_0:μ'=μ_0 Against H_1: μ'≠μ_0

##Import the data from the excel problem_6 setwd(C:\Users\Admin\OneDrive\Desktop\santosh)

```
setwd("C:\\Users\\Admin\\OneDrive\\Desktop\\santosh")
getwd()
```

```
## [1] "C:/Users/Admin/OneDrive/Desktop/santosh"
```

```
data=read.csv("problem_6.csv")
data
```

```
##    X1..Sweat.rate. X2..Sodium. X3..Potassium.
## 1              3.7        48.5            9.3
## 2              5.7        65.1            8.0
## 3              3.8        47.2           10.9
## 4              3.2        53.2           12.0
## 5              3.1        55.5            9.7
## 6              4.6        36.1            7.9
## 7              2.4        24.8           14.0
## 8              7.2        33.1            7.6
## 9              6.7        47.4            8.5
## 10             5.4        54.1           11.3
## 11             3.9        36.9           12.7
## 12             4.5        58.8           12.3
## 13             3.5        27.8            9.8
## 14             4.5        40.2            8.4
## 15             1.5        13.5           10.1
## 16             8.5        56.4            7.1
## 17             4.5        71.6            8.2
## 18             6.5        52.8           10.9
## 19             4.1        44.1           11.2
## 20             5.5        40.9            9.4
```

##given that the population mean and covariance matrix inverse of covariance matrix

```
Mu=c(4,50,10)
Mu
```

```
## [1]  4 50 10
```

```
x=colMeans(data)
x
```

```
## X1..Sweat.rate.     X2..Sodium.  X3..Potassium.
##           4.640          45.400           9.965
```

```
s=cov(data)
s
```

```
##                 X1..Sweat.rate. X2..Sodium. X3..Potassium.
## X1..Sweat.rate.        2.879368     10.0100      -1.809053
## X2..Sodium.           10.010000    199.7884      -5.640000
## X3..Potassium.        -1.809053     -5.6400       3.627658
```

```
inv=solve(s)
inv
```

```
##                 X1..Sweat.rate.  X2..Sodium. X3..Potassium.
## X1..Sweat.rate.      0.58615531 -0.022085719    0.257968742
## X2..Sodium.         -0.02208572  0.006067227   -0.001580929
## X3..Potassium.       0.25796874 -0.001580929    0.401846765
```

To finding the transpose matrix

```
y=t(x-Mu)
y
```

```
##      X1..Sweat.rate. X2..Sodium. X3..Potassium.
## [1,]            0.64        -4.6         -0.035
```

```
z=(x-Mu)
z
```

```
## X1..Sweat.rate.     X2..Sodium.  X3..Potassium.
##           0.640          -4.600          -0.035
```

To find the T square matrix

The sample mean x^- to the test value μ_0

$T^2=n(x^-—\mu_0)^2/s^2=n(x—\mu_0)'(s^2)^{-1}(x^-\mu_0)$

for natural generalizationof the square distance in multivariate analog

$T^2=n(x—\mu_0)'(s/n)^{-1}(x^-\mu_0)=n(x—\mu_0)'(s)^{-1}(x^-\mu_0)$

```
T2=20*y%*%inv%*%z
T2
```

```
##          [,1]
## [1,] 9.738773
```

Comparision of data with f distribution

```
tcal=T2*(20-3)/((20-1)*3)
tcal
```

```
##          [,1]
## [1,] 2.904546
```

Result:-tcal<Ftab there is evidence to fail the null hypothesis. i.e. reject the null hypothesis

SADS-Problem-7 **13-03-2024**

In the first phase of a study of the cost of transporting milk from farms to dairy plants, a survey was taken of firms engaged in milk transportations. Cost data on X_1 = fuel, X_2 = repair and X_3=capital, all measured on a per mile basis, are presented in below table for n_1=20 gasoline and n_2=15 diesel trucks.

Table-1: Milk Transportation Cost Data

Gasoline Trucks			Diesel Trucks		
X_1	X_2	X_3	X_1	X_2	X_3
16	12	11	8	12	9
7	3	4	7	5	17
10	2	10	10	3	11
4	6	8	10	15	6
11	5	11	13	4	29
14	6	10	10	13	11
13	11	11	6	9	19
13	14	9	11	10	14
29	15	3	9	3	14
13	8	10	10	5	21
7	6	8	11	18	35
10	4	9	12	12	17
10	5	10	9	13	21
11	6	8	8	10	17
12	14	14	8	6	16
10	3	6			
10	6	12			
9	3	12			
12	8	11			
8	14	12			

Test whether mean cost vectors for Gasoline Trucks and Diesel Trucks are same or not at $\alpha = 0.01$ level of significance [$F_{3,31}(0.01) = 4.48$].

Solution:

```
rm(list=ls())
x11=c(16,7,10,4,11,14,13,13,29,13,7,10,10,11,12,10,10,9,12,8)
x21=c(12,3,2,6,5,6,11,14,15,8,6,4,5,6,14,3,6,3,8,14)
x31=c(11,4,10,8,11,10,11,9,3,10,8,9,10,8,14,6,12,12,11,12)
x=c(x11,x21,x31)
X1=matrix(x,ncol=3,byrow=F)
X1
      [,1] [,2] [,3]
 [1,]   16   12   11
 [2,]    7    3    4
 [3,]   10    2   10
 [4,]    4    6    8
 [5,]   11    5   11
 [6,]   14    6   10
```

```
 [7,]   13   11   11
 [8,]   13   14    9
 [9,]   29   15    3
[10,]   13    8   10
[11,]    7    6    8
[12,]   10    4    9
[13,]   10    5   10
[14,]   11    6    8
[15,]   12   14   14
[16,]   10    3    6
[17,]   10    6   12
[18,]    9    3   12
[19,]   12    8   11
[20,]    8   14   12
```

```
n1=20
x12=c(8,7,10,10,13,10,6,11,9,10,11,12,9,8,8)
x22=c(12,5,3,15,4,13,9,10,3,5,18,12,13,10,6)
x32=c(9,17,11,6,29,11,19,14,14,21,35,17,21,17,16)
y=c(x12,x22,x32)
X2=matrix(y,ncol=3,byrow=F)
X2
```

```
      [,1] [,2] [,3]
 [1,]    8   12    9
 [2,]    7    5   17
 [3,]   10    3   11
 [4,]   10   15    6
 [5,]   13    4   29
 [6,]   10   13   11
 [7,]    6    9   19
 [8,]   11   10   14
 [9,]    9    3   14
[10,]   10    5   21
[11,]   11   18   35
[12,]   12   12   17
[13,]    9   13   21
[14,]    8   10   17
[15,]    8    6   16
```

```
n2=15

M1=colMeans(X1)
M1

[1] 11.45  7.55  9.45

M2=colMeans(X2)
M2

[1]  9.466667  9.200000 17.133333

#mean difference
M=M1-M2
M

[1]  1.983333 -1.650000 -7.683333

X1bar=matrix(M1,nrow=20,ncol=3,byrow=T)
X1bar

       [,1] [,2] [,3]
 [1,] 11.45 7.55 9.45
 [2,] 11.45 7.55 9.45
```

```
 [3,] 11.45 7.55 9.45
 [4,] 11.45 7.55 9.45
 [5,] 11.45 7.55 9.45
 [6,] 11.45 7.55 9.45
 [7,] 11.45 7.55 9.45
 [8,] 11.45 7.55 9.45
 [9,] 11.45 7.55 9.45
[10,] 11.45 7.55 9.45
[11,] 11.45 7.55 9.45
[12,] 11.45 7.55 9.45
[13,] 11.45 7.55 9.45
[14,] 11.45 7.55 9.45
[15,] 11.45 7.55 9.45
[16,] 11.45 7.55 9.45
[17,] 11.45 7.55 9.45
[18,] 11.45 7.55 9.45
[19,] 11.45 7.55 9.45
[20,] 11.45 7.55 9.45
```

```
X2bar=matrix(M2,nrow=15,ncol=3,byrow=T)
X2bar
```

```
          [,1] [,2]     [,3]
 [1,] 9.466667  9.2 17.13333
 [2,] 9.466667  9.2 17.13333
 [3,] 9.466667  9.2 17.13333
 [4,] 9.466667  9.2 17.13333
 [5,] 9.466667  9.2 17.13333
 [6,] 9.466667  9.2 17.13333
 [7,] 9.466667  9.2 17.13333
 [8,] 9.466667  9.2 17.13333
 [9,] 9.466667  9.2 17.13333
[10,] 9.466667  9.2 17.13333
[11,] 9.466667  9.2 17.13333
[12,] 9.466667  9.2 17.13333
[13,] 9.466667  9.2 17.13333
[14,] 9.466667  9.2 17.13333
[15,] 9.466667  9.2 17.13333
```

```
a=(X1-X1bar)
a
```

```
       [,1]  [,2]  [,3]
 [1,]  4.55  4.45  1.55
 [2,] -4.45 -4.55 -5.45
 [3,] -1.45 -5.55  0.55
 [4,] -7.45 -1.55 -1.45
 [5,] -0.45 -2.55  1.55
 [6,]  2.55 -1.55  0.55
 [7,]  1.55  3.45  1.55
 [8,]  1.55  6.45 -0.45
 [9,] 17.55  7.45 -6.45
[10,]  1.55  0.45  0.55
[11,] -4.45 -1.55 -1.45
[12,] -1.45 -3.55 -0.45
[13,] -1.45 -2.55  0.55
[14,] -0.45 -1.55 -1.45
[15,]  0.55  6.45  4.55
[16,] -1.45 -4.55 -3.45
[17,] -1.45 -1.55  2.55
[18,] -2.45 -4.55  2.55
[19,]  0.55  0.45  1.55
[20,] -3.45  6.45  2.55
```

```
k=t(a)%*%a
k
```

```
       [,1]   [,2]   [,3]
[1,] 466.95 222.05 -72.05
[2,] 222.05 342.95  32.05
[3,] -72.05  32.05 140.95
```

```
b=X2-X2bar
b
```

```
             [,1] [,2]        [,3]
 [1,] -1.4666667  2.8  -8.1333333
 [2,] -2.4666667 -4.2  -0.1333333
 [3,]  0.5333333 -6.2  -6.1333333
 [4,]  0.5333333  5.8 -11.1333333
 [5,]  3.5333333 -5.2  11.8666667
 [6,]  0.5333333  3.8  -6.1333333
 [7,] -3.4666667 -0.2   1.8666667
 [8,]  1.5333333  0.8  -3.1333333
 [9,] -0.4666667 -6.2  -3.1333333
[10,]  0.5333333 -4.2   3.8666667
[11,]  1.5333333  8.8  17.8666667
[12,]  2.5333333  2.8  -0.1333333
[13,] -0.4666667  3.8   3.8666667
[14,] -1.4666667  0.8  -0.1333333
[15,] -1.4666667 -3.2  -1.1333333
```

```
l=t(b)%*%b
l
```

```
         [,1]  [,2]      [,3]
[1,] 49.73333  14.6  61.06667
[2,] 14.60000 306.4  41.60000
[3,] 61.06667  41.6 779.73333
```

```
#sample covariance matrix
s=(k+l)/(n1+n2-2)
s
```

```
           [,1]      [,2]       [,3]
[1,] 15.6570707  7.171212 -0.3328283
[2,]  7.1712121 19.677273  2.2318182
[3,] -0.3328283  2.231818 27.8994949
```

```
#calculate T^2
T={(n1*n2)/(n1+n2)}*t(M)%*%solve(s)%*%M
T
```

```
         [,1]
[1,] 22.01801
```

```
#calculate F
F={T/(n1+n2-2)}*{(n1+n2-2-3+1)/3}
F
```

```
         [,1]
[1,] 6.894529
```

Result:- fcal > Ftab then reject the null hypothesis there is difference between gasolin trucks and diesel trucks.

CENTRAL UNIVERSITY OF SOUTH BIHAR

DATA SCIENCE AND APPLIED STATISTICS

(SADS ASSIGNMENT)

SEMESTER - II

[SESSION: 2023-2025]

SUBMITTED TO

Dr. KAMLESH KUMAR

ASSISTANT PROFESSOR

SUBMITTED BY

SURESH Kr. PRAJAPATI

Enroll. no. :CUSB2302222008

DEPARTMENT OF STATISTICS

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE CENTRAL UNIVERSITY OF SOUTH BIHAR

SADS-II problem-2 **Date: 31-01-2024**

The data of weight, height and chest circumference of 15 UG students selected from the population of 50 UG students are given below.

Weight(kg)	Height(cm)	Chest circumference(cm)
60	155	36
61	158	45
54	155	39
49	146	34
63	165	46
69	170	49
67	170	46
50	150	39
57	153	45
55	151	40
59	155	41
67	171	47
63	161	41
49	150	39
50	152	40

Find the sample mean vector $\bar{X}$, sample variance covariance matrix S and sample correlation matrix r of the above data.

Result:-

The sample mean vector:-

```
  # to make the matrix
> x=c(60,61,54,49,63,69,67,50,57,55,59,67,63,49,50)
> y=c(155,158,155,146,165,170,170,150,153,151,155,171,161,150,152)
> z=c(36,45,39,34,46,49,46,39,45,40,41,47,41,39,40)
> # TO make the column matrix
> Mtr=cbind(x,y,z)
> Mtr
         x   y  z
 [1,] 60 155 36
 [2,] 61 158 45
 [3,] 54 155 39
 [4,] 49 146 34
 [5,] 63 165 46
 [6,] 69 170 49
 [7,] 67 170 46
 [8,] 50 150 39
 [9,] 57 153 45
[10,] 55 151 40
[11,] 59 155 41
[12,] 67 171 47
[13,] 63 161 41
[14,] 49 150 39
[15,] 50 152 40
> # For finding the sample mean
> sample_mean=colMeans(Mtr)
```

```
> sample_mean
        x        y        z
 58.2000 157.4667  41.8000
> # For finding the covarience matrix
> covarience=cov(Mtr)
> covarience
           x        y        z
x 47.31429 52.11429 23.11429
y 52.11429 65.69524 28.95714
z 23.11429 28.95714 18.60000
> # To finding the correlation matrix
> correlation=cor(Mtr)
> correlation
             x         y         z
x 1.0000000 0.9347462 0.7791622
y 0.9347462 1.0000000 0.8283851
z 0.7791622 0.8283851 1.0000000
> # For sample mean
> sample_mean=colMeans(Mtr)
> sample_mean
        x        y        z
 58.2000 157.4667  41.8000
> # For variance covariance matrix
> covarience=cov(Mtr)
> covarience
           x        y        z
x 47.31429 52.11429 23.11429
y 52.11429 65.69524 28.95714
z 23.11429 28.95714 18.60000
> # for finding the correlation
> correlation=cor(Mtr)
> correlation
             x         y         z
x 1.0000000 0.9347462 0.7791622
y 0.9347462 1.0000000 0.8283851
z 0.7791622 0.8283851 1.0000000
  #for finding the diagonal and finding diagonal matrix
> Mat=diag(covarience)
> Mat
        x        y        z
47.31429 65.69524 18.60000
> v=diag(Mat,nrow=3,ncol=3,names=true)
> v
           [,1]     [,2] [,3]
[1,] 47.31429  0.00000  0.0
[2,]  0.00000 65.69524  0.0
[3,]  0.00000  0.00000 18.6
> # For finding the square root and inverse of matrix
> squa_root=sqrt(v)
> squa_root
           [,1]    [,2]     [,3]
[1,] 6.878538 0.00000 0.000000
[2,] 0.000000 8.10526 0.000000
[3,] 0.000000 0.00000 4.312772
> Inverse_matrix=solve(squa_root)
> Inverse_matrix
             [,1]      [,2]      [,3]
[1,] 0.1453797 0.0000000 0.0000000
[2,] 0.0000000 0.1233767 0.0000000
[3,] 0.0000000 0.0000000 0.2318694



> # To cheack for the correlation
> Rn=Inverse_matrix %*%covarience%*%Inverse_matrix
> Rn
```

```
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.9347462 0.7791622
[2,] 0.9347462 1.0000000 0.8283851
[3,] 0.7791622 0.8283851 1.0000000
>
```

SADS-4 Problem-3 **21-02-2024**

SURESH KUMAR PRAJAPATI

Generate the data for vector $X' = [X_1, X_2, X_3]$ using multivariate normal distribution with mean vector $\mu = [160, 60, 30]$ and covariance matrix $\Sigma = \begin{bmatrix} 4 & 3 & 2 \\ 3 & 6 & 5 \\ 2 & 5 & 9 \end{bmatrix}$. Also calculate the sample mean vector, covariance matrix and correlation matrix of vector *X*. Here X_1, X_2, X_3 are taken as height, weight and chest circumference of students.

SOLUTION:-

```
> #the mean vector given matrix
> Mean_vec=matrix(c(160,60,30),nrow=3,ncol=1)
> Mean_vec
     [,1]
[1,]  160
[2,]   60
[3,]   30
> #the covariance matrix is given
> Co_var_mat=matrix(c(4,3,2,3,6,5,2,5,9),nrow=3,ncol=3)
> Co_var_mat
     [,1] [,2] [,3]
[1,]    4    3    2
[2,]    3    6    5
[3,]    2    5    9
> #the diagonal matrix
> dia=diag(Co_var_mat)
> dia
[1] 4 6 9
> #the random sample to genrate the using mass function
> library(MASS)
> ran_genrate=mvrnorm(10,Mean_vec,Co_var_mat)
> ran_genrate
          [,1]     [,2]     [,3]
 [1,] 158.7836 58.74057 28.47533
 [2,] 162.6535 61.95354 29.79249
 [3,] 157.9313 57.70417 30.02561
 [4,] 161.7552 60.11172 29.51152
 [5,] 158.5643 61.58477 33.45692
 [6,] 160.0259 60.76642 33.90260
 [7,] 158.2184 59.25295 27.07845
 [8,] 157.6580 60.62638 27.28491
 [9,] 159.7754 64.59677 32.35139
[10,] 158.9140 61.13182 30.12351
> # to find the sample mean vector
```

```
> sam_mean_vec=matrix(c(mean(ran_genrate[,1]),
+ mean(ran_genrate[,2]),mean(ran_genrate[,3])))
> sam_mean_vec
          [,1]
[1,] 159.42795
[2,]  60.64691
[3,]  30.20027
> # with the help of random genrate for finding covariance and correlation
> Co_var_mat1=cov(ran_genrate)
> Co_var_mat1
         [,1]     [,2]     [,3]
[1,] 2.728836 1.189848 0.842079
[2,] 1.189848 3.662393 2.222868
[3,] 0.842079 2.222868 5.648654
> co_rr=cor(ran_genrate)
> co_rr
          [,1]      [,2]
[1,] 1.0000000 0.3763747
[2,] 0.3763747 1.0000000
[3,] 0.2144824 0.4887185
          [,3]
[1,] 0.2144824
[2,] 0.4887185
[3,] 1.0000000
>
```

CONCLUSION:-We have generate random data of sample size 10 using mvrnorm function and calculated sample mean vector, sample cov matrix, sample cor matrix from it. The values of sample mean vector and cov matrix are obtain nearest to the values of population mean vector and population cov matrix.

SADS PROBLEM-1 **24**-01-2024

SURESH KUMAR PRAJAPATI

The matrices A and B are given below:

$$A=\begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix} \text{ and } B=\begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Find the eigen values and the associated normalized eigen vectors of the above matrices using R software.

SOLUTION:-

Eigen value of given matrix is,

```
> #The eigen value of a matrix and eigen vector
> x=c(1,-5,-5,1)
> x
[1]  1 -5 -5  1
> y=matrix(x,nrow = 2,ncol = 2)
> y
      [,1] [,2]
[1,]    1   -5
[2,]   -5    1
> z=eigen(y)
> z
eigen() decomposition
$values
[1]  6 -4

$vectors
              [,1]       [,2]
[1,] -0.7071068 -0.7071068
[2,]  0.7071068 -0.7071068

> extr=z$vectors
> extr
              [,1]       [,2]
[1,] -0.7071068 -0.7071068
[2,]  0.7071068 -0.7071068
> # To extract the data column and also cheack the normality.
> extr=z$vectors
> extr
              [,1]       [,2]
[1,] -0.7071068 -0.7071068
[2,]  0.7071068 -0.7071068
> col1=extr[,1]
> col1
[1] -0.7071068  0.7071068
> sq=col1*col1
> sq
[1] 0.5 0.5
> sm=sum(sq)
> sm
[1] 1
> col2=extr[,2]
> col2
```

```
[1] -0.7071068 -0.7071068
> sq1=col2*col2
> sq1
[1] 0.5 0.5
> sm1=sum(sq1)
> sm1
[1] 1
```

[2]:→ For matrix B solve the eigen value and eigen vector and cheack the normality.

```
 > # second matrix eigen value and normal eigen vector

> T=c(1,-2,0,-2,5,0,0,0,2)
> T
[1]  1 -2  0 -2  5  0  0  0  2
> U=matrix(T,nrow=3,ncol=3)
> U
      [,1] [,2] [,3]
[1,]    1   -2    0
[2,]   -2    5    0
[3,]    0    0    2
> V=eigen(U)
> V
eigen() decomposition
$values
[1] 5.8284271 2.0000000 0.1715729

$vectors
              [,1] [,2]      [,3]
[1,] -0.3826834    0 0.9238795
[2,]  0.9238795    0 0.3826834
[3,]  0.0000000    1 0.0000000

> W=V$vectors
> W
              [,1] [,2]      [,3]
[1,] -0.3826834    0 0.9238795
[2,]  0.9238795    0 0.3826834
[3,]  0.0000000    1 0.0000000
 > # To extract the data in matrix by column and cheack normality

> X=W[,1]
> X
[1] -0.3826834  0.9238795  0.0000000
> SQ=X*X
> SQ
[1] 0.1464466 0.8535534 0.0000000
> SM=sum(SQ)
> SM
[1] 1
> X1=W[,2]
> X1
[1] 0 0 1
> SQ1=X1*X1
> SQ1
[1] 0 0 1
> SM1=sum(SQ1)
> SM1
[1] 1
> X2=W[,3]
> X2
[1] 0.9238795 0.3826834 0.0000000
> SQ2=X2*X2
> SQ2
[1] 0.8535534 0.1464466 0.0000000
```

```
> SM3=sum(SQ1)
> SM3
[1] 1
```

Conclusion:-Above these two matrices have eigen value and associated eigen vector are normalized.

SADS Problem 5

Suresh kumar prajapati

2024-02-23

OPERATION

The data of weight (kg), height (cm) and chest circumference (cm) of 15 UG students selected from the multivariate normal distribution with mean vector sigma and variance- covariance matrix data, are given below.

```
sigma=matrix(c(44,49,21,49,61,27,21,27,17),nrow=3,ncol=3)
sigma
```

```
##      [,1] [,2] [,3]
## [1,]   44   49   21
## [2,]   49   61   27
## [3,]   21   27   17
```

Make the matrix in data frame

```
WeightX1=c(59,63,57,47,66,68,70,52,56,53,56,64,62,48,51)
HeightX2=c(152,157,156,147,166,168,171,151,154,152,156,169,163,151,153)
ChestX3=c(35,46,40,33,47,42,45,38,44,41,42,48,42,39,41)
data=data.frame(WeightX1,HeightX2,ChestX3)
data
```

WeightX1 <dbl>	HeightX2 <dbl>	ChestX3 <dbl>
59	152	35
63	157	46
57	156	40
47	147	33
66	166	47
68	168	42
70	171	45
52	151	38
56	154	44
53	152	41

1-10 of 15 rows Previous **1** 2 Next

```
x=colMeans(data)
x
```

```
##  WeightX1  HeightX2   ChestX3
##  58.13333 157.73333  41.53333
```

Inverse of sigma

```
inv=solve(sigma)
 inv
```

```
##       [,1]       [,2]       [,3]
## [1,]  0.22 -0.1900000  0.0300000
## [2,] -0.19  0.2192857 -0.1135714
## [3,]  0.03 -0.1135714  0.2021429
```

to find the Xbar -mu

```
mu=c(58,157,42)
mu
```

```
## [1]  58 157  42
```

```
X= x-mu
 X
```

```
##   WeightX1   HeightX2    ChestX3
##  0.1333333  0.7333333 -0.4666667
```

chi square test

```
res=15*(t(X)%*%inv%*%X)
res
```

```
##          [,1]
## [1,] 3.040571
```

Conclusion:- At the given level of significane chi square is less the tabulated value of chi square test.i.e.The data of weight,height,chest circumference of 15 ug with sample is not varies,i.e. accepted

SADS Problem-6

Suresh kumar prajapati

2024-03-06

Perspiration from a sample of 20 healthy females was analyzed. Three components, X_1 = Sweet rate, X_2= Sodium content and X_3= Potassium content, were measured and the results, which we call the sweat data are given in the below table.

Sweat Data			
Individual	X_1 (Sweat rate)	X_2 (Sodium)	X_3 (Potassium)
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14.0
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Population mean vector for Sweat rate, Sodium content and Potassium content is given as $\mu_0^{'}=[4,50,10]$. Test the hypothesis $H_0:\mu^{'}=\mu_0^{'}$ against $H_1:\mu^{'}\neq\mu_0^{'}$ at given level of significance $\alpha=0.05$ $[F_{(3,17)}(0.05)=3.197]$.

H_0:μ'=μ_0 Against H_1: μ'≠μ_0

##Import the data from the excel problem_6 setwd(C:\Users\Admin\OneDrive\Desktop\santosh)

```
setwd("C:\\Users\\Admin\\OneDrive\\Desktop\\santosh")
getwd()
```

```
## [1] "C:/Users/Admin/OneDrive/Desktop/santosh"
```

```
data=read.csv("problem_6.csv")
data
```

```
##    X1..Sweat.rate. X2..Sodium. X3..Potassium.
## 1              3.7        48.5            9.3
## 2              5.7        65.1            8.0
## 3              3.8        47.2           10.9
## 4              3.2        53.2           12.0
## 5              3.1        55.5            9.7
## 6              4.6        36.1            7.9
## 7              2.4        24.8           14.0
## 8              7.2        33.1            7.6
## 9              6.7        47.4            8.5
## 10             5.4        54.1           11.3
## 11             3.9        36.9           12.7
## 12             4.5        58.8           12.3
## 13             3.5        27.8            9.8
## 14             4.5        40.2            8.4
## 15             1.5        13.5           10.1
## 16             8.5        56.4            7.1
## 17             4.5        71.6            8.2
## 18             6.5        52.8           10.9
## 19             4.1        44.1           11.2
## 20             5.5        40.9            9.4
```

##given that the population mean and covariance matrix inverse of covariance matrix

```
Mu=c(4,50,10)
Mu
```

```
## [1]  4 50 10
```

```
x=colMeans(data)
x
```

```
## X1..Sweat.rate.     X2..Sodium.  X3..Potassium.
##           4.640          45.400           9.965
```

```
s=cov(data)
s
```

```
##                 X1..Sweat.rate. X2..Sodium. X3..Potassium.
## X1..Sweat.rate.        2.879368     10.0100      -1.809053
## X2..Sodium.           10.010000    199.7884      -5.640000
## X3..Potassium.        -1.809053     -5.6400       3.627658
```

```
inv=solve(s)
inv
```

```
##                 X1..Sweat.rate.  X2..Sodium. X3..Potassium.
## X1..Sweat.rate.      0.58615531 -0.022085719    0.257968742
## X2..Sodium.         -0.02208572  0.006067227   -0.001580929
## X3..Potassium.       0.25796874 -0.001580929    0.401846765
```

To finding the transpose matrix

```
y=t(x-Mu)
y
```

```
##      X1..Sweat.rate. X2..Sodium. X3..Potassium.
## [1,]            0.64        -4.6         -0.035
```

```
z=(x-Mu)
z
```

```
## X1..Sweat.rate.     X2..Sodium.  X3..Potassium.
##           0.640          -4.600          -0.035
```

To find the T square matrix

The sample mean x^- to the test value μ_0

$T^2=n(x^-—\mu_0)^2/s^2=n(x—\mu_0)'(s^2)^{-1}(x^-\mu_0)$

for natural generalizationof the square distance in multivariate analog

$T^2=n(x—\mu_0)'(s/n)^{-1}(x^-\mu_0)=n(x—\mu_0)'(s)^{-1}(x^-\mu_0)$

```
T2=20*y%*%inv%*%z
T2
```

```
##          [,1]
## [1,] 9.738773
```

Comparision of data with f distribution

```
tcal=T2*(20-3)/((20-1)*3)
tcal
```

```
##          [,1]
## [1,] 2.904546
```

Result:-tcal<Ftab there is evidence to fail the null hypothesis. i.e. reject the null hypothesis

SADS-Problem-7 **13-03-2024**

In the first phase of a study of the cost of transporting milk from farms to dairy plants, a survey was taken of firms engaged in milk transportations. Cost data on X_1 = fuel, X_2 = repair and X_3=capital, all measured on a per mile basis, are presented in below table for n_1=20 gasoline and n_2=15 diesel trucks.

Table-1: Milk Transportation Cost Data

Gasoline Trucks			Diesel Trucks		
X_1	X_2	X_3	X_1	X_2	X_3
16	12	11	8	12	9
7	3	4	7	5	17
10	2	10	10	3	11
4	6	8	10	15	6
11	5	11	13	4	29
14	6	10	10	13	11
13	11	11	6	9	19
13	14	9	11	10	14
29	15	3	9	3	14
13	8	10	10	5	21
7	6	8	11	18	35
10	4	9	12	12	17
10	5	10	9	13	21
11	6	8	8	10	17
12	14	14	8	6	16
10	3	6			
10	6	12			
9	3	12			
12	8	11			
8	14	12			

Test whether mean cost vectors for Gasoline Trucks and Diesel Trucks are same or not at $\alpha = 0.01$ level of significance [$F_{3,31}(0.01) = 4.48$].

Solution:

```
rm(list=ls())
x11=c(16,7,10,4,11,14,13,13,29,13,7,10,10,11,12,10,10,9,12,8)
x21=c(12,3,2,6,5,6,11,14,15,8,6,4,5,6,14,3,6,3,8,14)
x31=c(11,4,10,8,11,10,11,9,3,10,8,9,10,8,14,6,12,12,11,12)
x=c(x11,x21,x31)
X1=matrix(x,ncol=3,byrow=F)
X1
      [,1] [,2] [,3]
 [1,]   16   12   11
 [2,]    7    3    4
 [3,]   10    2   10
 [4,]    4    6    8
 [5,]   11    5   11
 [6,]   14    6   10
```

```
 [7,]   13   11   11
 [8,]   13   14    9
 [9,]   29   15    3
[10,]   13    8   10
[11,]    7    6    8
[12,]   10    4    9
[13,]   10    5   10
[14,]   11    6    8
[15,]   12   14   14
[16,]   10    3    6
[17,]   10    6   12
[18,]    9    3   12
[19,]   12    8   11
[20,]    8   14   12
```

```
n1=20
x12=c(8,7,10,10,13,10,6,11,9,10,11,12,9,8,8)
x22=c(12,5,3,15,4,13,9,10,3,5,18,12,13,10,6)
x32=c(9,17,11,6,29,11,19,14,14,21,35,17,21,17,16)
y=c(x12,x22,x32)
X2=matrix(y,ncol=3,byrow=F)
X2
```

```
      [,1] [,2] [,3]
 [1,]    8   12    9
 [2,]    7    5   17
 [3,]   10    3   11
 [4,]   10   15    6
 [5,]   13    4   29
 [6,]   10   13   11
 [7,]    6    9   19
 [8,]   11   10   14
 [9,]    9    3   14
[10,]   10    5   21
[11,]   11   18   35
[12,]   12   12   17
[13,]    9   13   21
[14,]    8   10   17
[15,]    8    6   16
```

```
n2=15

M1=colMeans(X1)
M1

[1] 11.45  7.55  9.45

M2=colMeans(X2)
M2

[1]  9.466667  9.200000 17.133333

#mean difference
M=M1-M2
M

[1]  1.983333 -1.650000 -7.683333

X1bar=matrix(M1,nrow=20,ncol=3,byrow=T)
X1bar

       [,1] [,2] [,3]
 [1,] 11.45 7.55 9.45
 [2,] 11.45 7.55 9.45
```

```
 [3,] 11.45 7.55 9.45
 [4,] 11.45 7.55 9.45
 [5,] 11.45 7.55 9.45
 [6,] 11.45 7.55 9.45
 [7,] 11.45 7.55 9.45
 [8,] 11.45 7.55 9.45
 [9,] 11.45 7.55 9.45
[10,] 11.45 7.55 9.45
[11,] 11.45 7.55 9.45
[12,] 11.45 7.55 9.45
[13,] 11.45 7.55 9.45
[14,] 11.45 7.55 9.45
[15,] 11.45 7.55 9.45
[16,] 11.45 7.55 9.45
[17,] 11.45 7.55 9.45
[18,] 11.45 7.55 9.45
[19,] 11.45 7.55 9.45
[20,] 11.45 7.55 9.45
```

```
X2bar=matrix(M2,nrow=15,ncol=3,byrow=T)
X2bar
```

```
          [,1] [,2]     [,3]
 [1,] 9.466667  9.2 17.13333
 [2,] 9.466667  9.2 17.13333
 [3,] 9.466667  9.2 17.13333
 [4,] 9.466667  9.2 17.13333
 [5,] 9.466667  9.2 17.13333
 [6,] 9.466667  9.2 17.13333
 [7,] 9.466667  9.2 17.13333
 [8,] 9.466667  9.2 17.13333
 [9,] 9.466667  9.2 17.13333
[10,] 9.466667  9.2 17.13333
[11,] 9.466667  9.2 17.13333
[12,] 9.466667  9.2 17.13333
[13,] 9.466667  9.2 17.13333
[14,] 9.466667  9.2 17.13333
[15,] 9.466667  9.2 17.13333
```

```
a=(X1-X1bar)
a
```

```
       [,1]  [,2]  [,3]
 [1,]  4.55  4.45  1.55
 [2,] -4.45 -4.55 -5.45
 [3,] -1.45 -5.55  0.55
 [4,] -7.45 -1.55 -1.45
 [5,] -0.45 -2.55  1.55
 [6,]  2.55 -1.55  0.55
 [7,]  1.55  3.45  1.55
 [8,]  1.55  6.45 -0.45
 [9,] 17.55  7.45 -6.45
[10,]  1.55  0.45  0.55
[11,] -4.45 -1.55 -1.45
[12,] -1.45 -3.55 -0.45
[13,] -1.45 -2.55  0.55
[14,] -0.45 -1.55 -1.45
[15,]  0.55  6.45  4.55
[16,] -1.45 -4.55 -3.45
[17,] -1.45 -1.55  2.55
[18,] -2.45 -4.55  2.55
[19,]  0.55  0.45  1.55
[20,] -3.45  6.45  2.55
```

```
k=t(a)%*%a
k
```

```
       [,1]   [,2]   [,3]
[1,] 466.95 222.05 -72.05
[2,] 222.05 342.95  32.05
[3,] -72.05  32.05 140.95
```

```
b=X2-X2bar
b
```

```
             [,1] [,2]        [,3]
 [1,] -1.4666667  2.8  -8.1333333
 [2,] -2.4666667 -4.2  -0.1333333
 [3,]  0.5333333 -6.2  -6.1333333
 [4,]  0.5333333  5.8 -11.1333333
 [5,]  3.5333333 -5.2  11.8666667
 [6,]  0.5333333  3.8  -6.1333333
 [7,] -3.4666667 -0.2   1.8666667
 [8,]  1.5333333  0.8  -3.1333333
 [9,] -0.4666667 -6.2  -3.1333333
[10,]  0.5333333 -4.2   3.8666667
[11,]  1.5333333  8.8  17.8666667
[12,]  2.5333333  2.8  -0.1333333
[13,] -0.4666667  3.8   3.8666667
[14,] -1.4666667  0.8  -0.1333333
[15,] -1.4666667 -3.2  -1.1333333
```

```
l=t(b)%*%b
l
```

```
         [,1]  [,2]      [,3]
[1,] 49.73333  14.6  61.06667
[2,] 14.60000 306.4  41.60000
[3,] 61.06667  41.6 779.73333
```

```
#sample covariance matrix
s=(k+l)/(n1+n2-2)
s
```

```
           [,1]      [,2]       [,3]
[1,] 15.6570707  7.171212 -0.3328283
[2,]  7.1712121 19.677273  2.2318182
[3,] -0.3328283  2.231818 27.8994949
```

```
#calculate T^2
T={(n1*n2)/(n1+n2)}*t(M)%*%solve(s)%*%M
T
```

```
         [,1]
[1,] 22.01801
```

```
#calculate F
F={T/(n1+n2-2)}*{(n1+n2-2-3+1)/3}
F
```

```
         [,1]
[1,] 6.894529
```

SADS-II Lab Problem-9 **03-04-2024**

A doctor has collected data on cholesterol, blood pressure, and weight from 15 patients. He also collected data on the eating habits of the subjects (e.g., how many grams of red meat, fish, and dairy products consumed per week).

S.N. of Persons	Cholesterol (milligrams)	Systolic blood pressure (millimeter)	Weight (kg)	Red meat (grams)	Fish (grams)	Dairy Products(grams)
1	190	125	67	700	350	600
2	200	120	57	600	200	700
3	203	119	59	300	150	300
4	180	124	62	450	400	400
5	207	123	69	500	600	450
6	201	115	71	650	450	550
7	199	114	73	400	300	650
8	230	130	56	450	300	700
9	225	122	70	300	450	300
10	300	121	62	250	600	400
11	170	118	58	350	700	450
12	180	117	62	400	400	600
13	250	124	72	500	300	750
14	210	128	82	650	550	300
15	150	123	74	700	650	150

To test whether there are any relationship between the three measures of health and eating habits at 5 % level of significance.

Answer:-

Hypothesis:

Null: There is no relationship b/w the three measures of health and eating habits.

Alternative: There is significant relationship b/w the three measures of health and eating habits.

1. Open SPPS
2. Enter Data in SPSS
3. Go to analyze > General Linear Model > Multivariate
4. Move Cholesterol, Systolic BP and weight to dependent variable box and rest of the variable to covariate box.
5. Click on save and select unstandardized in predicted values and residuals.
6. Select Descriptive statistics and parameter estimates in option box.

7. Click on OK

Descriptive Statistics

	Mean	Std. Deviation	N
Cholesterol_ml	206.33	35.740	15
Systolic_ml	121.53	4.502	15
Weight	66.27	7.545	15

Bartlett's Test of Sphericity[a]

Likelihood Ratio	.000
Approx. Chi-Square	42.832
df	5
Sig.	.000

Tests the null hypothesis that the residual covariance matrix is proportional to an identity matrix.

a. Design: Intercept + Red_meat_gm + Fish_gm + Dairy_gm

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.965	82.939[b]	3.000	9.000	.000
	Wilks' Lambda	.035	82.939[b]	3.000	9.000	.000
	Hotelling's Trace	27.646	82.939[b]	3.000	9.000	.000
	Roy's Largest Root	27.646	82.939[b]	3.000	9.000	.000
Red_meat_gm	Pillai's Trace	.605	4.590[b]	3.000	9.000	.033
	Wilks' Lambda	.395	4.590[b]	3.000	9.000	.033
	Hotelling's Trace	1.530	4.590[b]	3.000	9.000	.033
	Roy's Largest Root	1.530	4.590[b]	3.000	9.000	.033
Fish_gm	Pillai's Trace	.025	.078[b]	3.000	9.000	.970
	Wilks' Lambda	.975	.078[b]	3.000	9.000	.970
	Hotelling's Trace	.026	.078[b]	3.000	9.000	.970
	Roy's Largest Root	.026	.078[b]	3.000	9.000	.970
Dairy_gm	Pillai's Trace	.260	1.051[b]	3.000	9.000	.416
	Wilks' Lambda	.740	1.051[b]	3.000	9.000	.416
	Hotelling's Trace	.350	1.051[b]	3.000	9.000	.416
	Roy's Largest Root	.350	1.051[b]	3.000	9.000	.416

a. Design: Intercept + Red_meat_gm + Fish_gm + Dairy_gm

b. Exact statistic

The effect of red_meat on measure of health is significant since the significant level is 0.033 which Is not sufficient to accept null hypothesis at 5% level of significance.

The effect of fish on measure of health is not significant since the significant level is 0.970 which Is sufficient to accept null hypothesis at 5% level of significance.

The effect of dairy_product on measure of health is not significant since the significant level is 0.416 which Is sufficient to accept null hypothesis at 5% level of significance.

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Cholesterol_ml	4604.839[a]	3	1534.946	1.272	.332
	Systolic_ml	24.134[b]	3	8.045	.341	.796
	Weight	266.615[c]	3	88.872	1.843	.198
Intercept	Cholesterol_ml	18844.387	1	18844.387	15.611	.002
	Systolic_ml	5412.038	1	5412.038	229.325	.000
	Weight	1241.174	1	1241.174	25.745	.000
Red_meat_gm	Cholesterol_ml	3736.289	1	3736.289	3.095	.106
	Systolic_ml	21.705	1	21.705	.920	.358
	Weight	172.543	1	172.543	3.579	.085
Fish_gm	Cholesterol_ml	86.668	1	86.668	.072	.794
	Systolic_ml	.109	1	.109	.005	.947
	Weight	12.967	1	12.967	.269	.614
Dairy_gm	Cholesterol_ml	1103.541	1	1103.541	.914	.360
	Systolic_ml	3.080	1	3.080	.131	.725

	Weight	34.786	1	34.786	.722	.414
Error	Cholesterol_ml	13278.494	11	1207.136		
	Systolic_ml	259.599	11	23.600		
	Weight	530.318	11	48.211		
Total	Cholesterol_ml	656485.000	15			
	Systolic_ml	221839.000	15			
	Weight	66666.000	15			
Corrected Total	Cholesterol_ml	17883.333	14			
	Systolic_ml	283.733	14			
	Weight	796.933	14			

a. R Squared = .257 (Adjusted R Squared = .055)

b. R Squared = .085 (Adjusted R Squared = -.164)

c. R Squared = .335 (Adjusted R Squared = .153)

There is no significant impact on cholesterol level by eating habits since P value is 0.332 which is not sufficient to reject null hypothesis at 5% level of significance

There is no significant impact on systolic BP by eating habits since P value is 0.796 which is not sufficient to reject null hypothesis at 5% level of significance

Conclusion:-

There is no significant impact on weight by eating habits since P value is 0.198 which is not sufficient to reject null hypothesis at 5% level of significance

Principle Component

Suresh kumar prajapati

2024-04-13

CAS-II problem-10 Date: 10-04-2024

Determine the first and second principal components Y1 and Y2for the covariance matrix of the random vector X'= [X1 X2 X3] sigma =matrix(c(1,-2,0,-2,5,0,0,0,2),nrow=3,ncol=3)

Also calculate the variance of the first and second principal components and proportion of the total population variance explained by the first principal component.Also calculate correlation coefficient between first and second(y1,y2) and random variable x1,x2,x3.

formulation of Var Cov Matrix

```
SIGMA=matrix(c(1,-2,0,-2,5,0,0,0,2),nrow = 3,ncol=3)
SIGMA
```

```
##      [,1] [,2] [,3]
## [1,]    1   -2    0
## [2,]   -2    5    0
## [3,]    0    0    2
```

finding Eigen Value and Eigen Vector

```
EIGEN_VALUE=eigen(SIGMA)
EIGEN_VALUE
```

```
## eigen() decomposition
## $values
## [1] 5.8284271 2.0000000 0.1715729
##
## $vectors
##            [,1] [,2]      [,3]
## [1,] -0.3826834    0 0.9238795
## [2,]  0.9238795    0 0.3826834
## [3,]  0.0000000    1 0.0000000
```

for normalize eigen vector

```
 VECTORS=EIGEN_VALUE$vectors
VECTORS
```

```
##            [,1] [,2]      [,3]
## [1,] -0.3826834    0 0.9238795
## [2,]  0.9238795    0 0.3826834
## [3,]  0.0000000    1 0.0000000
```

```
E1=VECTORS[,1]
E1
```

```
## [1] -0.3826834  0.9238795  0.0000000
```

```
E2=VECTORS[,2]
E2
```

```
## [1] 0 0 1
```

```
E3=VECTORS[,3]
E3
```

```
## [1] 0.9238795 0.3826834 0.0000000
```

##Finding varaiance of Principal components

```
#Y1 =E1.X=-0.383(X1)-.924(X3)
#VAR_Y1=var(Y1)
VAR_Y1=(((-0.3826834)^2)*(1)+((0.9238795 )^2)*(5)+2*(-0.3826834)*(0.9238795 )*(-2))
VAR_Y1
```

```
## [1] 5.828427
```

```
#Y3=E3.X=0.924(X1)+0.383(X3)
VAR_Y3=(((0.9238795)^2)*(1)+((0.3826834)^2)*(5)+2*(0.9238795)*(0.3826834)*(-2))
VAR_Y3
```

```
## [1] 0.1715729
```

```
#Y2=E2.X=1(X3)
  VAR_Y2=1*2
  VAR_Y2
```

```
## [1] 2
```

###correlation=((eii*sqrt(lamda1))/sqrt(var1)

```
var_x1=1
var_x2=5
var_x3=2
cov_x1_x2=-2
cov_x1_x3=0
cov_x2_x3=0

cor_y1_x1=(-0.3826834 *(sqrt(5.8284271))/sqrt(1))
cor_y1_x1
```

```
## [1] -0.9238795
```

```
cor_y1_x2=(0.9238795*(sqrt(5.8284271))/sqrt(5))
cor_y1_x2
```

```
## [1] 0.9974842
```

```
cor_y1_x3=0*(sqrt(5.8284271))/sqrt(2)
cor_y1_x3
```

```
## [1] 0
```

SADS-II problem-11 **Date: 24-04-2024**

Perspiration from a sample of 20 healthy females was analyzed. Three components, X_1 = Sweet rate, X_2= Sodium content and X_3= Potassium content, were measured and the results, which we call the sweat data are given in the below table.

Sweat Data			
Individual	X_1 (Sweat rate)	X_2 (Sodium)	X_3 (Potassium)
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14.0
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Find the principal components of the above data and also find the proportion of total variation explained by first and second principal components.

Solution:-

```
> setwd("D:\\dataset\\")

> data=read.csv("principalcom.csv")
> data
     X1   X2   X3
1   3.7 48.5  9.3
2   5.7 65.1  8.0
3   3.8 47.2 10.9
4   3.2 53.2 12.0
5   3.1 55.5  9.7
6   4.6 36.1  7.9
7   2.4 24.8 14.0
8   7.2 33.1  7.6
9   6.7 47.1  8.5
10 5.4 54.1 11.3
11 3.9 36.9 12.7
12 4.5 58.8 12.3
13 3.5 27.8  9.8
```

```
14 4.5 40.2  8.4
15 1.5 13.5 10.1
16 8.5 56.4  7.1
17 4.5 71.6  8.2
18 6.5 52.8 10.9
19 4.1 44.1 11.2
20 5.5 40.9  9.4
> head(data)
    X1   X2   X3
1 3.7 48.5  9.3
2 5.7 65.1  8.0
3 3.8 47.2 10.9
4 3.2 53.2 12.0
5 3.1 55.5  9.7
6 4.6 36.1  7.9
> setwd("D:\\dataset\\")
> data=read.csv("principalcom.csv")
> head(data)
    X1   X2   X3
1 3.7 48.5  9.3
2 5.7 65.1  8.0
3 3.8 47.2 10.9
4 3.2 53.2 12.0
5 3.1 55.5  9.7
6 4.6 36.1  7.9
> ##step2: Standardize the data
> scale_data=scale(data)
> head(scale_data)
                  X1         X2         X3
[1,] -0.55396106  0.2204127 -0.3491471
[2,]  0.62467949  1.3950038 -1.0316904
[3,] -0.49502903  0.1284267  0.4909061
[4,] -0.84862119  0.5529777  1.0684427
[5,] -0.90755322  0.7157222 -0.1391338
[6,] -0.02357281 -0.6569927 -1.0841937
> ##Step 3: Perform Principal Component Analysis (PCA)
> # Perform PCA
> pca_result <- prcomp(scale_data, scale. = TRUE)
> pca_result
Standard deviations (1, .., p=3):
[1] 1.3440666 0.8955140 0.6257313

Rotation (n x k) = (3 x 3):
             PC1        PC2        PC3
X1  0.6534351 -0.1023336 -0.7500336
X2  0.4870226  0.8153574  0.3130516
X3 -0.5795097  0.5698422 -0.5826219
> # Summary of PCA
> summary(pca_result)
Importance of components:
                                  PC1    PC2
Standard deviation     1.3441 0.8955
Proportion of Variance 0.6022 0.2673
Cumulative Proportion  0.6022 0.8695
                                  PC3
Standard deviation     0.6257
Proportion of Variance 0.1305
Cumulative Proportion  1.0000
> ##Step 4: Interpret Principal Components Extract and Interpret Loadings
> # Extract loadings
> loadings <- pca_result$rotation
>
> # Print loadings
> print(loadings)
             PC1        PC2        PC3
X1  0.6534351 -0.1023336 -0.7500336
```

```
X2  0.4870226  0.8153574  0.3130516
X3 -0.5795097  0.5698422 -0.5826219
> ##step5:Proportion of Variance
> # Proportion of variance explained
> prop_variance <- pca_result$sdev^2 / sum(pca_result$sdev^2)
>
> # Print proportion of variance
> print(prop_variance)
[1] 0.6021717 0.2673151 0.1305132
> ##step6: percentage
> per =((0.6021717+0.2673151)/sum(prop_variance))*100
> per
[1] 86.94868
```

CAS Problem-12 24-04-2024

The owner of Pizza Corner, Varanasi collected data on six variables such as no. of delivery boys (X_1), cost of advertisements in rupees (X_2), number of outlets (X_3), varieties of pizzas (X_4), competitor's activities index (X_5) from past 15 months.

X1	X2	X3	X4	X5	X6
15	20	35	17	4	70
10	12	10	13	4	43
7	11	14	14	3	31
2	6	9	13	3	10
4	10	11	12	4	17
1	5	6	12	5	8
4	14	15	15	2	39
7	12	16	16	3	40
5	10	18	15	4	30
3	5	8	13	2	16
13	17	20	14	2	30
2	9	10	12	3	20
5	12	15	12	3	25
12	18	30	15	4	50
1	5	6	12	5	20

By using factor analysis to reduce these six variables into factors.

Solution:

Factor Analysis

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.707
Bartlett's Test of Sphericity	Approx. Chi-Square	76.451
	df	15
	Sig.	.000

Communalities

	Initial	Extraction
X1	1.000	.852
X2	1.000	.907
X3	1.000	.910
X4	1.000	.744
X5	1.000	.998
X6	1.000	.911

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.302	71.708	71.708	4.302	71.708	71.708	4.278	71.300	71.300
2	1.018	16.974	88.682	1.018	16.974	88.682	1.043	17.382	88.682
3	.386	6.434	95.116						
4	.147	2.453	97.569						
5	.105	1.755	99.324						
6	.041	.676	100.000						

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component	
	1	2
X1	.923	.028
X2	.951	-.056
X3	.949	.096
X4	.859	-.074
X5	-.138	.989
X6	.943	.145

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Rotated Component Matrix[a]

	Component	
	1	2
X1	.922	-.052
X2	.942	-.138
X3	.954	.014
X4	.850	-.148
X5	-.052	.998
X6	.952	.063

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Component Transformation Matrix

Component	1	2
1	.996	-.086
2	.086	.996

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

Conclusion: - KMO and Bartletts test .707>0.05 so it is good for factor analysis there are two factor percentage of variance exist.

SADS-Problem-13
01-05-2024

The data on weight (kg) and height (cm) of 10 PG students are given below.

Weight	67 65 63 66 70 63 62 45 54 55 65
Height	167 166 165 166 168 162 156 153 160 152 163

Find the clusters of objects of the above data using agglomerative hierarchical clustering methods and also plot the dendrogram.

SOLUTION:-

Step:-

- Go to analyze
- Choose classify
- Select K- Means cluster
- Select Variable
- Iterative by Default & Save in Cluster membership & Distance from cluster center
- Go to options – Initial cluster centers & missing value Exclude cases listwise
- OK

LINK:-

```
QUICK CLUSTER Weight Heght
 /MISSING=LISTWISE
 /CRITERIA=CLUSTER(3) MXITER(10) CONVERGE(0)
 /METHOD=KMEANS(NOUPDATE)
 /SAVE CLUSTER
 /PRINT INITIAL CLUSTER DISTAN.
```

Initial Cluster Centers

	Cluster		
	1	2	3
Weight	70	45	62
Heght	168	153	156

Intial clusture

Devide in three cluster of 11 items

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	4.549	.000	3.808
2	.695	.000	2.121
3	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 14.422.

Itreration see the minumum distance

Cluster Membership

Case Number	Cluster	Distance
1	1	2.231
2	1	.915
3	1	2.587
4	1	.833
5	1	5.194
6	1	4.172
7	3	5.000
8	2	.000
9	3	5.000
10	3	4.472
11	1	2.356

In Cluster Membership in first cluster with 1,2,3,4,5,6 and 11

2nd only 8

3rd 7,9 and 10

Final Cluster Centers

	Cluster		
	1	2	3
Weight	66	45	57
Heght	165	153	156

#Final cluster centers

Distances between Final Cluster Centers

Cluster	1	2	3
1		23.961	12.637
2	23.961		12.369
3	12.637	12.369	

Number of Cases in each Cluster

Cluster	1	7.000
	2	1.000
	3	3.000
Valid		11.000
Missing		.000

Distances between Final Cluster Centers

Cluster	1	2	3
1		23.961	12.637
2	23.961		12.369
3	12.637	12.369	

Hierarchical Cluster

Steps :-

» Go to analyze and then classify

» Go to Hierarchical cluster select the variable.

» Go to statistics to choose agglomeration schedule and Proximity matrix use single solution of 3 cluster.

»Go to plot and set Dendrogram.

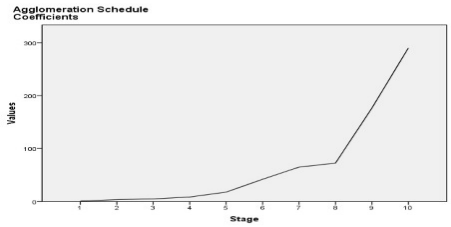
»Go to method and choose Between -groups linkage

»Square Euclidean distance & also in save for single solution.

Average Linkage (Between Groups)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	4	1.000	0	0	2
2	1	2	3.500	0	1	5
3	6	11	5.000	0	0	4
4	3	6	8.500	0	3	5
5	1	3	17.778	2	4	6
6	1	5	42.000	5	0	10
7	9	10	65.000	0	0	8
8	7	9	72.500	0	7	9
9	7	8	176.333	8	0	10
10	1	7	289.821	6	9	0



Cluster Membership

Case	3 Clusters
1	1
2	1
3	1
4	1
5	1
6	1
7	2
8	3
9	2
10	2
11	1

Quick Cluster

Case Processing Summary[a]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
11	100.0	0	.0	11	100.0

a. Average Linkage (Between Groups)

Proximity Matrix

Case	Squared Euclidean Distance								
	1	2	3	4	5	6	7	8	9
1	.000	5.000	20.000	2.000	10.000	41.000	146.000	680.000	218.000
2	5.000	.000	5.000	1.000	29.000	20.000	109.000	569.000	157.000
3	20.000	5.000	.000	10.000	58.000	9.000	82.000	468.000	106.000
4	2.000	1.000	10.000	.000	20.000	25.000	116.000	610.000	180.000
5	10.000	29.000	58.000	20.000	.000	85.000	208.000	850.000	320.000
6	41.000	20.000	9.000	25.000	85.000	.000	37.000	405.000	85.000
7	146.000	109.000	82.000	116.000	208.000	37.000	.000	298.000	80.000
8	680.000	569.000	468.000	610.000	850.000	405.000	298.000	.000	130.000
9	218.000	157.000	106.000	180.000	320.000	85.000	80.000	130.000	.000
10	369.000	296.000	233.000	317.000	481.000	164.000	65.000	101.000	65.000
11	20.000	9.000	8.000	10.000	50.000	5.000	58.000	500.000	130.000

This is a dissimilarity matrix

Number of Cases in each Cluster

Cluster	1	7.000
	2	1.000
	3	3.000
Valid		11.000
Missing		.000

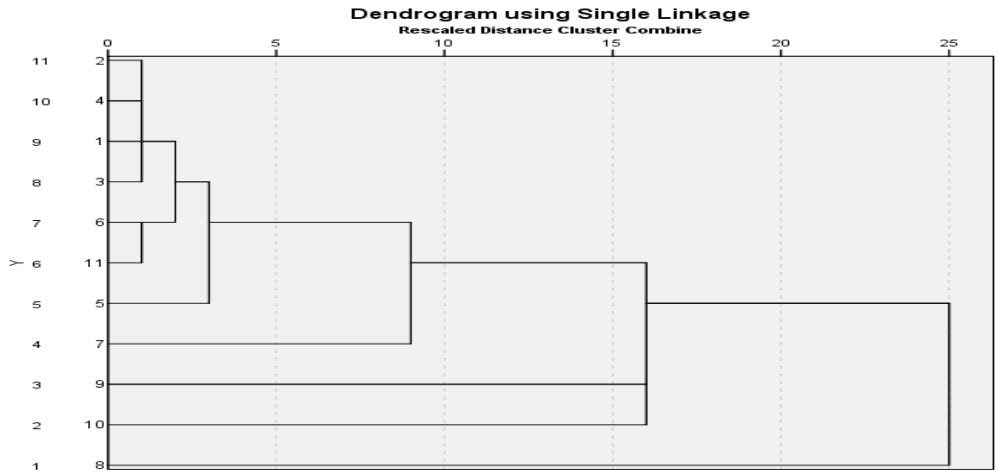
Number of Cases in each Cluster

Cluster	1	7.000
	2	1.000
	3	3.000
Valid		11.000
Missing		.000

Single Linkage

Cluster Membership

Case	3 Clusters
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	2
9	3
10	3
11	1



Conclusion:- Plot Dendrogram Average linkage method