PRACTICAL-I

MULTIVARIATE ANALYSIS

&

TESTING OF HYPOTHESIS

M.Sc., STATISTICS, First Year

SEMESTER-II, PAPER - V

(205ST24)

LESSON WRITER
Dr. Syed Jilani
Guest Faculty
Department of Statistics
Acharya Nagarjuna University

EDITOR
Dr. A. Vasudeva Rao
Honorary Professor,
Department of Statistics,
Acharya Nagarjuna University,

ACADEMIC ADVISOR
Prof. G. V. S. R. Anjaneyulu
Professor of Statistics (Retd.)
Acharya Nagarjuna University

DIRECTOR, I/c.
Prof. V. Venkateswarlu
M.A., M.P.S., M.S.W., M.Phil., Ph.D.
Centre for Distance Education
Acharya Nagarjuna University
Nagarjuna Nagar 522 510

Ph: 0863-2346222, 2346208
0863- 2346259 (Study Material)
Website www.anucde.info
E-mail: anucdedirector@gmail.com

M.Sc., STATISTICS: Multivariate Analysis & Testing of Hypothesis
PRACTICAL-I (205ST24)

First Edition : 2025

No. of Copies :

Published by:

Prof. V. VENKATESWARLU
Director, I/c
Centre for Distance Education,
Acharya Nagarjuna University

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A+' grade from the NAAC in the year 2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the door step of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof. K. Gangadhara Rao
M.Tech., Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University.

CONTENTS

MULTIVARIATE ANALYSIS - PRACTICAL

TESTING OF HYPOTHESIS - PRACTICAL

LAB EXERCISE 1:

Two sample Hotelling T^2 Statistic

Problem:

In the first phase of a study of the coast of transporting milk from farms to dairy plants, a survey was taken of firms engaged in milk transportation. 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=15 gasoline and n_2=23 diesel trucks.

Gasoline trucks			Diesel trucks		
X_1	X_2	X_3	X_1	X_2	X_3
16.44	12.43	11.23	8.50	12.26	9.11
7.19	2.70	3.92	7.42	5.13	17.15
9.92	1.35	9.75	10.28	3.32	11.23
4.24	5.78	7.78	10.16	14.72	5.99
11.20	5.05	10.67	12.79	4.17	29.28
14.25	5.78	9.88	9.60	12.72	11.00
13.50	10.98	10.60	6.47	8.89	19.00
13.32	14.27	9.45	11.35	9.95	14.53
29.11	15.09	3.28	9.15	2.94	13.68
12.68	7.61	10.23	9.70	5.06	20.84
7.51	5.80	8.13	9.77	17.86	35.18
9.90	3.63	9.13	11.61	11.75	17.00
10.25	5.07	10.17	9.09	13.25	20.66
11.11	6.15	7.61	8.53	10.14	17.45
12.17	14.26	14.9	8.29	6.22	16.38

Test for differences in the mean cost vectors of gasoline and diesel trucks at α=5% l.o.s. by applying Hotelling's T^2.

Aim:

To test whether gasoline and diesel trucks have same mean cost vectors or not using two sample Hotelling's T^2 statistic.

Procedure:

In this problem, we have to examine whether the mean vector of one MVN populations equations to the mean vector of another MVN populations

$$H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)} \text{ i.e., } \underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)} = \underset{\sim}{0} ,$$

Let

$$\Pi_1 : \underset{\sim}{X}_1, \underset{\sim}{X}_2 .... \underset{\sim}{X}_m \sim N_p\left(\underset{\sim}{\mu}_1, \Sigma\right)$$

$$\Pi_2 : \underset{\sim}{Y}_1, \underset{\sim}{Y}_2 .... \underset{\sim}{Y}_n \sim N_p\left(\underset{\sim}{\mu}_2, \Sigma\right)$$

Now the Hotelling T^2statistic for testing the above null hypothesis is given as

$$\text{Hotelling's } T^2 = \left(\frac{\text{mn}}{\text{m+n}}\right)\left(\underset{\sim}{\bar{\mathbf{X}}} - \underset{\sim}{\bar{Y}}\right)' \mathbf{S}^{-1}\left(\underset{\sim}{\bar{\mathbf{X}}} - \underset{\sim}{\bar{Y}}\right) \sim T^2_{\text{m+n-2}}$$

where

$$\underset{\sim}{\bar{\mathbf{x}}} = \text{ Sample mean vector of } \underset{\sim}{\mathbf{x}}_1, \underset{\sim}{\mathbf{x}}_2 .... \underset{\sim}{\mathbf{x}}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} \underset{\sim}{\mathbf{x}}_i$$

$$\underset{\sim}{\bar{\mathbf{y}}} = \text{ Sample mean vector of } \underset{\sim}{\mathbf{y}}_1, \underset{\sim}{\mathbf{y}}_2 .... \underset{\sim}{\mathbf{y}}_{\mathbf{n}} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{\mathbf{y}}_i$$

and the pooled sample dispersion (variance-covariance) matrix is given by

$$S = \frac{(m-1)S_1 + (n-1)S_2}{m+n-1}$$

$$\mathbf{S}_1 = \frac{1}{(m-1)}\sum_{i=1}^{m}(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\bar{\mathbf{x}}})(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\bar{\mathbf{x}}})'$$

$$\mathbf{S}_2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(\underset{\sim}{\mathbf{y}}_i - \underset{\sim}{\bar{\mathbf{y}}})(\underset{\sim}{\mathbf{y}}_i - \underset{\sim}{\bar{\mathbf{y}}})'$$

Now the critical value of T^2 at α level of significance is given by

$$T_0^2 = \frac{\text{p(m+n- 2)}}{\text{m+ n- p - 1}} F_{\text{p , m+n- p -1}}(\alpha).$$

Conclusion:

If calculated $T^2 > T_0^2$, we reject H_0, otherwise we accept $H_0 =$

R-CODE:

```
hot2samp=function(data){
nc=ncol(data);
x=subset(data[,-nc],data[,nc]==1);
y=subset(data[,-nc],data[,nc]==2);
cat("\n First Sample:\n"); print(x);
cat("\n Second Sample:\n"); print(y);
p=ncol(x);q=ncol(y);
m=nrow(x); n=nrow(y);
xbar=colMeans(x); ybar=colMeans(y);
sx=cov(x); sy=cov(y);
s=((m-1)*sx+(n-1)*sy)/(m+n-2);
cat("\n x-bar=[",xbar,"]");
cat("\n y-bar=[",ybar,"]");
cat("\n S-matrix:\n");
print(round(s,4));
tsq=(m*n/(m+n))*(t(xbar-ybar))%*%solve(s,xbar-ybar);
#hot t^2 value.
t0sq=(m+n-2)*p/(m+n-p-1)*qf(0.95,p,m+n-1);  #critical  T^2
value.
cat("\n Hotellings' calculated T^2=",tsq,"\n");
cat("\n Hotellings' critical T^2(5% los)=",t0sq,"\n");
if(tsq<=t0sq) {cat("Based on the given data we accept H0
at 5% los \n");
               cat("That is we conclude that the given
two samples have been drawn from MVN population. \n");
}
if(tsq>t0sq) {cat("Based on the given data we reject H0
at 5% los \n");
cat("That is we conclude that the given two samples have
not been drawn from MVN population. \n");
}
}
data=read.csv("hotT2SAMPLES_TURTLES.csv",header=T);
hot2samp(data);
```

INFERENCE :

Calculated T^2 value = 20.26095

Critical T^2 value = 9.612036

Since the calculated T^2 value is greater than the critical T^2 value, we may conclude that the average milk transport cost of diesel trucks is different from gasoline trucks.

LAB EXERCISE 2:

Repeated Measures Design

Problem:

Improved anesthetics are often developed by first studying their effects on animals. In one study 19 dogs were initially given the drug pentobarbital. Each dog was then administered carbon dioxide (co_2) at each of two pressure levels. Next halothane (H) was added and the administration of co_2 was repeated. The response, milliseconds between heartbeats, was measured for the 4 treatments combinations.

Now the treatments are as follows:

T_1=high co_2 pressure without H
T_2=low co_2 pressure without H
T_3=high co_2 pressure with H
T_4=low co_2 pressure with H

The four measurements for each of the 19 dogs-data as follows

Dogs	**Treatments**			
	1	2	3	4
1	426	609	566	600
2	253	236	392	395
3	359	433	349	367
4	432	431	542	600
5	415	426	513	513
6	324	438	507	539
7	315	312	410	456
8	326	329	350	504
9	375	447	540	548
10	286	286	403	422
11	349	382	473	457
12	429	410	488	547
13	348	367	447	514
14	412	473	472	446
15	397	326	455	468
16	434	458	637	524
17	394	367	432	469
18	420	395	418	431
19	397	566	645	625

Analyze the anesthetizing effects of CO_2 pressure and Halothane (H) using repeating measures designs.

Aim: To test the equality of treatments and analyze the anesthetizing effects of CO_2 presure and halothane (H) using repeated measures designs.

Procedure:

Consider a multivariate normal population $N_p(\underset{\sim}{\mu},\Sigma)$. Let 'C' be a matrix of known constants

Null hypothesis: Under Null hypothesis we have $H_0 : \mathbf{C}\underset{\sim}{\boldsymbol{\mu}} = \underset{\sim}{\mathbf{0}}$

Alternative Hypothesis: Under alternative hypothesis we have $H_1 : C\underset{\sim}{\boldsymbol{\mu}} \neq \underset{\sim}{\mathbf{0}}$

TEST STATISTIC: To test the hypothesis we have the test statistic

$$T^2 = n(C\underset{\sim}{\bar{\mathbf{X}}})'(CSC')^{-1}(C\underset{\sim}{\bar{\mathbf{X}}}) \square T^2_{n-1}$$

Where

$$\underset{\sim}{\bar{\mathbf{X}}} = \frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{\mathbf{x}}_i$$

$$\mathbf{S} = \frac{1}{(n-1)}\sum_{i=1}^{n}(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\mathbf{x}})(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\mathbf{x}})'$$

$$\mathbf{C} = \begin{bmatrix} -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

At the given α level of significance, H_0 may be rejected in favor of H_1 if

$T^2 > T_0^2$, where $T_0^2 = \dfrac{(n - 1)\,(q-1)}{n - q+1} F_{q-1,n-q+1}\,(\alpha)$ and $F_{q-1,n-q+1}\,(\alpha)$ is the upper $100\,\alpha^{th}$ percentile of the F-distribution and can be obtained from the F-tables.

R-CODE: -

```
# R-CODE FOR COMPUTING HOTELLING'S T^2 STATISTIC FOR REPEATED
MEASURES DESIGN
data=read.csv("rmd_dogs.csv",header=T);
#data=data[,-1];
RMD=function(data){
X=as.matrix(data);
n=nrow(X);q=ncol(X);
C=matrix(c(-1,1,1,-1,-1,-1,1,1,-1,1,-1,1),3);
cat("\n C-matrix:\n");print(C);
xbar=colMeans(X); # COMPUTES XBAR
cat("x-bar=[",xbar,"]\n");
S=cov(X);#computes S-matrix
print("S-matrix:"); print(S);
CSC=C%*%S%*%t(C);
zbar=C%*%xbar;
Tsq=n*t(zbar)%*%solve(CSC,zbar);
cat("Hotelling's CALCULATED T^2=",Tsq,"\n");
T0sq=(n-1)*(q-1)/(n-q+1)*qf(0.95,q-1,n-q+1);
cat("Hotelling's CRITICAL T^2(5% los) =",T0sq,"\n");
if (Tsq<=T0sq) {cat("Based on the given data we Accept H0 AT
5% LOS.\n");
cat("That is we conclude that the given TREATMENTS are
equal\n");}
if(Tsq>T0sq) {cat("Based on the given data we Reject H0 at 5%
LOS.\n");
cat("That is we conclude that the given TREATMENTS are NOT
equal \n");      }
}
RMD(data)
```

OUTPUT:

```
C-matrix:
     [,1] [,2] [,3] [,4]
[1,]   -1   -1    1    1
[2,]    1   -1    1   -1
[3,]    1   -1   -1    1
x-bar=[ 373.2105 404.7895 475.7368 496.0526 ]
[1] "S-matrix:"
         T1       T2       T3       T4
T1 2865.398 3292.769 2656.836 1920.488
T2 3292.769 8147.398 5567.553 4248.012
T3 2656.836 5567.553 7158.871 4759.848
T4 1920.488 4248.012 4759.848 5031.386

Hotelling's CALCULATED T^2= 77.22728
Hotelling's CRITICAL T^2(5% los) = 10.93119
```

Based on the given data we Reject H0 at 5% LOS.

That is we conclude that the given TREATMENTS are NOT equal

INFERENCE:

Hotelling's CALCULATED T^2= 77.22728

Hotelling's CRITICAL T^2(5% los) = 10.93119

Thus, calculated T^2 is greater than the critical T^2 i.e., $T\text{^}2_{(cal)}$>$T\text{^}2_{(crit)}$

Therefore, we conclude that the given Treatments T_1, T_2, T_3 and T_4 are NOT equal.

LAB EXERCISE 3:

Paired Sample in Hotelling T^2 Statistics

Problem:

Municipal waste water treatment plants are required by law to monitor their discharges into rivers and streams on a regular basis. Concern about the reliability of data from one of these self monitoring programmers lead to a study in which samples of efficient are divided and sent to two laboratories for testing. One half of each sample was sent to the Wisconsin state laboratories of hygiene and another half was sent to a private commercial lab routinely used in the monitoring program. Measurement of bio-chemical oxygen demand (BOD) and suspended solids (S.S) were obtained for n=11 samples splits from the two laboratories the data displayed below.

Efficient Data:- Sample	Commercial lab		State Lab of hygiene	
	$X_1^{(1)}$(BOD)	$X_1^{(1)}$(S.S)	$X_1^{(2)}$*(BOD)*	$X_1^{(2)}$(S.S)
1	10	27	25	15
2	6	23	28	30
3	15	64	36	29
4	8	42	45	29
5	11	30	15	36
6	34	79	49	64
7	28	26	48	30
8	70	24	54	68
9	43	54	34	56
10	30	30	29	32
11	25	14	35	21

Does the two laboratories chemical analysis agree? Apply Hotelling's T^2 test.

Aim: To test whether the two laboratories chemical analysis agree or not by using Hotelling's T^2 Statistic procedure.

Procedure:

Let $\underset{\sim}{\mathbf{x}}_j^{(1)}$ denote the response of an individual before the test and $\underset{\sim}{\mathbf{x}}_j^{(2)}$ denotes the response of the individual after the treatment i.e., $\left(\underset{\sim}{\mathbf{x}}_j^{(1)}, \underset{\sim}{\mathbf{x}}_j^{(2)}\right)$ are p measurements recorded on j^{th} unit. The 'n' differences $\underset{\sim}{\mathbf{d}}_j = \underset{\sim}{\mathbf{x}}_j^{(1)} - \underset{\sim}{\mathbf{x}}_j^{(2)}$, j = 1,2,3,...,n represents independent observations from an $N_p(\underset{\sim}{\boldsymbol{\mu}}_d, \Sigma_d)$. Now, for testing the hypothesis $H_0 : \underset{\sim}{\boldsymbol{\mu}}_d = \underset{\sim}{0}$ *vs* $H_1 : \underset{\sim}{\boldsymbol{\mu}}_d \neq \underset{\sim}{0}$ we have the following test statistic.

TEST STATISTIC:

To test the hypothesis we have the test statistic

$$T_d^2 = n\bar{\underset{\sim}{\mathbf{d}}}' S_d^{-1}\bar{\underset{\sim}{\mathbf{d}}} \sim T_{n-1}^2$$

where

$$\bar{\underset{\sim}{\mathbf{d}}} = \frac{1}{n}\sum_{i=1}^{n}\bar{\underset{\sim}{\mathbf{d}}}_i$$

$$\mathbf{S}_d = \frac{1}{(n-1)}\sum_{i=1}^{n}(\bar{\underset{\sim}{\mathbf{d}}}_i - \bar{\underset{\sim}{\mathbf{d}}})(\bar{\underset{\sim}{\mathbf{d}}}_i - \bar{\underset{\sim}{\mathbf{d}}})'$$

At the given α % level of significance, H_0 may be rejected in favor of H_1 if $T_d^2 > T_0^2$, where,

$T_0^2 = \frac{(n-1)p}{(n-p)}F_{p,n-p}{}^{(\alpha)}$. The upper 100 α^{th} percentile of the F – distribution and can be obtained from the F – Tables.

R-CODE:

```
# R-CODE FOR COMPUTING HOTELLING'S T^2 STATISTIC IN CASE OF
PAIRED SAMPLES
pairedhT2=function(X,Y){
data=read.csv("PairedT_MUNICAPAL.csv",header=T);
nc=ncol(data);
data=as.matrix(data);
X=subset(data[,-nc],data[,nc]==1);
Y=subset(data[,-nc],data[,nc]==2);
D=X-Y;
n=nrow(D);
i=rep(1,n);
dbar=t(D)%*%i/n;# COMPUTES DBAR
Sd=(t(D)%*%D-n*dbar%*%t(dbar))/(n-1);#computes covariance-
matrix
p=ncol(D);
cat("d-bar=[",dbar,"]\n");
print("Sample variance-covariance matrix:"); print(Sd);
print("Inverse of the matrix:");print(solve(Sd));

Tsq=n*sum(dbar*solve(Sd,dbar));# Hotelling T^2 value
cat("Hotelling's CALCULATED T^2=",Tsq,"\n");
T0sq=(n-1)*p/(n-p)*qf(0.95,p,n-p);
cat("Hotelling's CRITICAL T^2(5% los) =",T0sq,"\n");
if (Tsq<=T0sq) {cat("Based on the given data we Accept H0 AT
5% LOS.\n");
cat("That is we conclude that there is NO significant
difference between two sample mean vectors\n");}
if(Tsq>T0sq) {cat("Based on the given data we Reject H0 at 5%
LOS.\n");
cat("That is we conclude that there is SIGNIFICANT DIFFERENCE
between two sample mean vectors\n");}
}
pairedhT2(X,Y)
```

OUTPUT:-

```
d-bar=[ -10.72727 0.2727273 ]
[1] "Sample variance-covariance matrix:"
        BOD       SS
BOD  233.2182 -198.4818
SS  -198.4818  385.6182
[1] "Inverse of the matrix:"
          BOD          SS
BOD 0.007630241 0.003927367
SS  0.003927367 0.004614697
Hotelling's CALCULATED T^2= 9.409494
Hotelling's CRITICAL T^2(5% los) = 9.458877
Based on the given data we Accept H0 AT 5% LOS.
That is we conclude that there is NO significant difference between two sample mean vectors
```

INFERENCE:

Hotelling's CALCULATED T^2= 9.409494

Hotelling's CRITICAL T^2(5% los) = 9.458877

Since, Cal T^2 < Critical T^2, we may conclude that the two laboratories (Commercial lab and State Lab) chemical analysis agreed with each other.

LAB EXERCISE 4:

Mahalnobis's D^2 Statistic.

Problem:

Researchers interested in assessing pulmonary function in no pathological populations asked subjects to run on a treadmill until exhaustion. Samples of air were collected at definite intervals and the gas contents analyzed. The results on 4 measures of oxygen consumption for 15 males and 15 females are given in below table. The variables were

x_1 = Resting Volume O_2 (L/min)
x_2 = Resting Volume O_2 (mL/kg/min)
x_3 = maximum Volume O_2 (L/min)
x_4 = maximum Volume O_2 (mL/kg/min)

Female				Male			
x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
0.34	3.71	2.87	30.87	0.29	5.04	1.93	33.85
0.39	5.08	3.38	43.85	0.28	3.95	2.51	35.82
0.48	5.13	4.13	44.51	0.31	4.88	2.31	36.40
0.31	3.95	3.60	46.00	0.30	5.97	1.90	37.87
0.36	5.51	3.11	47.02	0.28	4.57	2.32	38.30
0.33	4.07	3.95	48.50	0.11	1.74	2.49	39.19
0.43	4.77	4.39	48.75	0.25	4.66	2.12	39.12
0.48	6.69	3.50	48.86	0.26	5.28	1.98	39.94
0.21	3.71	2.82	48.92	0.39	7.32	2.25	42.41
0.32	4.35	3.59	48.38	0.37	6.22	1.71	28.97
0.54	7.89	3.47	50.56	0.31	4.20	2.76	37.80
0.32	5.37	3.07	51.15	0.35	5.10	2.10	31.10
0.40	4.95	4.43	55.34	0.29	4.46	2.50	38.30
0.31	4.97	3.56	56.67	0.33	5.60	3.06	51.80
0.44	6.68	3.86	58.49	0.18	2.80	2.40	37.60

Look for gender differences testing for equality of groups means up at 5% l.o.s. by applying Mahalnobis's D^2 Statistic.

Aim: To carry out Mahalnobis's D^2 Statistic for the given data.

Procedure: Suppose

$$\pi_1 : \underset{\sim}{X}_{11}, \underset{\sim}{X}_{12}, \dots, \underset{\sim}{X}_{1n_1} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}_1, \Sigma)$$

$$\pi_2 : \underset{\sim}{X}_{21}, \underset{\sim}{X}_{22}, \dots, \underset{\sim}{X}_{2n_2} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}_2, \Sigma)$$

Now our problem is test $H_0 : \underset{\sim}{\boldsymbol{\mu}_1} = \underset{\sim}{\boldsymbol{\mu}_2}$ vs $H_1 : \underset{\sim}{\boldsymbol{\mu}_1} \neq \underset{\sim}{\boldsymbol{\mu}_2}$

The Mahalanobis D^2 Statistic is given by

$$D^2 = (\underset{\sim}{\bar{\mathbf{X}}}_{\mathbf{1}} - \underset{\sim}{\bar{\mathbf{X}}}_2)' S^{-1} (\underset{\sim}{\bar{\mathbf{X}}}_{\mathbf{1}} - \underset{\sim}{\bar{\mathbf{X}}}_2)$$

where

$$\underset{\sim}{\bar{\mathbf{x}}}_1 = \text{ Sample mean vector of } \underset{\sim}{X}_{11}, \underset{\sim}{X}_{12}, \ldots, \underset{\sim}{X}_{1n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} \underset{\sim}{X}_{1i}$$

$$\underset{\sim}{\bar{\mathbf{x}}}_2 = \text{ Sample mean vector of } \underset{\sim}{X}_{21}, \underset{\sim}{X}_{22}, \ldots, \underset{\sim}{X}_{2n_1} = \frac{1}{n_2}\sum_{i=1}^{n_2} \underset{\sim}{X}_{2i}$$

and the pooled sample dispersion (variance-covariance) matrix is given by

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 1}$$

$$\mathbf{S}_1 = \frac{1}{(n_1 - 1)}\sum_{i=1}^{m} (\underset{\sim}{\mathbf{x}}_{1i} - \underset{\sim}{\bar{\mathbf{x}}}_1)(\underset{\sim}{\mathbf{x}}_{1i} - \underset{\sim}{\bar{\mathbf{x}}}_1)'$$

$$\mathbf{S}_2 = \frac{1}{(n_2 - 1)}\sum_{i=1}^{n} (\underset{\sim}{\mathbf{x}}_{2i} - \underset{\sim}{\bar{\mathbf{x}}}_2)(\underset{\sim}{\mathbf{x}}_{2i} - \underset{\sim}{\bar{\mathbf{x}}}_2)'$$

$$D^2 > \frac{(n_1 + n_2)}{n_1 n_2}\frac{(n_1 + n_2 - 2)*p}{n_1 + n_2 - p - 1}, F_{p,\, n_1 + n_2 - p - 1}(\alpha)$$

R-CODE:

```
#R CODE FOR MAHALNOBIS D^2 STATISTIC FOR TESTING EQUALITY OF
TWO MVN POPULATION MEAN VECTORS.
d2.test=function(data){
nc=ncol(data);
x=subset(data[,-nc],data[,nc]==1);
y=subset(data[,-nc],data[,nc]==2);
cat("\n First Sample:\n"); print(x);
cat("\n Second Sample:\n"); print(y);
p=ncol(x);q=ncol(y);
m=nrow(x); n=nrow(y);
xbar=colMeans(x); ybar=colMeans(y);
sx=cov(x); sy=cov(y);
s=((m-1)*sx+(n-1)*sy)/(m+n-2);
cat("\n x-bar=[",xbar,"]");
cat("\n y-bar=[",ybar,"]");
cat("\n S-matrix:\n");
print(round(s,4));
dsq=t(xbar-ybar)%*%solve(s,xbar-ybar);    #mahalanobis d^2
test statistic.
d0sq=(m+n)/(m+n)*p/(m+n-p-1)*qf(0.95,p,m+n-p-1);    #critical
d^2 value.
cat("\n Mahalanobis D^2 statistic(distance)value=",dsq,"\n");
cat("\n Mahalanobis critical D^2 value (at 5%
los)=",d0sq,"\n");
if(dsq<=d0sq) {cat("Based on the given data we accept H0 at 5%
los \n");
cat("That is we conclude that the given two samples have been
drawn from MVN population. \n");}
if(dsq>d0sq) {cat("Based on the given data we reject H0 at 5%
los \n");
cat("That is we conclude that the given two samples have not
been drawn from MVN population. \n");
}
}
data=read.csv("MAHALANOBIS_TREADMILL.csv",header=T);
data=as.matrix(data);
#sink("Y24ST20024_mahalnobis");
d2.test(data);
#sink();
```

OUTPUT:

```
First Sample:
     X1   X2   X3    X4
[1,] 0.34 3.71 2.87 30.87
[2,] 0.39 5.08 3.38 43.85
[3,] 0.48 5.13 4.13 44.51
[4,] 0.31 3.95 3.60 46.00
[5,] 0.36 5.51 3.11 47.02
```

```
 [6,] 0.33 4.07 3.95 48.50
 [7,] 0.43 4.77 4.39 48.75
 [8,] 0.48 6.69 3.50 48.86
 [9,] 0.21 3.71 2.82 48.92
[10,] 0.32 4.35 3.59 48.38
[11,] 0.54 7.89 3.47 50.56
[12,] 0.32 5.37 3.07 51.15
[13,] 0.40 4.95 4.43 55.34
[14,] 0.31 4.97 3.56 56.67
[15,] 0.44 6.68 3.86 58.49
Second Sample:
       X1   X2   X3    X4
 [1,] 0.29 5.04 1.93 33.85
 [2,] 0.28 3.95 2.51 35.82
 [3,] 0.31 4.88 2.31 36.40
 [4,] 0.30 5.97 1.90 37.87
 [5,] 0.28 4.57 2.32 38.30
 [6,] 0.11 1.74 2.49 39.19
 [7,] 0.25 4.66 2.12 39.12
 [8,] 0.26 5.28 1.98 39.94
 [9,] 0.39 7.32 2.25 42.41
[10,] 0.37 6.22 1.71 28.97
[11,] 0.31 4.20 2.76 37.80
[12,] 0.35 5.10 2.10 31.10
[13,] 0.29 4.46 2.50 38.30
[14,] 0.33 5.60 3.06 51.80
[15,] 0.18 2.80 2.40 37.60

 x-bar=[ 0.3773333 5.122 3.582 48.52467 ]
 y-bar=[ 0.2866667 4.786 2.289333 37.898 ]
 S-matrix:
      X1      X2      X3      X4
X1 0.0061  0.0831  0.0079  0.0123
X2 0.0831  1.6212 -0.0526  1.8783
```

```
X3 0.0079 -0.0526  0.1878  1.2868
X4 0.0123  1.8783  1.2868 34.0008


 Mahalanobis D^2 statistic(distance)value= 9.463008
 Mahalanobis critical D^2 value (at 5% los)= 0.4413937
Based on the given data we reject H0 at 5% los
```

INFERENCE:

Mahalanobis D^2 statistic (distance)value= 9.463008

Mahalanobis critical D^2 value (at 5% los)= 0.4413937

Based on the given data we reject H0 at 5% los

That is, we conclude that the given two samples have not been drawn from MVN population.

LAB EXERCISE 5:

MANOVA

Problem:

Carry out Manova for GPA and GMAT data given in the following table Admission data for graduate school of business.

Π_1:Admit			Π_2:Not Admit			Π_3:Border line		
App No.	GPA (X_1)	GMAT (X_2)	App No.	GPA (X_1)	GMAT (X_2)	App No.	GPA (X_1)	GMAT (X_2)
1	2.96	596	21	2.54	446	36	2.86	494
2	3.14	473	22	2.43	425	37	2.85	496
3	3.22	472	23	2.29	460	38	3.14	425
4	3.59	527	24	2.36	531	39	3.28	371
5	3.69	515	25	2.57	542	40	3.89	447
6	3.46	693	26	2.35	416	41	3.15	313
7	3.13	626	27	2.60	412	42	3.50	412
8	3.19	663	28	2.51	458	43	3.00	485
9	3.63	447	29	2.36	389	44	2.80	444
10	3.65	598	30	2.98	482	45	3.13	430
11	3.30	563	31	2.66	420			
12	3.40	553	32	2.68	420			
13	3.55	588	33	2.48	533			
14	3.78	591	34	2.90	519			
15	3.44	692	35	2.63	504			
16	3.48	538						
17	3.90	552						
18	3.35	520						
19	3.39	555						
20	3.35	523						

Aim: To investigate whether the population means vectors are same using the random samples. And carryout MANOVA for the given data.

Procedure:

Suppose we have ‘g’ populations each is distributed multivariate normal with mean vector $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, ..., \underset{\sim}{\mu}_g$ respectively. Let us suppose that all populations have the same covariance

matrix ∑

Thus, we have ‘g’ population.

$\pi_1 \sim NP(\underset{\sim}{\mu}_1, \Sigma)$

$\pi_2 \sim NP(\underset{\sim}{\mu}_2, \Sigma)$

.

.

.

$\pi_g \sim NP(\underset{\sim}{\mu}_g, \Sigma)$

Now, we have a sample ‘g’ size n$_i$ from ith population π_i thus, we have ‘g’ sample form ‘g’ population follows:

$$\pi_1 : \underset{\sim}{X}_{11}, \underset{\sim}{X}_{12}, ..., \underset{\sim}{X}_{1n_1} \sim NP(\underset{\sim}{\mu}_1, \Sigma)$$
$$\pi_2 : \underset{\sim}{X}_{21}, \underset{\sim}{X}_{22}, ..., \underset{\sim}{X}_{2n_2} \sim NP(\underset{\sim}{\mu}_2, \Sigma)$$
$$.$$
$$.$$
$$.$$
$$\pi_g : \underset{\sim}{X}_{g1}, \underset{\sim}{X}_{g2}, ..., \underset{\sim}{X}_{gn_1} \sim NP(\underset{\sim}{\mu}_g, \Sigma)$$

Null hypothesis:

Under the Null Hypothesis we have $H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = ... = \underset{\sim}{\mu}_g$

Alternative hypothesis:

Under Alternative hypothesis we have $H_1 : \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2 \neq ... \neq \underset{\sim}{\mu}_g$

To test the hypothesis we have the test statistic is $\Lambda^* = \dfrac{|W|}{|B+W|}$

The quantity Λ^* is called Wilk’s lamda and related to likelihood ratio criterion.

where,

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}_i})(\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}_i})'$$

$$= (n_1 - 1)S_1 + (n_2 - 1)S_2 + ... + (n_g - 1)S_g$$

Where S

Where S_i is sample covariance matrix of the sample

$$B=\sum_{i=1}^{g} n_i(\underset{\sim}{X}_i-\overline{\underset{\sim}{X}})(\underset{\sim}{X}_i-\overline{\underset{\sim}{X}})'$$

B+W =(n-1)s pooled

W = Error sum of squares

S = Total sum of squares

Where ‘n’ is total number of sample pooled it’s the samples co-variance matrix of the pooled samples.

The exact distribution of Λ^* can be derived for the special cases as listed in the below.

$p \geq 1, g = 3$

$$\left[\sum_{i=1}^{g} n_i - g - 2\right]\left[\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right] \square\ F_{\left(2p,2\sum_{i=1}^{g} n_i - g - 2\right)}$$

R-CODE:

```
# R-CODE FOR MANOVA (FOR TESTING THE EQUALITY OF SEVERAL
MULTIVARIATE NORMAL SAMPLE MEAN VECTORS)
data=read.csv("manova_gmat.csv",header=T);
print(data);
#data=data[,-1];
MANOVA=function(data){
data=as.matrix(data);
nc=ncol(data);g=max(data[,nc]);
W=0;
for (i in 1:g) {
X=subset(data[,-nc],data[,nc]==i);
n=nrow(X);
A=(n-1)*cov(X);
W=W+A;}
print("W-matrix:");print(W);
X=data[,-nc];n=nrow(X);
BPW=(n-1)*cov(X);
print("B+W-matrix:");print(BPW);
lamda=sqrt(det(W)/det(BPW));
lamda=(1-lamda)/lamda*(n-g-1)/(g-1);
cat("lamda=",lamda,"\n");
Fcrit=qf(0.95,2*(g-1),2*n-g-1);# critical F value
cat("F-critical value=",Fcrit,"\n");
if (lamda<=Fcrit) cat(" The given samples have come from the
same population\n")
if(lamda>Fcrit) cat(" The given samples have come from
different populations\n");
}
MANOVA(data);
```

OUTPUT:-

```
> source("F:\\CHANDU MVA 2nd SEM\\MANOVA.txt")
   GPA GMAT SAMPLES
1  2.96  596      1
2  3.14  473      1
3  3.22  472      1
4  3.59  527      1
5  3.69  515      1
6  3.46  693      1
7  3.13  626      1
8  3.19  663      1
9  3.63  447      1
10 3.65  598      1
11 3.30  563      1
12 3.40  553      1
13 3.55  588      1
14 3.78  591      1
15 3.44  692      1
```

```
16 3.48  538     1
17 3.90  552     1
18 3.35  520     1
19 3.39  555     1
20 3.35  523     1
21 2.54  446     2
22 2.43  425     2
23 2.29  460     2
24 2.36  531     2
25 2.57  542     2
26 2.35  416     2
27 2.60  412     2
28 2.51  458     2
29 2.36  389     2
30 2.98  482     2
31 2.66  420     2
32 2.68  420     2
33 2.48  533     2
34 2.90  519     2
35 2.63  504     2
36 2.86  494     3
37 2.85  496     3
38 3.14  425     3
39 3.28  371     3
40 3.89  447     3
41 3.15  313     3
42 3.50  412     3
43 3.00  485     3
44 2.80  444     3
45 3.13  430     3
[1] "W-matrix:"
          GPA       GMAT
GPA    2.62336    -50.092
GMAT -50.09200 155134.250
```

```
[1] "B+W-matrix:"
        GPA       GMAT
GPA    9.25592    629.6287
GMAT 629.62867 303923.6444
lamda= 29.60971
F-critical value= 2.47774
The given samples have come from different populations
```

INFERENCE:

F-calculated value = 3.113445

F-critical value= 2.47774

Therefore, the given samples are not being drawn from the same population.

LAB EXERCISE 6:

Linear Discriminant Analysis for Two Multivariate Populations

Problem:

Jolicoueur and Mosimann studied the relationship of size and shape for painted turtles. The following table contains their measurements on carapaces of 23 female and 23 male turtles. Test whether the female turtles and male turtles have the same measurements with respect to carapaces by applying Hotelling's T^2at 5% l.o.s.

Female			**Male**		
Length (x_1)	Width (x_2)	Height (x_3)	Length (x_1)	Width (x_2)	Height (x_3)
98	81	40	93	74	37
103	84	38	96	78	39
105	86	45	101	84	39
119	88	44	102	85	38
123	92	50	103	81	37
103	100	46	104	83	39
133	99	51	111	102	39
133	102	51	107	82	40
133	102	51	112	89	40
134	100	48	113	88	40
136	102	49	114	86	40
145	98	65	116	90	43
138	99	51	117	90	41
141	105	53	117	99	41
147	108	57	119	93	41
149	107	55	120	89	45
153	107	56	120	93	44
160	115	63	121	95	42
155	117	65	125	93	45
148	115	62	127	96	45
159	118	63	128	103	45
162	124	61	131	95	46
175	132	67	135	106	47

Carry out the discriminant analysis to find the Fisher's linear discriminant.

Aim: To carry out the discriminant analysis to find the linear discriminant functions and using these functions allocate the new observations to an appropriate group.

Procedure: -

The linear discriminant for ith population is given by $\hat{d}_i(\underset{\sim}{x}) = -\frac{1}{2}\overline{\underset{\sim}{x}_i}' S^{-1}\overline{\underset{\sim}{x}} + \overline{\underset{\sim}{x}_i}' S^{-1}\underset{\sim}{x}$

$where\ \overline{\underset{\sim}{x}_i} = mean\ of\ the\ i^{th} sample$

$$S = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$$

$n = n_1 + n_2$

S_i = Sample dispersion matrix for ith sample

Allocation of new observation $\underset{\sim}{x}_0$:- using the above linear discriminant function we find linear discriminant scores for ith group $\hat{d}_i(\underset{\sim}{x}_0) = \overline{\underset{\sim}{x}_i}' S^{-1}\underset{\sim}{x}_0 - \frac{1}{2}\overline{\underset{\sim}{x}_i}' S^{-1}\overline{\underset{\sim}{x}_i}$; i=1,2

Allocate the new observation $\underset{\sim}{x}_0$ to that group for which the discriminant score is maximum.

R-CODE:

```
DA2=function(data){

X=data;X=as.matrix(X)

nc=ncol(X)

X1=subset(X[,-nc],X[,nc]==1)

X2=subset(X[,-nc],X[,nc]==2)

cat("\n Enter new observation values:\n")

xnew = matrix(scan(), nrow = 1)

p=ncol(X1)

n1=nrow(X1);x1bar=round(colMeans(X1),4);S1=cov(X1)

n2=nrow(X2);x2bar=round(colMeans(X2),4);S2=cov(X2)

S=((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)

cat("\n Discriminant analysis for allocating a new observation")

cat("\n Between two multivariate normal populations")

cat("\n Mean vectors of the given samples:\n")
cat(" X1bar=[",x1bar,"]\n")

cat(" X2bar=[",x2bar,"]\n")

cat(" New observation Xnew=[",xnew,"]\n")

cat("\n Sample variance covariance matrix
(pooled):\n");print(round(S,4))

w=round(solve(S,x1bar-x2bar),4)

d1=solve(S,x1bar)

d2=solve(S,x2bar)

k1=-sum(x1bar*solve(S,x1bar))/2

k2=-sum(x2bar*solve(S,x2bar))/2

cat("\n Linear discriminant functions:\n")

cat(" Y1=",k1)

for(i in 1:p) if(d1[i]>0) cat("+",d1[i],names(data)[i])

else cat("-",-d1[i],names(data)[i])

cat("\n\n")

cat(" Y2=",k2)

for(i in 1:p) if(d2[i]>0) cat("+",d2[i],names(data)[i])

else cat("-",-d2[i],names(data)[i])

cat("\n\n")

D1=k1+sum(xnew*d1)

D2=k2+sum(xnew*d2)

cat("\n Discriminant score of first population=",D1)

cat("\n Discriminant score of second population=",D2)

if(D1<D2) cat("\n\n Conclusion: New observation x0 is
allocated to second MVN population\n")
```

```
else cat("\n\n Conclusion: New observation x0 is allocated to
first MVN population\n")

}
```

OUT PUT:

```
Discriminant analysis for allocating a new observation

 Between two multivariate normal populations

 Mean vectors of the given samples:

 X1bar=[ 137.0435 103.5217 53.5217 ]

 X2bar=[ 114.4348 90.1739 41.4348 ]

 New observation Xnew=[ 0.45 5.75 4 40 ]

 Sample variance covariance matrix (pooled):

        x1       x2      x3

x1 279.1957 159.8804 93.9575

x2 159.8804 114.6146 54.5909

x3  93.9575  54.5909 38.6680

 Linear discriminant functions:

 Y1= -50.28727- 0.42671 x1+ 1.054256 x2+ 0.9325939 x3

 Y2= -36.92879- 0.2961669 x1+ 1.058552 x2+ 0.2967467 x3

 Discriminant score of first population= -57.75534

 Discriminant score of second population= -41.63508

 Conclusion: New observation x0 is allocated to second MVN population
```

INFERENCE:

Linear discriminant functions:

Y1= -50.28727- 0.42671 x1+ 1.054256 x2+ 0.9325939 x3

Y2= -36.92879- 0.2961669 x1+ 1.058552 x2+ 0.2967467 x3

Discriminant score of first population= -57.75534

Discriminant score of second population= -41.63508

New observation $\underset{\sim}{x}_0$ is allocated to first MVA Population.

LAB EXERCISE 7:

Linear discriminant Analysis for several multivariate populations

Problem:

In a diabetic centre the fasting blood sugar levels of three groups of patients are recorded two times, one before the treatment(X1) and another after the treatment(X2).

Group1:<40years		Group2:40-50Years		Group3:50+Years	
X_1	X_2	X_1	X_2	X_1	X_2
162	174	329	310	110	198
222	210	314	303	112	105
110	206	228	343	294	328
233	218	215	320	213	230
356	148	159	215	221	170
181	366	179	303	160	176
185	215	156	130	369	319
144	236	196	167	157	180
250	206	253	279	186	160
241	217	218	234	236	235

Carry out the discriminant analysis to find the linear discriminant functions (scores) and using these functions (scores) allocate the new observation X=[185 200]′ to an appropriate group.

Aim: -To carry out discriminant analysis for the 3 groups of diabetic patient to find the linear discriminant functions(scores) to 3 groups and to allocate the given new observation to an appropriate group(population).

Procedure:-

The simple linear discriminant function (score) for ith group (population) is given by $\hat{d}_i(\underset{\sim}{x}) = -\frac{1}{2}\bar{\underset{\sim}{x}}_i{}' S^{-1}\bar{\underset{\sim}{x}} + \bar{\underset{\sim}{x}}_i{}' S^{-1}\underset{\sim}{x}$

where $\underset{\sim}{x}_i$ = Mean of the ith sample

$$S = \frac{(n_1-1)S_1 + (n_2-1)S_2 + ... + (n_g-1)S_g}{n-g}$$

where $n = \sum_{i=1}^{g} n_i$

S_i = Sample dispersion matrix for ith sample

Allocation of new observation $\underset{\sim}{x}_0$:- using the above linear discriminant function we find linear discriminant scores for ith group $\hat{d}_i(\underset{\sim}{x}_0) = \underset{\sim}{\bar{x}}_i' S^{-1} \underset{\sim}{x}_0 - \frac{1}{2} \underset{\sim}{\bar{x}}_i' S^{-1} \underset{\sim}{\bar{x}}_i$; i=1,2,..

Allocate the new observation $\underset{\sim}{x}_0$ to that group for which the discriminant score is maximum.

Conclusion:-New observation $\underset{\sim}{x}_0$ is allocated to MVA population.

R-CODE: -

```
# R-CODE FOR DISCRIMINANT ANALYSIS FOR THE CASE OF SEVERAL
MULTIVARIATE
# NORMAL NORMAL POPULATIONS WITH EQUAL VARIANCE-COVARIANCE
MATRIX
data=read.csv("ldf_sugar_sugar_levels.csv",header=T);
nc=ncol(data);
#data=data[,-nc];
cat("\n ENTER NEW Observatioin x0:");
x0=scan();
DISCANA=function(data,x0){
k=array();D=array();
nc=ncol(data);
g=max(data[,nc]);
data1=data[,-nc]
data=as.matrix(data);
mean=matrix(,g,nc-1);
d=matrix(,g,nc-1);
S=0;N=0;
for (i in 1:g) {
X=subset(data[,-nc],data[,nc]==i);
mean[i,]=colMeans(X);
n=nrow(X);N=N+n
S=S+(n-1)*cov(X);
}
S=S/(N-g);
cat(" MEAN VECTORS OF THE GIVEN SAMPLES:\n");
for (i in 1:g) cat(" [",mean[i,],"]\n");
cat("\n New Observation X0= [",x0,"]\n");
cat("\nSample Dispersion matrix (pooled):\n");print(S);
cat("\n LINEAR DISCRIMINANT FUNCTIONS:\n");
cat("        const.    ");
for (i in 1:(nc-1)) cat(names(data1)[i],"      ");
cat("\n");
for (i in 1:g) {
d[i,]=solve(S,mean[i,]);
k[i]=-sum(mean[i,]*d[i,])/2;
cat("LDF",i,format(round(k[i],4),width=8,nsmall=2),format(roun
d(d[i,],4),width=8,nsmall=2),"\n");
D[i]=k[i]+sum(x0*d[i,]);
}
cat("\nDiscriminant scores of POPULATIONs:\n");
cat(round(D,4));
cat("\n\n CONCLUSION:\n");
cat(" New observation x0 is allocated to MVN
Population",which(D==max(D)),"\n")
}
DISCANA(data,x0)
```

OUTPUT:-

```
> source("C:\\Jilani\\DA_several_pop.R")

 ENTER NEW Observatioin x0:1: 185
2: 200
3:
Read 2 items
 MEAN VECTORS OF THE GIVEN SAMPLES:
 [ 208.4 219.6 ]
 [ 224.7 260.4 ]
 [ 205.8 219.1 ]
 New Observation X0= [ 185 200 ]
Sample Dispersion matrix (pooled):
        X1       X2
X1 4920.670 2050.778
X2 2050.778 4837.915
 LINEAR DISCRIMINANT FUNCTIONS:
       const.    X1       X2
LDF 1   -6.625   0.0285   0.0333
LDF 2  -8.6208   0.0282   0.0419
LDF 3  -6.5351   0.0279   0.0335
Discriminant scores of POPULATIONs:
5.3058 4.9721 5.316
New observation x0 is allocated to MVN Population 3
```

Inference:

Discriminant scores of POPULATIONs:

5.3058 4.9721 5.316

New observation x0 is allocated to MVN Population 3

LAB EXERCISE 8:

Fisher's Linear discriminant Analysis for several multivariate populations

Problem:

Bhuyan and Othman(1991) utilized the data on vertebrae of two fish specimens belong to families Serranidae(1),Carangidae(2) in the Dept of Zoology, Garyounis University, Libya. Only 12 observations of the variables are presented on each of the specimens of commercial fishes recorded The recorded data are

X_1=length of the centrum,

X_2=width of the centrum taken from anterior,

X_3=width of the centrum taken from posterior,

The variables are measured in mm.

SLNO	X1	X2	X3	FAMILY
1	7.5	6.7	6.5	1
2	6.8	6.2	6.1	1
3	8.5	7.1	6.8	1
4	5.8	6.0	6.3	1
5	5.2	5.8	5.4	1
6	7.0	7.2	5.9	1
7	8.2	7.5	7.0	1
8	6.9	7.3	6.7	1
9	7.4	6.8	6.6	1
10	8.4	7.3	6.7	1
11	7.6	7.0	6.7	1
12	9.2	7.8	6.9	1
13	9.4	6.3	6.5	2
14	9.2	6.0	6.2	2
15	8.7	6.1	6.0	2
16	7.5	5.2	6.7	2
17	8.2	6.6	6.3	2
18	7.2	5.3	5.8	2
19	6.7	5.8	6.9	2
20	7.2	5.9	6.0	2
21	7.7	6.3	7.1	2
22	6.7	5.2	5.9	2
23	9.4	7.2	6.9	2
24	8.1	6.9	7.0	2

Carry out the discriminant analysis to find the Fisher's linear discriminant and using this function allocate the new observation X0=[**7.8** **5.5** **4.8**]′ to an appropriate group.

Aim:-To carry out the discriminant analysis to find the Fisher's linear discriminant and using this allocate the new observation.

Procedure:

We have the Fisher's linear discriminant function

$y = \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}},$ where $\underset{\sim}{\mathbf{w}} = S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)$

(1)

Let 'm' be the midpoint between $\bar{y}_1$ and $\bar{\mathrm{y}}_2$ and is given by

$$m = (\bar{y}_1 + \bar{y}_2)/2 \qquad (2)$$
$$=1/2\ (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)'\mathrm{S}^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 + \underset{\sim}{\bar{\mathbf{x}}}_2)$$

Now , the allocation rule or classification rule based on Fisher's discriminant function is as follows:

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$,if

$y_0 = (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)'S^{-1}\underset{\sim}{\mathbf{x}}_0 \geq m$ or $\mathrm{y}_0 - m \geq 0$

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$,if

$y_0 < m$ or $\mathrm{y}_0 - m < 0$ (3)

R-CODE: -

```
# R-CODE FOR FINDING FISHER LINEAR DISCRIMINANT IN CASE OF TWO MULTIVARIATE
# POPULATIONS WITH EQUAL VARIANCE-COVARIANCE MATRIX
# DATA OF TWO SAMPLES SHOULD BE GIVEN IN THE SAME FILE
# R-CODE FOR FINDING FISHER LINEAR DISCRIMINANT IN CASE OF TWO MULTIVARIATE
# POPULATIONS WITH EQUAL VARIANCE-COVARIANCE MATRIX
# DATA OF TWO SAMPLES SHOULD BE GIVEN IN THE SAME FILE WITH SAMPLE CODES 1 & 2
# ENTER NEW OBSERVATION AS LAST OBSERVATION IN THE SAME FILE WITH SAMPLE CODE 3
FISHERLDF=function(data){
nc=ncol(data);
X1=subset(data[,-nc],data[,nc]==1);X2=subset(data[,-nc],data[,nc]==2);
xnew = matrix(scan(), nrow = 1)
p=ncol(X1);
n1=nrow(X1);x1bar=round(colMeans(X1),4);S1=cov(X1);
n2=nrow(X2);x2bar=round(colMeans(X2),4);S2=cov(X2);
S=((n1-1)*S1+(n2-1)*S2)/(n1+n2-2);
#sink("FISHERLDF-OUTPUT");
cat("\n FISHER LINEAR DISCRIMINANT ANALYSIS FOR ALLOCATING A
NEW OBSERVATION");
cat(" BETWEEN TWO MULTIVARIATE POPULATIONS");
cat("\n MEAN VECTORS OF THE GIVEN SAMPLES:\n");
cat("X1bar= [",x1bar,"]\n");
cat("X2bar= [",x2bar,"]\n");
cat("New Observation X0= [",xnew,"]\n");
cat("\nSample Variance-Covariance matrix
(pooled):\n");print(round(S,4));
w=round(solve(S,x1bar-x2bar),4); # Computes Inv(S)(x1bar-
x2bar)
cat("\n FISHER's LINEAR DISCRIMINANT FUNCTION:\n");
cat(" Y=");
for (i in 1:p) if (w[i]>0) cat("  +",w[i],"X",i) else cat("  -
",-w[i],"X",i);
cat("\n\n");
y1bar=sum(w*x1bar);
y2bar=sum(w*x2bar);
m=(y1bar+y2bar)/2;
y0=sum(w*xnew);
cat("\n y1bar=",y1bar);cat("\t y2bar=",y2bar);
cat("\n m-value=",m);cat("\t y0-value=",y0);
if (y0>m) cat("\n\n CONCLUSION: New observation x0 is
allocated to FIRST Multivariate  Population\n")
if(y0<=m) cat("\n\n CONCLUSION: New observation x0 is
allocated to  SECOND Multivariate  Population\n");
}
```

OUT PUT: -

```
data=read.csv("FISHDATA.csv", header=TRUE)
data
X1  X2  X3 X

1  7.5 6.7 6.5 1

2  6.8 6.2 6.1 1

3  8.5 7.1 6.8 1

4  5.8 6.0 6.3 1

5  5.2 5.8 5.4 1

6  7.0 7.2 5.9 1

7  8.2 7.5 7.0 1

8  6.9 7.3 6.7 1

9  7.4 6.8 6.6 1

10 8.4 7.3 6.7 1

11 7.6 7.0 6.7 1

12 9.2 7.8 6.9 1

13 9.4 6.3 6.5 2

14 8.7 6.0 6.2 2

15 7.5 6.1 6.0 2

16 8.2 5.2 6.7 2

17 7.2 6.6 6.3 2

18 7.7 5.3 5.8 2

19 6.7 5.8 6.9 2

20 7.2 5.9 6.0 2
```

```
21 7.7 6.3 7.1 2

22 6.7 5.2 5.9 2

23 9.4 7.2 6.9 2

24 8.1 6.9 7.0 2
```

```
FISHERLDF(data)
1: 7.8
2: 5.5
3: 4.8
4:
Read 3 items

 FISHER LINEAR DISCRIMINANT ANALYSIS FOR ALLOCATING A NEW
OBSERVATION BETWEEN TWO MULTIVARIATE POPULATIONS
 MEAN VECTORS OF THE GIVEN SAMPLES:
           X1bar= [ 7.375 6.8917 6.4667 ]
           X2bar= [ 7.875 6.0667 6.4417 ]
New Observation X0= [ 7.8 5.5 4.8 ]

Sample Variance-Covariance matrix (pooled):
      X1     X2     X3
X1 1.0757 0.4444 0.2842
X2 0.4444 0.3980 0.1883
X3 0.2842 0.1883 0.218

 FISHER's LINEAR DISCRIMINANT FUNCTION:
    Y=  - 2.2564 X 1  + 5.3221 X 2  - 1.5415 X 3

 y1bar= 10.06895        y2bar= 4.588554
 m-value= 7.328751       y0-value= 4.27243

 CONCLUSION: New observation x0 is allocated to  SECOND Multivariate  Population
```

Inference:-

The Fisher's linear discriminant function is

Y= - 2.2564 X 1 + 5.3221 X 2 - 1.5415 X 3

y1bar= 10.06895 y2bar= 4.588554

New Observation X0= [7.8 5.5 4.8]

The new observation is allocated to second MV population.

LAB EXERCISE 9:

Principle Component Analysis

Problem:

Jolicoueur and Mosimann studied the relationship of size and shape for painted turtles. The following table contains their measurements on carapaces of 23 female and 23 male turtles. Test whether the female turtles and male turtles have the same measurements with respect to carapaces.

Female			Male		
Length (x_1)	Width (x_2)	Height (x_3)	Length (x_1)	Width (x_2)	Height (x_3)
98	81	40	93	74	37
103	84	38	96	78	39
105	86	45	101	84	39
119	88	44	102	85	38
123	92	50	103	81	37
103	100	46	104	83	39
133	99	51	111	102	39
133	102	51	107	82	40
133	102	51	112	89	40
134	100	48	113	88	40
136	102	49	114	86	40
145	98	65	116	90	43
138	99	51	117	90	41
141	105	53	117	99	41
147	108	57	119	93	41
149	107	55	120	89	45
153	107	56	120	93	44
160	115	63	121	95	42
155	117	65	125	93	45
148	115	62	127	96	45
159	118	63	128	103	45
162	124	61	131	95	46
175	132	67	135	106	47

Carry out the principal component analysis for the turtle data and find the first two principal components.

Aim:-To carry out principal component analysis for the given data.

Procedure: -

From the given data we have to calculate the sample dispersion matrix. Now, we can compute the first PC Y_1 and its variance.

$$Y_1 = \underset{\sim}{w}_{11}\underset{\sim}{x}_1 + \underset{\sim}{w}_{12}\underset{\sim}{x}_2 + \ldots + \underset{\sim}{w}_{1p}\underset{\sim}{x}_p = \underset{\sim}{w}_1^{'}\underset{\sim}{x}$$

Where $\underset{\sim}{w}_1{}^{'}\underset{\sim}{w}_1 = 1$ $\quad and \quad Var(x_1) = \lambda_1$ (1)

From the following iterative equations

$$S_1\underset{\sim}{w}_1 = \lambda_1\underset{\sim}{w}_1 \quad where, S_1 = S \tag{2}$$

Equation (2) can be written as an iterative equation is given by

$$\lambda_1^{(i+1)}\underset{\sim}{w}^{(i+1)} = \underset{\sim}{\beta} = S_1\underset{\sim}{w}_1^{(i)}; i = 0,1,2,..... \tag{3}$$

From equation (3) we can compute

$$\lambda_1^{(i+1)} = \sqrt{\underset{\sim}{\beta}^{'}\underset{\sim}{\beta}} \quad and \quad \underset{\sim}{w}^{(i+1)} = \frac{\underset{\sim}{\beta}}{\lambda_1^{(i+1)}} \tag{4}$$

Now, the above equation (3) will be irritated with $\underset{\sim}{w}_1^{(0)} = \begin{bmatrix} 1 \\ 0 \\ : \\ : \\ 0 \end{bmatrix}_{PX1}$

Equation (3) will be solved iteratively until two successive values of λ_1 (computed using eq (4)) do agree upto 4 decimal places the corresponding $\underset{\sim}{w}$ is the first PC and its variance is λ_1

Computing 2nd PC: -

We have to replace the sample dispersion matrix S_1 with the adjusted dispersion matrix S_2 is given by

$$S_2 = S_1 - \lambda_1\underset{\sim}{w}_1\underset{\sim}{w}_1^{'} \tag{5}$$

Now, the 2nd PC can be computed in the same way as we computed the 1st PC by solving the following equations iteratively.

$$S_2\underset{\sim}{w}_2 = \lambda_2\underset{\sim}{w}_2 \tag{6}$$

Thus, the 2nd PC and it variance are given by

$$Y_2 = \underset{\sim}{w}_2^{'}\underset{\sim}{x}_2 \quad and \quad V(Y_2) = \lambda_2 \tag{7}$$

Computing 3rd PC: -

We have to replace the matrix S_2 with the adjusted matrix S_3 is given by

$$S_3 = S_2 - \lambda_2\underset{\sim}{w}_2\underset{\sim}{w}_2^{'} \tag{8}$$

Now, the 3rd PC can be computed in the same way as computed the 2nd PC by solving the following equations iteratively.

$$S_3 \underset{\sim}{w}_3 = \lambda_3 \underset{\sim}{w}_3 \qquad (9)$$

Thus, the 3rd PC and its variance are given below

$$Y_3 = \underset{\sim}{w}_3' \underset{\sim}{x} \quad and \quad V(Y_3) = \lambda_3 \qquad (10)$$

Similarly, one can compute the remaining PC's iteratively.

R-CODE:

```
# R-CODE FOR PRINCIPAL COMPONENT ANALYSIS BASED ON
# SAMPLE VARIANCE-COVARIANCE MATRIX OF MULTIVRIATE POPULATION
 PCA=function(data,k){
 vnames=names(data)
 data=as.matrix(data);
 p=ncol(data);
 S=cov(data);S1=S;
 cat("\nPC ANALYSIS BASED ON SAMPLE VARIANCE-COVARIANCE MATRIX:\n");
 cat("\nSample Variance-Covariance matrix based on the given data:\n");print(round(S,2));
 totvar=0;
 for (i in 1:p) totvar=totvar+S[i,i];
 cat("\n Total variance of the original components=",round(totvar,2),"\n\n");
 for (i in 1:k) {
 x=rep(0,p);x[i]=1;
 lamdaold=0;lamdanew=1;
 while(abs(lamdaold-lamdanew)>0.00001) {
 lamdaold=lamdanew;
 x=S%*%x;
 lamdanew=sqrt(sum(x*x));
 x=x/lamdanew;}
 cat("PC",i,": ");
 for (j in 1:p) {if (x[j]>0) cat("+",round(x[j],2)) else cat(round(x[j],2));
 cat(" ",vnames[j]," ");}
 cat("\nVar(PC",i,")=",round(lamdanew,2));
 cat("\n% of total
variation=",round(lamdanew/totvar*100,2),"\n\n");
 S=S-lamdanew*x%*%t(x);
 }
 }
 data=read.table("PCA.CSV",header=T);
 data=data[,-1];
 nc=ncol(data);
 data=data[,-nc];
 print(data);
 cat("\n ENTER No of PCs required:");k=scan();
 S=PCA(data,k);
```

OUT PUT:

```
> source("C:\\Users\\Jilani\\Desktop\\MVA\\jilani MVA 2nd SEM\\pca.R")

   LENGTH WIDTH HEIGHT
1     98    81     38
2    103    84     38
3    115    90     42
4    109    88     45
5    123    92     50
6    113    95     46
7    133   109     53
8    133   102     51
9    133   102     51
10   134    95     48
11   130   102     55
12   138    95     51
13    93    74     37
14    96    78     35
15   101    84     45
16   112    75     38
```

```
17   103   81   37
18   104   83   39
19   116   83   39
20   107   82   49
21   112   95   40
22   113   88   40
23   104   86   40
24   116   90   50

ENTER No of PCs required:1: 2
2:
Read 1 item
PC ANALYSIS BASED ON SAMPLE VARIANCE-COVARIANCE MATRIX:
Sample Variance-Covariance matrix based on the given data:
     LENGTH  WIDTH HEIGHT
LENGTH 179.16 102.97  66.17
WIDTH  102.97  83.21  45.13
HEIGHT  66.17  45.13  37.26
Total variance of the original components= 299.63
PC[ 1 ]: [ 0.8 0.51 0.32 ]
Var(PC[ 1 ])= 272.08
 %of total variation= 90.8
PC[ 2 ]: [ -0.59 0.76 0.26 ]
Var(PC[ 2 ])= 18.26
 %of total variation= 6.09
```

Inference:-

Total variance of the original components = 299.63

PCA of measurements of 15 trucks for: -

PC[1]: [0.8 0.51 0.32]

Var(PC[1])= 272.08

%of total variation= 90.8

PC[2]: [-0.59 0.76 0.26]

Var(PC[2])= 18.26

%of total variation= 6.09

LAB EXERCISE 1:

Power Curve for Most Powerful Test

Problem: Draw a power curve for the most powerful test the sample size 10

i) $H_0 : \mu = 4$ vs $H_1 : \mu > 4$

ii) $H_1 : \mu = 4$ vs $H_1 : \mu < 4$

μ, σ whereis the mean of normal population having σ = 2 , with level of significance 5%.

Aim:

To draw a power curve for the most powerful test based on the sample size 10

Procedure :-

$$\square_0 = (c - \mu) * \sqrt{n} / \sigma$$

$$c = \mu_0 + 1.645 * 2 / \sqrt{n}$$

Let, the problem of testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_1 : \theta = \theta_1$.

The critical region is the most powerful critical region of size for α testing

$H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$

$$P\{x \in \omega / H_0\} = \int_{\omega} L_0 dx = \alpha \qquad (1)$$

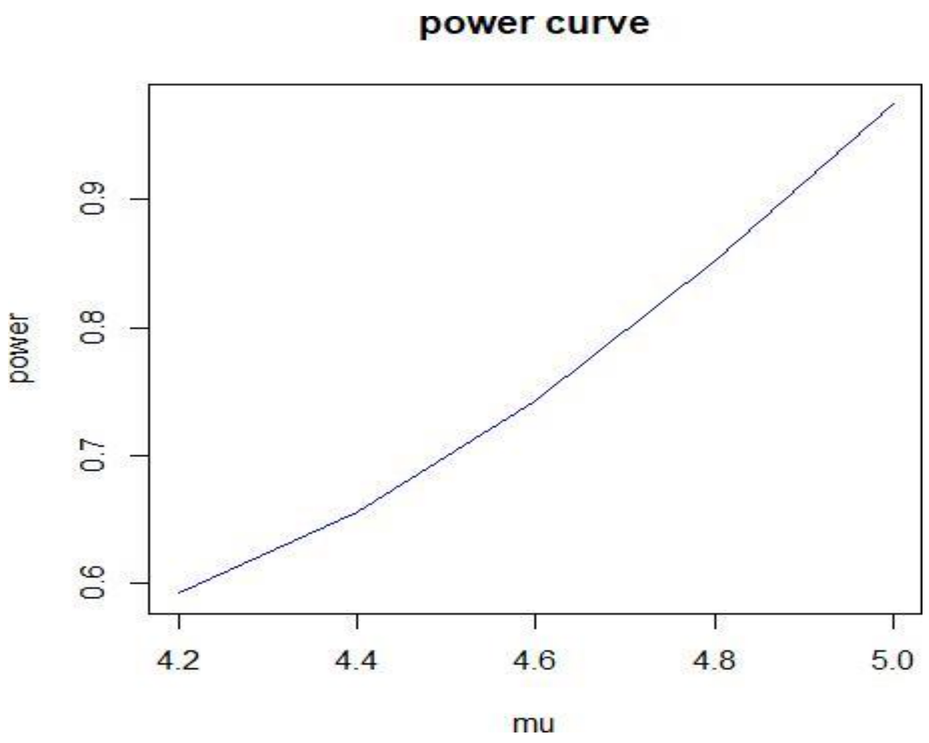
$$P\{x \in \omega / H_1\} \geq P\{x \in \omega_1 / H_1\}$$

for every other critical region , satisfying equation (1). The corresponding that is called as Most Powerful Test.

i) **R-CODE:**

```
mu0=4
n=10
s=2
c=mu0+1.645*2/sqrt(10)
cat("\n c \n",c)
mu=seq(4.2,5,by=0.2)
```

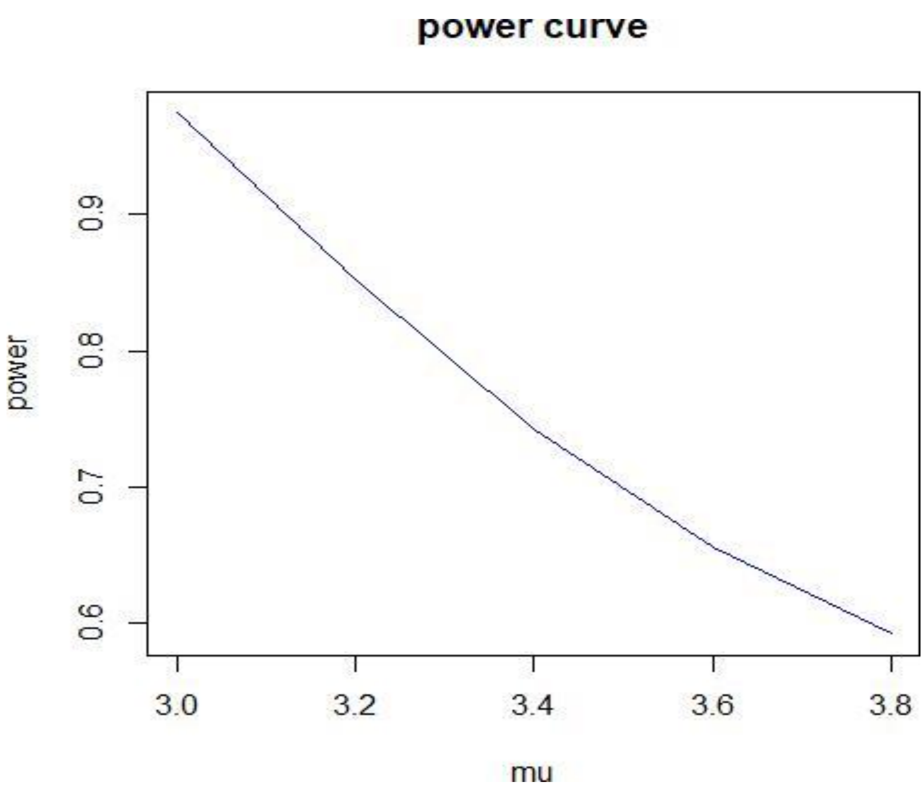
```
z0=(c-mu)*sqrt(n)/s
cat("\n z0 \n",z0)
z=(-0.5+pnorm(z0))
power=1-z
cat("\n power \n",power)
plot(mu,power,type="l",col="blue",lwd=0.01,xlab="mu",ylab="p
ower",main="power curve")
```



ii) **R-CODE: - 0**

```
mu0=4
n=10
s=2
c=mu0-1.645*2/sqrt(10)
cat("\n c \n",c)
mu=seq(3.8,3,by=-0.2)
z0=(c-mu)*sqrt(n)/s
cat("\n z0 \n",z0)
z=(0.5+pnorm(z0))
power=z
```

```
cat("\n power \n",power)
plot(mu,power,type="l",col="blue",lwd=0.01,xlab="mu",ylab="power",main="power curve")
```



Inference:

i) Output:

$$c = 5.040389$$

$$\square_0 = 1.3287772 \quad 1.012544 \quad 0.6963167 \quad 0.3800889 \quad 0.06386117$$

$$power = 0.59191616 \quad 0.65556389 \quad 0.7431153 \quad 0.8519397 \quad 0.9745404$$

ii) Output :

$$c = 2.959611$$

$$\square_0 = -1.3287772 \quad -1.012544 \quad -0.6963167 \quad -0.3800889 \quad -0.06386117$$

$$power = 0.59191616 \quad 0.65556389 \quad 0.7431153 \quad 0.8519397 \quad 0.9745404$$

LAB EXERCISE 2:

Power Curve

Problem:

Power analysis for a large sample, hypothesis test where the test statistics has no approximately normal distribution

Aim:

To draw a power curve for paver analysis for a large sample hypothesis test where the test statistics has an approximately normal distribution.

Procedure:-

Let, the problem of testing $H_0 : \theta = \theta_0$ a simple null hypothesis against a simple alternative $H_1 : \theta = \theta_1$ hypothesis.

The critical region is the most powerful critical region of α size for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$

$$P\{x \in \omega / H_0\} = \int_{\omega} L_0 \, dx = \alpha \qquad (1)$$

$$P\{x \in \omega / H_1\} \geq P\{x \in \omega_1 / H_1\} \qquad (2)$$

for every other critical region , satisfying equation (1).

$$\square_0 = (c - \mu) * \sqrt{n} / \sigma$$

$$c = \mu_0 + 1.645 * 2 / \sqrt{n}$$

R-CODE: -

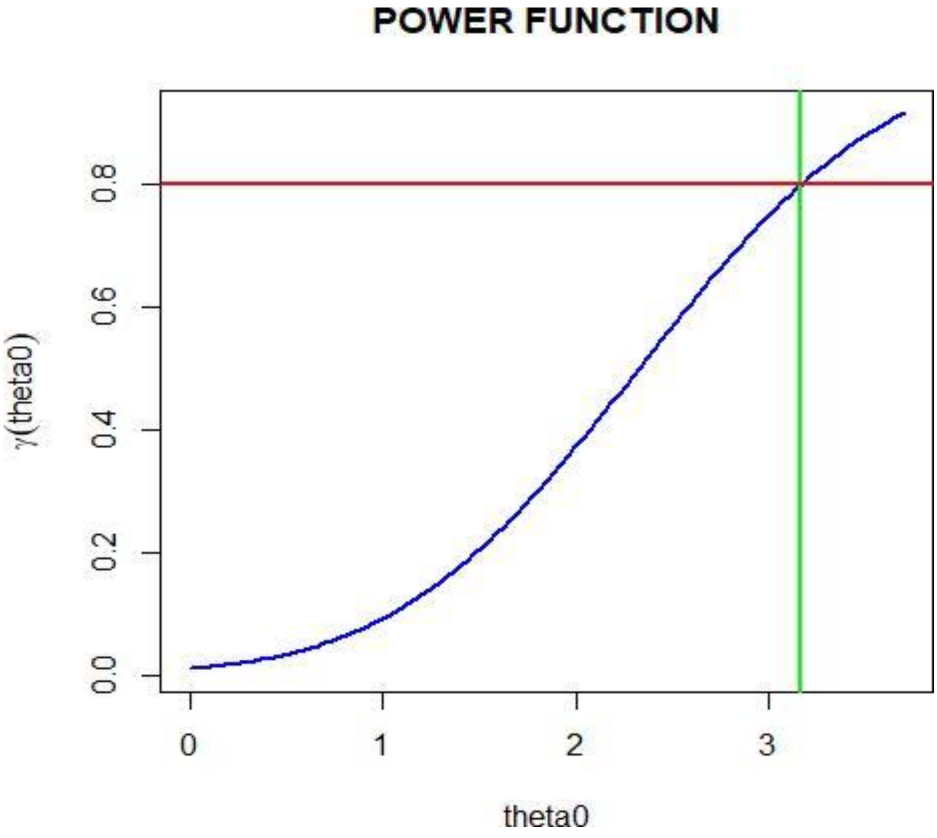
```
sigma=5
n=25
theta0=0
power=0.80
alpha=0.01
beta=1-power
z.alpha=qnorm(1-alpha)
cat("\n z.alpha:\n",z.alpha)
```

```
z.beta=qnorm(1-beta)
cat("\n z.beta: \n",z.beta)
curve(pnorm(sqrt(n)*(x-theta0)/sigma-z.alpha),
from=theta0,to=theta0+3.7*sigma/sqrt(n),
col="blue", main="POWER FUNCTION",
xlab=expression(theta0),
ylab=expression(gamma(theta0)), Iwd=2)
abline(v=theta0+(z.alpha+z.beta)*sigma/sqrt(n),
col="green", lwd=2)
abline(h=power,col="red", lwd=2)
```

Inference:-

$\square_{\alpha} = 2.326348$

$\square_{\beta} = 0.8416212$



LAB EXERCISE 3:

Kolmogrov-Smirnov Test

Problem:-

i) Let's say you have a sample data set and you want to test if it comes from a normal distribution with mean 0 and standard deviation 1.

Aim:-

To test a sample data set and want to test if it comes from a normal distribution with mean 0 and standard deviation 1.

Procedure:-

You can use ks.test() to compare your sample data against the normal distribution.

Construct the K-S test statistics.

$$ks\,test = \max\{x : | f_x(x) - F_x(x) |\}$$

R-CODE:

```
sample data=rnorm(100)
ks_test_result= ks.test(sample_data,"pnorm",mean=0,sd=1)
print(ks_test_result)
```

OUT PUT:

```
Asymptotic one-sample Kolmogorov-Smirnov test
data: sample_data
D = 0.072, p-value = 0.6777
alternative hypothesis: two-sided
```

Inference:

The ks test statistic (D) is 0.053159, and the p-value is 0.9401. Since the p-value is greater than 0.05, there is not enough evidence to reject the null hypothesis. So we might conclude that the sample data could come from a normal distribution with mean 0 and standard deviation 1.

ii) You have a sample of data set and you want to test if it follows a normal distribution by using ks_test.

Aim:

To test a sample data and want to test if it follows a normal distribution using ks_test.

Procedure:-

You can use ks.test() to compare your sample data against the normal distribution.

Construct the K-S test statistics.

$$ks\,test = \max\{x : | f_x(x) - F_x(x) |\}$$

R-CODE: -

```
sample_data=rnorm(500,mean=0,sd=1)
ks_test=ks.test(sample_data,"pnorm",mean=mean(sample_data),sd=
sd(sample_data))
print(ks_test)
```

OUT PUT:

```
Asymptotic one-sample Kolmogorov-Smirnov test
data: sample_data
D = 0.030957, p-value = 0.7241
alternative hypothesis: two-sided
```

Inference:

The ks test statistic (D) is 0.030957, and the p-value is 0.7241. Since the p-value is greater than 0.05, there is not enough evidence to reject the null hypothesis. So we might conclude that the sample data could come from a normal distribution with mean 0 and standard deviation 1.

iii) You have a sample of data set and you want to test if it follows a normal distribution by using ks_test.

Aim:-

To test a sample data and want to test if it follows a normal distribution using ks_test.

Procedure:

You can use ks.test() to compare your sample data against the normal distribution.

Construct the K-S test statistics.

$$ks\,test = \max\{x : | f_x(x) - F_x(x) |\}$$

R-CODE:

```
data1=rnorm(100); data2=rnorm(100)
print(ks.test(data1, data2))
```

OUT PUT:

```
Asymptotic two-sample Kolmogorov-Smirnov test
data: data1 and data2
D = 0.1, p-value = 0.6994
alternative hypothesis: two-sided
```

Inference:

The ks test statistic (D) is 0.1, and the p-value is 0.6994. Since the p-value is greater than 0.05, there is not enough evidence to reject the null hypothesis. So we might conclude that the sample data could come from a normal distribution with mean 0 and standard deviation 1.

LAB EXERCISE 4(a):

Kruskal Walli's Test

Problem:

Assume you have a study whether you measure the effectiveness of 3 different groups. To test the measure of effectiveness of 3 different groups (groups 1. group 2, group 3) on the same subjects by using Kruskal-Walli's test.

S.no	Group1	Group2	Group3
1	23	22	20
2	45	44	42
3	67	62	64
4	34	33	32
5	56	55	54
6	78	77	76
7	23	22	20
8	45	44	42
9	67	66	64
10	34	33	32

Aim: To test a study whether you measure of effectiveness of 3 different groups (groups 1. group 2, group 3) on the same subjects by using Kruskal-Walli's test.

Procedure:

Kruskal Walli's test statistic is defined on

$$H = \frac{12}{N(N-1)} \sum_{i=1}^{k} \frac{1}{n_i} \left[R_i - \frac{n_i(N+1)}{2} \right]^2$$

R_i = The expected sum of ranks which is The expected sum of ranks which is $\mathrm{n_i}\left(N+1\right)$

$\mathrm{n_i}$ =Total number of observations in the i^{th} sample

K : Number of groups

N :Total number of observations across all groups

In order to perform Kruskal – Walli's test we use Kruskal.test()

R-CODE:

```
group1=c(23,45,67,34,56,78,23,45,67,34)
group2=c(22,44,66,33,55,77,22,44,66,33)
group3=c(20,42,64,32,54,76,20,42,64,32)
data=data.frame(group1,group2,group3) print(data)
kruskall_test_result=kruskal.test(data)
print(kruskall_test_result)
```

OUT PUT:

```
    group1 group2 group3
1    23     22     20
2    45     44     42
3    67     66     64
4    34     33    32
5   56      55     54
6   78      77     76
7   23      22     20
8   45      44     42
9   67      66     64
10   34     33     32
Kruskal-Walli’s rank sum test data: data
Kruskal-Walli’s chi-squared = 0.83837, df = 2, p-value = 0.6576
```

Inference:

The test statistic Chi-squared value=0.83837 and p-value=0.6567. As p-value is greater than 0.05, Then we conclude that we accept the null hypothesis i.e., there is no significant difference between the 3 groups.

LAB EXERCISE 4(b):

Friedman's Two WayAnalysis of Variance by Ranks

Problem:

You have a study where you measure the effectiveness of 3 different treatments (A,B,C) on the same subjects. Each subject receives all the three treatments at different times and you want to compare the effectiveness of 3 treatments.

Aim:

To study the effectiveness of 3 different treatments (A, B, C) on the same subjects. Each subject receives all the 3 treatments at different times and you want to compare the effectiveness of 3 treatments.

Procedure:-

Ranks by data:- For each subject, ranks the treatments assigned ranks to the treatments, such that the lowest value get rank 1, the second lowest rank 2 and so on. If these are ties, assign the average rank to the tied values.

Sum of the ranks for each treatment:-For each treatment, sum of ranks across all subjects.

Calculate the mean rank for each treatment:-Compute the mean rank for each treatment.

Sum of squares of treatment ranks:-For each treatment, calculate the squared difference between its mean rank & overall rank, then sum this squared difference.

Friedman's test statistic is given below

$$Q = \frac{12}{nk(k+1)}\left[\sum_{i=1}^{k} R_i^2 - 3n(k+1)\right]$$

R-CODE: -

```
subject=factor(rep(1:5,each=3))
treatment=factor(rep(1:3,times=5))
response=c(10,20,30,15,25,35,12,22,32,18,28,38,16,26,36)
data=data.frame(subject, treatment, response)
data
data$rank=ave(data$response, data$subject,FUN=rank)
data
sum_ranks=aggregate(rank~treatment, data=data, sum)
sum_ranks
```

```
n=length(unique(data$subject))
K=length(unique(data$treatment))
R_j=sum_ranks$rank
Q=(12/(n*K*(K+1)))*sum(R_j^2)-3*n*(K+1)
df=K-1
P_value=pchisq(Q,df,lower.tail=FALSE)
cat("Friedman Test statistic(Q):",Q,"\n")
cat("P_value:",P_value, "\n")
```

OUT PUT: -

```
Friedman Teststatistic(Q): 10
P_value: 0.006737947
```

Inference:

Friedman chi-squared test statistic:

The test statistic used to determine if there are differences between the treatments.

P-value:

Friedman test statistic is 10 and the p-value is 0.006737947. This value is less than 0.05 it suggests that atleast one of the treatments differ significance from the others.

LAB EXERCISE 5(a):

Chi-Square Test for Homogeneity of Correlation Coefficient

Problem:

Compare the correlation coefficient for 2 groups using Fisher's Z-transformation and chi-square test.

GROUP1		GROUP2	
X1	X2	X1	X2
10	27	25	15
6	23	28	30
15	64	36	29
8	42	45	29
11	30	15	36
34	79	49	64
28	26	48	30
70	24	54	68
43	54	34	56
30	30	29	32
25	14	35	21

Aim:

To compare the correlation coefficient for the two groups using Fisher's Z-transformation and chi-square test.

Procedure:

Calculate the Pearson rank correlation coefficient for each group or sample.

The sampling distribution of correlation coefficient is not normally distributed Fisher's Z-transformationis applied to each correlation coefficient. This transformation convert the correlation coefficient into normal distribution.

$$\square = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

where,

r :- sample correlation coefficient.

The test statistic is calculated using the transformed coefficient

$$\chi^2 = \frac{(z_1 - z_2)^2}{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

where,

z_1 & z_2 :Fisher's Z-transformation correlation coefficient.

n_1 & n_2 :Sample sizes of two groups.

Conclusion:

The calculated chi square statistic is compared to a value from the chi square distribution table with d.f. equals to 1. If the statistic is greater than the critical value, you reject the null hypothesis.

R-CODE: -

```
data=read.csv("pair.csv", header=T)
group1=subset(data[,1:2])
group2=subset(data[,3:4])
r1=cor(group1$X1,group1$X2);
r2=cor(group2$Y1,group2$Y2);
cat("Correlation for group1:",r1)
cat("\nCorrelation for group2:",r2)
n1=nrow(group1)
n2=nrow(group2)
z1=0.5*log((1+r1)/(1-r1))
z2=0.5*log((1+r2)/(1-r2))
cat ("\n Fisher z transformation for group1:″, z1,"\n")
cat ("\n Fisher z transformation for group2:", z2,"\n")
K1=1/(n1-3)
K2=1/(n2-3)
chisqcal=(z1-z2)^2/(K1+K2)
cat("chisquarecal=",chisqcal,"\n")
p_value=pchisq(chisqcal,df=1,lower.tail=F)
cat("p-value=",p_value, "\n")
```

```
if(p_value<0.05) {
cat("Reject null hypothesis: Correlation are not homogeneous\n")}
if(p_value>0.05) {
cat("Accept null hypothesis: Correlation are homogeneous\n")}
```

OUT PUT: -

```
Correlation for group1: 0.04902549
Correlation for group2: 0.5259224
Fisherztransformation for group1: 0.04906483
Fisher z transformation for group2: 0.5844917
chisquarecal= 1.146728
p-value= 0.2842352
Accept null hypothesis: Correlation are homogeneous
```

Inference:

Chi-squared value=1.146728 and p-value=0.2842352

As p-value greater than 0.05 then we accept the null hypothesis i.e., correlation coefficients are homogeneous.

LABEXERCISE 5(b):

Chi-Square Test for Homogeneity of Correlation Coefficient

Problem:

If the correlation coefficient between sepal length &sepal width the same for different species of iris.

Aim:

To test whether the correlation between sepal length and sepal width is same for different species of iris.

Procedure:-

Calculate the Pearson rank correlation coefficient for each group or sample.

The sampling distribution of correlation coefficient is not normally distributed Fisher's Z-transformation is applied to each correlation coefficient. This transformation converts the correlation coefficient into normal distribution.

$$\square = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

where,

r :- sample correlation coefficient.

The test statistic is calculated using the transformed coefficient

$$\chi^2 = \frac{\left[(n_1-3)*\left(z_1-\left(\frac{z_2+z_3}{2}\right)\right)\right]^2}{\frac{1}{n_1-3}+\frac{1}{n_2-3}+\frac{1}{n_3-3}}$$

where,

$z_1 , z_2 \,\&\, z_3$:Fisher's Z-transformation correlation coefficient.

$n_1 , n_2 \,\&\, n_3$:Sample sizes of two groups.

Conclusion:

The calculated chi square statistic is compared to a value from the chi square distribution table with d.f. equals to 1. If the statistic is greater than the critical value, you reject the null hypothesis.

R-CODE:

```
data(iris)
setosa=iris[iris$Species=="setosa",]
versicolor=iris[iris$Species=="versicolor",]
virginica-iris[iris$Species=="virginica",]
r_setosa=cor(setosa$Sepal.Length,setosa$Sepal.Width)
cat(" R_SETOSA:",r_setosa)
r_versicolor=cor(versicolor$Sepal.Length,
versicolor$Sepal.Width)
cat("\n R_VERSICOLOR:",r_versicolor)
r_virginica=cor(virginica$Sepal.Length, virginica$Sepal.Width)
cat("\n R_VIRGINICA:",r_virginica)
n_setosa=nrow(setosa)
cat("\n N_SETOSA:",n_setosa)
n_versicolor=nrow(versicolor)
cat("\n N_VERSICOLOR:",n_versicolor)
n_virginica=nrow(virginica)
cat("\n N_VIRGINICA:",n_virginica)
z_setosa=0.5*log((1+r_setosa)/(1-r_setosa))
cat("\n Z_SETOSA:",z_setosa)
z_versicolor=0.5*log((1+r_versicolor)/(1-r_versicolor))
cat("\n Z_VERSICOLOR:",z_versicolor)
z_virginica=0.5*log((1+r_virginica)/(1-r_virginica))
cat("\n Z_VIRGINICA:",z_virginica)
chi_square=((n_setosa-3)*(z_setosa-
(z_versicolor+z_virginica)/2)^2/(1/(n_setosa-
3)+1/(n_versicolor-3)+1/(n_virginica-3)))
p_value=pchisq(chi_square,df=2,lower.tail=FALSE)
cat("\n Chi-square statistic:",chi_square, "\n")
cat(" p-value:",p_value, "\n")
```

OUT PUT:

```
R_SETOSA: 0.7425467
R_VERSICOLOR: 0.5259107
R_VIRGINICA: 0.4572278
N_SETOSA: 50
N_VERSICOLOR: 50
N_VIRGINICA: 50
Z_SETOSA: 0.9561323
Z_VERSICOLOR: 0.5844755
Z_VIRGINICA: 0.4938007
Chi-square statistic: 128.0367
p-value: 1.574675e-28
```

Inference:

Chi-squared value=128.0367 and p-value=1.574675e-28.As p-value greater than 0.05 then we accept the null hypothesis i.e., correlation coefficients are homogeneous.

LABEXERCISE 5(c):

Chi-Square Test for Homogeneity of Correlation Coefficient

Problem:

Is the correlation between Miles per gallon and displacement the same for cars with different number of cylinders?

Aim:-

To test whether the correlation between miles per gallon and displacement the same for class with different no. of cylinders.

Procedure:-

Calculate the Pearson rank correlation coefficient for each group (or) sample.

The sampling distribution of correlation coefficient is not normally distributed. Fishers z-transformation is applied to each correlation coefficient. This transformation converts the correlation coefficient into a form that approximately follow a normal distribution.

$$\square = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

where,

r = sample correlation coefficient

The test statistic is calculated using the transformed coefficients.

$$\chi^2 = \frac{\left[(n_1-3)*\left(z_1-\left(\frac{z_2+z_3}{2}\right)\right)\right]^2}{\frac{1}{n_1-3}+\frac{1}{n_2-3}+\frac{1}{n_3-3}}$$

where,

$z_1, z_2 \,\&\, z_3$ = Fishers z-transformation correlation coefficient

$n_1, n_2 \,\&\, n_3$ = Sample Sizes of the 3 groups

Conclusion:-

The calculated Chi-squared statistic is compared to a critical value of x-distribution table with d.f. equal to 1

If the statistic is greater than the critical value, you reject the null hypothesis.

R-CODE: -

```
data(mtcars)
cyl4=mtcars[mtcars$cyl==4,]
cyl6=mtcars[mtcars$cyl==6,]
cyl8=mtcars[mtcars$cyl==8,]
r4=cor(cyl4$mpg,cyl4$disp)
cat("\n r4:",r4)
r6=cor(cyl6$mpg,cyl6$disp)
cat("\n r6:",r6)
r8=cor(cyl8$mpg,cyl8$disp)
cat("\n r8:",r8)
n4=nrow(cyl4)
cat("\n n4:",n4)
n6=nrow(cyl6)
cat("\n n6:",n6)
n8=nrow(cyl8)
cat("\n n8:",n8)
z4=0.5*log((1+r4)/(1-r4))
cat("\n z4:",z4)
z6=0.5*log((1+r6)/(1-r6))
cat("\n z6:",z6)
z8=0.5*log((1+r8)/(1-r8))
cat("\n z8:",z8)
chi_square=((n4-3)*(z4-(z6+z8)/2)^2/(1/(n4-3)+1/(n6-3)+1/(n8-
3)))
p_value=pchisq(chi_square,df = 2,lower.tail=FALSE)
cat("\n Chi-Square Statistic:",chi_square,"\n")
cat(" P-Value:",p_value,"\n")
```

OUTPUT: -

```
r4: -0.8052361
r6: 0.1030827
r8: -0.519767
```

```
n4: 11
n6: 7
n8: 14
z4: -1.113329
z6: 0.1034502
z8: -0.5760205
Chi-Square Statistic: 13.20783
P-Value: 0.001355049
```

Inference:

The Chi-squared statistic is 13.20783 &p-value is 0.001355049, which is less than the significance level of 0.05. This indicates that the correlation between milesper gallon and displacement is significantly differ for different number of cylinders.

LABEXERCISE 6(a):

F-Test to check the Homogeneity of Regression Coefficients

Problem:

A company wants to investigate if the relationship between the amount of money spent on advertising (x) and the sales (y) is the same for different regions. The company has data from these 3 regions: North, South & East

Aim:-

To investigate if the relationship between the amount of money spent on advertising (x) and sales (y) is the same for the three regions North, South & East.

Procedure:-

Step 1:Formulate the null & alternative hypothesis.

$H_0 : \beta_1 = \beta_2 = ... = \beta_k$ (regression coefficients are equal acrossgroups)

$H_1 : Not\, all\, \beta_i$ (regression coefficients are not equal across groups)

Step2:- Estimate the regression models.

Estimate the restricted model (assuming equal steps):

$-y = \beta_0 + \beta^1 x + \xi$ (single regression line of all groups).

Estimate the unrestricted model (allowing different slopes).

$-y = \beta_0 + \beta^1 x + \beta^2 x + ... + \beta^K x + \xi$ (separate regression lines for each group).

Step3:- Calculate the Test statistic

Calculate the residual sum of squares for the restricted model (SSE-R)

Calculate the residual sum of squares for the unrestricted model (SSE-U)

Calculate the degrees of freedom for the restricted model(df-R) and unrestricted model(df-U)

Calculate the F_statistic:

$$-F = \frac{(SSE - R) - (SSE - U)}{(df - R - df - U)}$$

Step4:- Determine the critical region and p-value:

Choose a significance level(ex:α=0.05)

Determine the critical value from the F-distribution table (or) using software

Calculate the p-value associated with the F-statistic.

Step 5:- Make a decision:

If the p-value is less than the level of significance(α), reject H₀ and conclude that the regression coefficients are not equal across groups.

If the p-value is greater than the significance level(α)fail to reject H₀ and conclude that the regression coefficients are equal across groups.

R-CODE: -

```
### Step 0: Create the data

region = c(rep("North", 10), rep("South", 10), rep("East", 
10))

advertising = c(10,20,30,40,50,60,70,80,90,100,

                 15,25,35,45,55,65,75,85,95,105,

                 20,30,40,50,60,70,80,90,100,110)

sales = c(100,120,140,160,180,200,220,240,260,280,

          110,130,150,170,190,210,230,250,270,290,

          120,140,160,180,200,220,240,260,280,300)

df = data.frame(region, advertising, sales)

### ----------------------------------------------------

### Step 1: Hypothesis

### ----------------------------------------------------

# H0: Relationship between advertising and sales is SAME in 
all regions.

#     (i.e., slopes are equal)

# H1: Relationship is DIFFERENT across regions.

#     (i.e., slopes differ)
```

```
### ----------------------------------------------------
### Step 2: Estimate models
### ----------------------------------------------------
# Restricted model: ONE regression line for all regions
restricted_model = lm(sales ~ advertising, data = df)
# Unrestricted model: DIFFERENT slopes for each region (interaction model)
unrestricted_model = lm(sales ~ advertising * region, data = df)
### ----------------------------------------------------
### Step 3: Calculate the Test Statistic
### ----------------------------------------------------
# SSE for restricted and unrestricted models
SSE_R = sum(residuals(restricted_model)^2)
SSE_U = sum(residuals(unrestricted_model)^2)
# Degrees of freedom
df_R = df.residual(restricted_model)
df_U = df.residual(unrestricted_model)
# Numerator and denominator df
df_num = df_R - df_U
df_den = df_U
# F-statistic
F_stat = ((SSE_R - SSE_U) / df_num) / (SSE_U / df_den)
```

```
# p-value

p_val = pf(F_stat, df_num, df_den, lower.tail = FALSE)

### ----------------------------------------------------

### Step 4: Print results

### ----------------------------------------------------

cat("SSE (Restricted)   =", SSE_R, "\n")

cat("SSE (Unrestricted) =", SSE_U, "\n")

cat("F statistic        =", F_stat, "\n")

cat("p-value            =", p_val, "\n")

### ----------------------------------------------------

### Step 5: Using built-in ANOVA comparison

### ----------------------------------------------------

anova(unrestricted_model, restricted_model)
```

OUT PUT: -

```
Analysis of Variance Table

Model 1: sales ~ advertising * region
Model 2: sales ~ advertising
  Res.Df        RSS Df   Sum of Sq      F Pr(>F)
1     24 1.7073e-27
2     28 2.2206e-27 -4 -5.1334e-28 1.804  0.161
```

Inference:-

If the p-value < 0.05, slopes are NOT equal → relationship differs by region.

If the p-value > 0.05, slopes are equal → same linear relationship.

LABEXERCISE 6(b):

F-Test to check the Homogeneity of Regression Coefficients

Problem:

Suppose you have two groups of data and you want to test whether the regression slopes of a response variable y on a predictor variable x all the same across the two groups.

Aim:-

To test whether the regression slopes of a response variable y on a predictor variable x all the same across the two groups.

Procedure:

Step 1: Formulate the null & alternative hypothesis.

$H_0 : \beta_1 = \beta_2 = ... = \beta_k$ (regression coefficients are equal acrossgroups)

$H_1 : Not\ all\ \beta_i$ (regression coefficients are not equal across groups)

Step2:- Estimate the regression models.

Estimate the restricted model (assuming equal steps):

$-y = \beta_0 + \beta^1 x + \xi$ (Single regression line of all groups).

Estimate the unrestricted model (allowing different slopes).

$-y = \beta_0 + \beta^1 x + \beta^2 x + ... + \beta^K x + \xi$ (Separate regression lines for each group).

Step 3:- Calculate the Test statistic

Calculate the residual sum of squares for the restricted model (SSE-R)

Calculate the residual sum of squares for the unrestricted model (SSE-U)

Calculate the degrees of freedom for the restricted model(df-R) and unrestricted model(df-U)

Calculate the F_statistic:

$$-F = \frac{(SSE - R) - (SSE - U)}{(df - R - df - U)}$$

Step 4:- Determine the critical region and p-value:

Choose a significance level (ex:α=0.05)

Determine the critical value from the F-distribution table (or) using software

Calculate the p-value associated with the F-statistic.

Step 5:- Make a decision:

If the p-value is less than the level of significance (α), reject H₀ and conclude that the regression coefficients are not equal across groups.

If the p-value is greater than the significance level (α) fail to reject H₀ and conclude that the regression coefficients are equal across groups.

R-CODE:

```
group = rep(c("A","B"), each = 10)
x = c(2.1,2.3,2.7,3.0,3.2,3.5,3.8,4,4.3,4.5,
      2,2.4,2.8,3.1,3.3,3.6,3.9,4.2,4.4,4.6)
y = c(4.5,4.8,5.1,5.3,5.7,6,6.2,6.5,6.8,7,
      5,5.2,5.4,5.7,6,6.3,6.5,6.8,7,7.2)
# Separate models
model_A = lm(y[group=="A"] ~ x[group=="A"])
model_B = lm(y[group=="B"] ~ x[group=="B"])
# Full and reduced model
combined_model = lm(y ~ x * group)      # includes interaction
reduced_model  = lm(y ~ x + group)      # no interaction
# Compute statistics
RSS_full = sum(resid(combined_model)^2)
RSS_reduced = sum(resid(reduced_model)^2)

df_interaction = length(coef(combined_model)) -
                 length(coef(reduced_model))
df_error = df.residual(combined_model)
F_stat = ((RSS_reduced - RSS_full) / df_interaction) /
         (RSS_full / df_error)
p_value  =  pf(F_stat,  df_interaction,  df_error,  lower.tail  =
FALSE)
cat("F_statistic:", F_stat, "\n")
cat("P_value:", p_value, "\n")
```

OUT PUT:

```
F_statistic: 10.75738
P_vaue: 0.004715281
```

Inference:-

The p-value is less than the level of significance (α), we reject H_0 and conclude that the regression coefficients are not equal across groups.

LABEXERCISE 7(a):

Bartlett's Test for Homogeneity of Several Variances

Problem:

Bartlett's test is used to check the homogeneity of various across groups

Aim:

To test the Bartlett's test this is used to check the homogeneity of variances across groups.

Procedure:

1)Compute the sample variance for each group

2)Compute the pooled variance has weighted the average of sample variance.

3)Calculate the Bartlett's test statistic using the formula

$$T = \frac{\left(N-k\right)*\log(S_p^2) - \sum_{i=1}^{k}(n_i - 1)*\log(S_i^2)}{1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{(n_i-1)} - \frac{1}{(N-k)}\right]}$$

where,

N: Total number of observations across all the groups

k: Number of groups

S_p^2 : Pooled variance

$\mathrm{S}_\mathrm{i}^2 = \text{Sample variance for group i}$

$\mathrm{n}_\mathrm{i} = \text{Numberof observations in group i}$

Compare the test statistic T with the critical value from the Chi-square distribution with k-1 d.f.

R-CODE: -

```
Group1 = c(7.9, 7.6, 7.1, 6.8)
Group2 = c(8.4, 8.6, 8.3, 8.7)
Group3 = c(9.2, 9.4, 9.1, 9.3)

K = 3
```

```
var1 = var(Group1)
var2 = var(Group2)
var3 = var(Group3)

n1 = length(Group1)
n2 = length(Group2)
n3 = length(Group3)

N = n1 + n2 + n3

# Pooled variance
pooled_variance = ((n1-1)*var1 + (n2-1)*var2 + (n3-1)*var3) /
(N - K)

# Bartlett's test statistic (before correction factor)
T = (N - K) * log(pooled_variance) -
    ((n1-1)*log(var1) + (n2-1)*log(var2) + (n3-1)*log(var3))

# Correction factor
C = 1 + (1 / (3 * (K - 1))) * ( (1/(n1-1)) + (1/(n2-1)) +
(1/(n3-1)) - 1/(N-K) )

# Corrected test statistic
T = T / C

# p-value
p_value = 1 - pchisq(T, K - 1)

cat("Bartlett’s test statistic:", T, "\n")
cat("P_value:", p_value, "\n")
```

OUTPUT:

```
Bartlett'steststatistic: 5.052548
 P_value:  0.07995639
```

Inference:

Bartlett’s test statistic T=5.052548 & p-value=0.07995639. As the p-value is greater than 0.05, the we accept the null hypothesis and we conclude that the variances are not significant different across all groups.

LABEXERCISE 7(b):

Bartlett's Test for Homogeneity of Several Variances

Problem:

Bartlett's test is used to check the homogeneity of various across groups

Aim:

To test the Bartlett's test which is used to check the homogeneity of variances across groups.

Procedure:

1) Compute the sample variance for each group

2) Compute the pooled variance has weighted the average of sample variance.

3) Calculate the Bartlett's test statistic using the formula

$$T = \frac{(N-k)*\log(S_p^2) - \sum_{i=1}^{k}(n_i - 1)*\log(S_i^2)}{1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{(n_i-1)} - \frac{1}{(N-k)}\right]}$$

where,

N: Total number of observations across all the groups

k: Number of groups

S_p^2 : Pooled variance

$\mathrm{S}_\mathrm{i}^2 =$ Sample variance for group i

$\mathrm{n}_\mathrm{i} =$ Numberof observations in group i

Compare the test statistic T with the critical value from the Chi-square distribution with k-1 d.f.

R-CODE:

```
Group1=c(5.2,5.8,6.1,5.5,5.9)
Group2=c(7.3,7.8,7.5,7.6,7.7)
Group3=c(4.8,5,5.3,5.1,5.2)
K=3
var1=var(Group1)
var2=var(Group2)
```

```
var3=var(Group3)
n1=length(Group1)
n2=length(Group2)
n3=length(Group3)
N=n1+n2+n3
pooled_variance=((n1-1)*var1+(n2-1)*var2+(n3-1)*var3)/(N-K)
T=(N-K)*log(pooled_variance)-((n1-1)*log(var1)+(n2-
1)*log(var2)+(n3-
1)*log(var3))
T=T/(1+(1/3*(K-1)))*(1/(n1-1)+1/(n2-1)+1/(n3-1)-1/(N-K))
p_value=1-pchisq(T,K-1)
cat("Bartlett's test statistic:", T,"\n")
cat("P_value:",p_value, "\n")
```

OUTPUT:

```
Bartlett's test statistic: 0.8542849
P_value: 0.6523706
```

Inference:

Bartlett's test (T = 0.8543, p = 0.6524) shows that p > 0.05. Therefore, we fail to reject the null hypothesis and conclude that the variances are not significantly different across the groups.

LABEXERCISE 8(a):

SPRT Binomial OC Curve

Problem:-

Draw a SPRT Binomial OC Curve.

Aim:

To draw a SPRT Binomial OC Curve.

Procedure:

Step 1:- Define parameters P_0, P_1, A, B

Step 2:- Calculate critical value S_m which is used in the rule for SPRT

$$S_m = \frac{\log \dfrac{(1-p_0)}{(1-p_1)}}{\log\left(\dfrac{p_1}{p_0}\right) - \log \dfrac{(1-p_1)}{(1-p_0)}}$$

Step 3:- Define probability values for plotting which are o, p_0, S_m, l

Step 4:- Define likelihood ratio.

Step5:- Calculate h₀& h₁

$$h_0 = \left(\frac{\log\left(\dfrac{b}{1-a}\right)}{\log\left(\dfrac{p_1}{p_0}\right) - \log\left(\dfrac{1-p_1}{1-p_0}\right)} \right)$$

$$h_1 = \left(\frac{\log\left(\dfrac{1-b}{a}\right)}{\log\left(\dfrac{p_1}{p_0}\right) - \log\left(\dfrac{1-p_1}{1-p_0}\right)} \right)$$

Step 6: Calculate likelihood ratio values for plotting

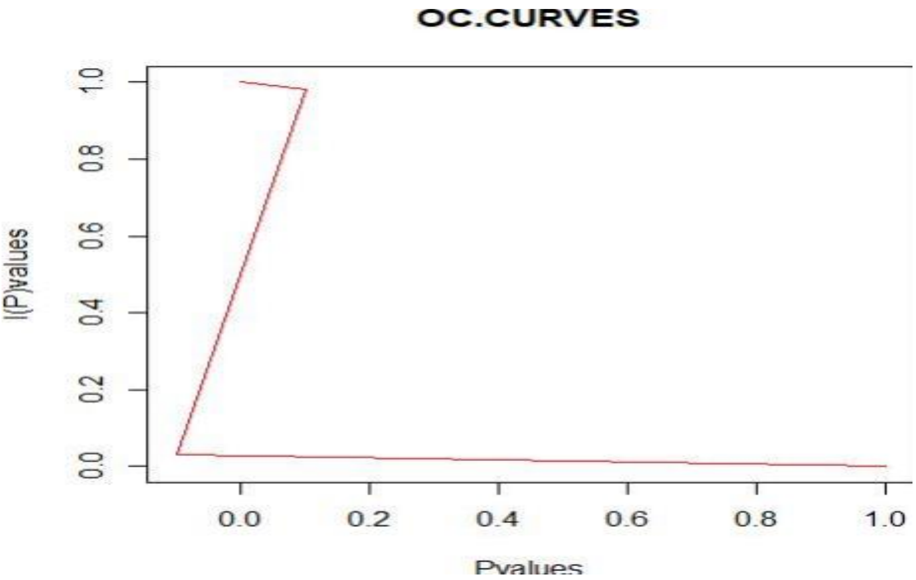
Step 7- Plot the OC Curve

R-CODE: -

```
P0=0.1
P1=0.3
a=0.02
b=0.03
Sm=(log(1-P0)/(1-P1))/(log(P1/P0)-(log(1-P1)/(1-P0)))
P=c(0,P0,Sm,1)
l0=1
lP0=1-a
lP1=b
l1=0
d=c("f","f","s","f","f","s","f","f","f","f","s","f","s","s","f
","s","f","f","f","s", "f","f","s")
h0=(log(b/(1-a)))/(log(P1/P0)-(log(1-P1)/(1-P0)))
h1=(log((1-b)/a))/(log(P1/P0)-(log(1-P1)/(1-P0)))
ls=h1/(h1+abs(h0))
Ez=P*log(P1/P0)+(1-P)*log((1-P1)/(1-P0))
lp=c(l0,lP0,lP1,l1)
plot(P,lp,type="l",col="red", lwd=0.01,xlab="Pvalues",
ylab="l(P)values", main="OC.CURVES")
```

OUTPUT:

```
> Sm
> h0
[1] -0.1006845
[1] -2.332138
> h1
> ls
[1] 2.596507
[1] 0.5268196
> Ez
[1] -0.2513144 -0.1163218 -0.3872311 1.0986123
> lp
[1] 1.00 0.98 0.03 0.00
```



Inference:

h₀ = -2.332138, h₁ = 2.596507

S_m =-0.1006845

ls=0.5268196

Ez=-0.2513144 -0.1163218 -0.3872311 1.098612

lp=1 0.98 0.03 0

LAB EXERCISE 8(b):

SPRT Binomial ASN Curve

Problem:

Draw a SPRT Binomial ASN Curve.

Aim:

To draw a SPRT Binomial ASN Curve.

Procedure:

ASN-Average Sample Number

Step1:- Define parameters a₀, a₁, a, b,d, v

Step2:- Calculate h₀, h₁& S_m

$$h_0 = \left(\frac{v}{a_1 - a_0}\right) * \log\left(\frac{b}{1-a}\right)$$

$$h_1 = \left(\frac{v}{a_1 - a_0}\right) * \log\left(\frac{1-b}{a}\right)$$

$$S_m = \frac{h_0 + h_1}{2}$$

Step 3:- Define threshold values t₁, t₂.

Step4:- Calculate likelihood ratios for each threshold values

$$lt_1 = \left[\left(e^{\frac{2}{v}} * \left(S_m - t_1\right) * h_1\right) - \frac{1}{\left(e^{\frac{2}{v}} * \left(S_m - t_1\right) * h_1\right)} - e^{\frac{2}{v}} * \left(S_m - t_1\right) * h_0\right]$$

$$lt_2 = \left(\frac{\log\left(\frac{1-b}{a}\right)}{\log\left(\frac{1-b}{a}\right) - \log\left(\frac{b}{1-a}\right)}\right)$$

Step5:- Calculate average samples numbers E_{n_1}, E_{n_2}

$$En_1 = \left[\frac{lt_1 * (h_0 - h_1) + h_1}{t_1 - S_m}\right]$$

$$En_2 = h_0 * \frac{h_1}{v}$$

Step 6:- Combine threshold values & ASN values

$$t_3 = c(t_1, t_2)\ ,\ lt_3 = c(lt_1, lt_2)\ ,\ En = c(En_1, En_2)$$

Step 7:- Plot the ASN curve

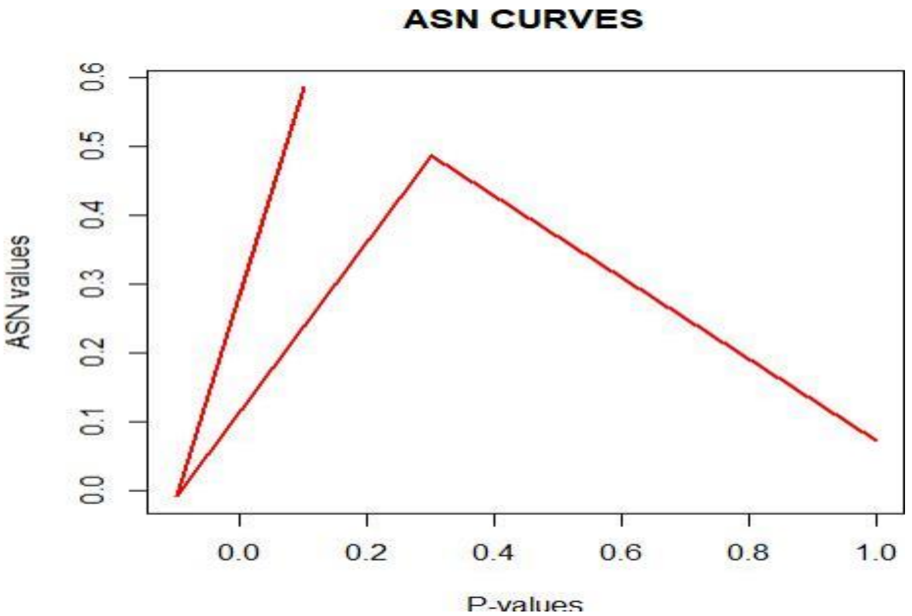
R-CODE:

```
P0=0.1
P1=0.3
a=0.02
b=0.03
Sm=(log(1-P0)/(1-P1))/(log(P1/P0)-(log(1-P1)/(1-P0)))
P=c(0,P0,Sm,P1,1)
l0=1
lP0=1-a
lP1=b
l1=0
lP=c(l0,lP0,lP1,l1)
lP=c(l0,lP0,(lP0+lP1)/2,lP1,l1)
Ez=P*log(P1/P0)+(1-P)*log((1-P1)/(1-P0))
Em1=(lP*log(b/(1-a))+(1-lP)*log((1-b)/a))/Ez
Em2=log(b/(1-a))*log((1-b)/a)/(log(P1/P0)*log((1-P1)/(1-P0)))
En=Em1/Em2
plot(P,En,type="l",col="red", lwd=2,xlab="P-values", ylab="ASN
values", main="ASN
CURVES")
```

OUTPUT:

```
> Ez
1] 49.01359
> Em2
[1] -0.1006845
```

```
[1] 49.01359
> Ez
[1] -0.2513144 -0.1163218 -0.3872311 0.1536636 1.0986123
> Em1
[1] 13.8724832 28.7048348 -0.4151648 23.8216893 3.5331516
> En
[1] 0.283033423 0.585650567 -0.008470402 0.486022161 0.072085148
```



Inference:

S_m = -0.1006845

Ez = -0.2513144 -0.1163218 -0.3872311 0.1536636 1.098612

En_1 = 13.87248 28.70483 -0.4151648 23.82169 3.533152

En_2 = 49.01359

En = 0.2830334 0.5856506 -0.008470402 0.4860222 0.07208515

LAB EXERCISE 9:

SPRT Normal OC Curve

Problem:

Draw a SPRT Normal OC Curve.

Aim:

To draw a SPRT Normal OC Curve.

Procedure:

Step 1:- Define parameters p_0, p_1, A, B

Step 2:- Calculate critical value S_m which is used in the rule for SPRT

$$S_m = \frac{\log\frac{(1-p_0)}{(1-p_1)}}{\log\left(\frac{p_1}{p_0}\right) - \log\frac{(1-p_1)}{(1-p_0)}}$$

Step 3:- Define probability values for plotting which are o, p_0, S_m, l

Step 4:- Define likelihood ratio.

Step5:- Calculate h_0 & h_1

$$h_0 = \left(\frac{\log\left(\frac{b}{1-a}\right)}{\log\left(\frac{p_1}{p_0}\right) - \log\left(\frac{1-p_1}{1-p_0}\right)} \right)$$

$$h_1 = \left(\frac{\log\left(\frac{1-b}{a}\right)}{\log\left(\frac{p_1}{p_0}\right) - \log\left(\frac{1-p_1}{1-p_0}\right)} \right)$$

Step 6: Calculate likelihood ratio values for plotting

Step 7- Plot the OC Curve

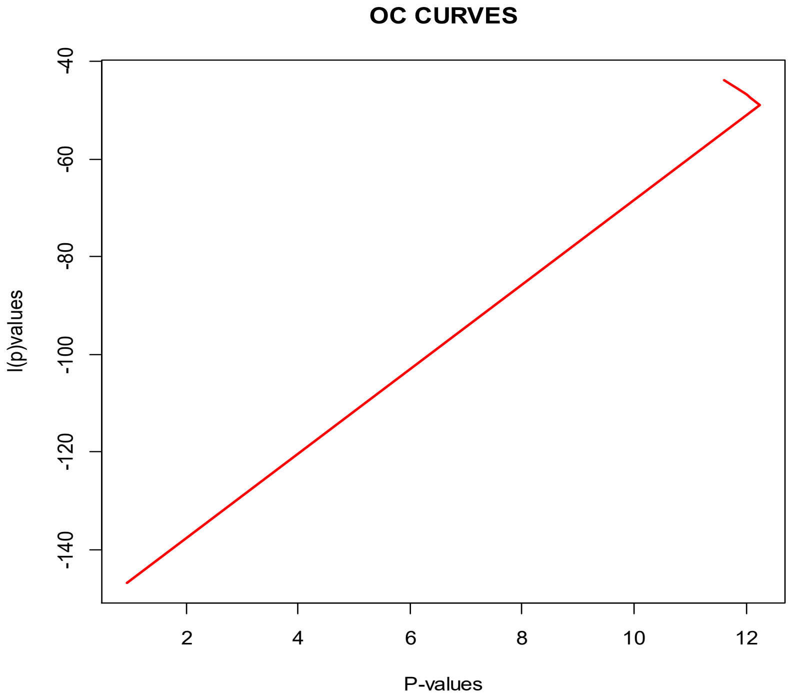
R-CODE:

```
a0=135
a1=150
a=0.01
b=0.03
d=c(151,144,121,137,130,136,155,130,142,136,125,145,106,108)
v=25
t1=c(135,140,144,146,150)
h0=(v/(a1-a0))*log(b/(1-a))
h1=(v/(a1-a0))*log((1-b)/a)
Sm=(h0+h1)/2
lt1=((exp(2/v)*(Sm-t1)/h1))-1/((exp(2/v)*(Sm-t1)/h1))-
(exp(2/v)*(Sm-t1)/h0)
t2=Sm
lt2=(log((1-b)/a))/(log(1-b))-log(b/(1-a))
t3=sqrt(c(t1,t2))
lt3=c(lt1,lt2)
plot(t3,lt3,type="l",col="red",lwd=2,xlab="P-values",
ylab="l(p)values", main="OC CURVES")
```

OUTPUT:

```
> h0
[1] -5.827513
> h1
[1] 7.624518
> Sm
[1] 0.8985028
> lt1
[1] -43.92895 -45.57069 -46.88399 -47.54061 -48.85380
> lt2
[1] -146.6949
```

```
> lt3
[1] -43.92895 -45.57069 -46.88399 -47.54061 -48.85380 -146.69489
```



Inference:

h₀ = -5.827513

h₁ = 7.624518

S_m = 0.8985028

lt₁	=	-43.92895	-45.57069	-46.88399	-47.54061	-48.8538	
lt₃	=	-43.92895	-45.57069	-46.88399	-47.54061	-48.8538	-146.6949