Submitted by-Aditya Gautam

Roll No-12

M.sc. 3rd semester

Date of Assignment-03/12/2020

Date of Submission-11/12/2020

Experiment No-08

Topic- DISCRIMINANT ANALYSIS AND MAHALANOBIS D^2 STATISTIC.

Problem- The following table shows the marks obtained in Mathematics(x) and marks obtained in statistics (y) by two batches of students.

BATCH-1	
x	y
12	34
45	56
44	35
52	63
8	32
39	48
71	84
38	57
38	51
47	62

BATCH-2	
x	Y
56	54
67	66
49	72
89	97
58	76
53	32
56	81
78	98
64	78
58	40

Compute the MAHALANOBIS- D^2 statistic and hence perform D^2 test for testing the equality of marks obtained by the two groups. Also, classify to which batch a student scoring 52 in mathematics and 70 in statistic will belong.

Theory-

The Fisher's discriminant function is given by

$\hat{y}=(\overline{X_{\sim 1}}-\overline{X_{\sim 2}})'\ S^{-1}_{pooled}\underset{\sim}{X}$ $\qquad \underset{\sim}{X}=\begin{bmatrix}X_1\\X_2\end{bmatrix}$

Where, $\overline{X_{\sim 1}}=((\overline{X_1})_1, (\overline{X_2})_1,.....(\overline{X_p})_1)$; $(\overline{X_i})_1=\frac{1}{n_1}\sum_{j=1}^{n_1}(X_{ij})_1$ i=1,2,.....p

$\overline{X_{\sim 2}}=((\overline{X_1})_2, (\overline{X_2})_2,.....(\overline{X_p})_2)$; $(\overline{X_i})_2=\frac{1}{n_2}\sum_{j=1}^{2}(X_{ij})_2$ i=1,2,.....p

$S_{pooled}=\frac{1}{n_1+n_2-1}[(n_1-1)S_1+(n_2-1)S_2]$

Where, $S_1=((S_{ij})_{1)},(S_{ij})_1)=\frac{1}{n_1-1}\sum_{k=1}^{n_1}\{(X_{ik})_1-(\overline{X}_i)_1\}\{(X_{jk})_1-(\overline{X}_j)_1\}$

$S_2=((S_{ij})_2),(S_{ij})_2)=\frac{1}{n_2-1}\sum_{k=1}^{n_2}\{(X_{ik})_2-(\overline{X}_i)_2\}\{(X_{jk})_2-(\overline{X}_j)_2\}$

The function $\hat{y}$ is a function which maximally separates the two populations and the maximum separation in the two sample from the population is $D^2=(\overline{X_{\sim 1}}-\overline{X_{\sim 2}})'\ S^{-1}_{pooled}(\overline{X_{\sim 1}}-\overline{X_{\sim 2}})$

Which is the MAHALANOBIS D^2 STATISTIC .

$$\begin{bmatrix}(X_{11})_1 & (X_{21})_1 & \cdots & (X_{p1})_1\\(X_{12})_1 & (X_{22})_1 & \cdots & (X_{p2})_1\\\vdots & \vdots & \vdots & \vdots\\(X_{1n})_1 & (X_{2n})_1 & \cdots & (X_{pn})_1\end{bmatrix} \text{ and } \begin{bmatrix}(X_{11})_2 & \cdots & (X_{p1})_2\\(X_{12})_2 & \cdots & (X_{p2})_2\\\vdots & \vdots & \vdots\\(X_{1n})_n & \cdots & (X_{pn})_2\end{bmatrix}$$

Here, are two samples from the multivariate normal populations $N_p(\underset{\sim}{\mu_1}, \Sigma)$, $N_p(\underset{\sim}{\mu_2}, \Sigma)$ respectively. Further, it is assumed that two population have the same variance covariance matrix.

Here, we are to test the hypothesis $H_0:(\underset{\sim}{\mu_1}-\underset{\sim}{\mu_2})^2=\underset{\sim}{0}$ against $H_1:(\underset{\sim}{\mu_1}-\underset{\sim}{\mu_2})^2\neq\underset{\sim}{0}$

i.e. the hypothesis of equality of the two population means . It can be accomplished from the basis of the D^2 statistic for the D^2 test is $\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}(\frac{n_1n_2}{n_1+n_2})D^2\sim F_{p,n1+n2-p-1}$

Conclusions are drawn accordingly

For allocating the observation $\underset{\sim}{X}=\begin{bmatrix}X_{01}\\X_{02}\end{bmatrix}$ into one of the two groups , the allocation rule based on $\hat{y}$ is –

1) Allocate $\underset{\sim}{X}_0$ to $N_p(\underset{\sim}{\mu}_1, \Sigma)$ (group 1) if $\hat{y}_0 = (\overline{X_{\sim 1}} - \overline{X_{\sim 2}})' S^{-1}_{pooled} \underset{\sim}{X}_0 \geq \frac{D^2}{2} = (\hat{m})$

2) Allocate $\underset{\sim}{X}_0$ to $N_p(\underset{\sim}{\mu}_2, \Sigma)$ (group 2) if $\hat{y}_0 = (\overline{X_{\sim 1}} - \overline{X_{\sim 2}})' S^{-1}_{pooled} \underset{\sim}{X}_0 < \frac{D^2}{2}$

Calculation-

The R-programming to obtain the solution for the given problem-

```
x1=c(12,45,44,52,8,39,71,38,38,47,34,56,35,63,32,48,84,57,51,62)
dim(x1)=c(10,2)
x2=c(56,67,49,89,58,53,56,78,64,58,54,66,72,97,76,32,81,98,78,40)
dim(x2)=c(10,2)
dim(x2)
mean1=mat.or.vec(2,1)
mean2=mat.or.vec(2,1)
for(i in 1:2){
mean1[i]=mean(x1[,i])
mean2[i]=mean(x2[,i])}
mean=array(c(mean1-mean2),dim=c(2,1))
mean
n1=10
n2=10
var11=mat.or.vec(2,1)
var12=mat.or.vec(2,1)
var21=mat.or.vec(2,1)
var22=mat.or.vec(2,1)
for(i in 1:2){
var11[i]=cov(x1[,1],x1[,i])*((n1-1)/(n1+n2-2))
var12[i]=cov(x2[,1],x2[,i])*((n2-1)/(n1+n2-2))}
for(i in 1:2){
var21[i]=cov(x1[,2],x1[,i])*((n1-1)/(n1+n2-2))
var22[i]=cov(x2[,2],x2[,i])*((n2-1)/(n1+n2-2))}
s1=c(var11,var21)
s1
dim(s1)=c(2,2)
dim(s1)
s2=c(var12,var22)
s2
dim(s2)=c(2,2)
dim(s2)
```

```
s_p=s1+s2
s_p
D2=t(mean)%*%solve(s_p)%*%mean
D2
p=2
cal_value=((n1+n2-p-1)/(p*(n1+n2-2)))*((n1*n2)/(n1+n2))*D2
cal_value
tab_value=qf(0.95,2,17,0)
tab_value
x0=array(c(52,70),dim=c(2,1))
x0
y0=t(mean)%*%solve(s_p)%*%x0
y0
m=D2/2
m
```

Conclusion-

The MAHALANOBIS D^2- STATISTIC is 2.333933 . Since the calculated value (i.e. 5.510676) is more than the tabulated value (i.e. 3.591531) of F we reject our null hypothesis at 5% level of significance and conclude that the equality of marks obtained by the two groups are significantly different.
And a student scoring 52 in Mathematics and 70 in statistics will belong to batch 2 since

$$\hat{y}_0(=-4.515759) < \frac{D^2}{2}(=1.166967)$$