Applied Multivariate Analysis Assignment 2

Souvik Das (23414110016)

2024-08-24

Problem No 1

In a study pertaining to woman's nutrition, daily nutrient intake was measured for a random sample of 737 women aged 25-50 years. Five nutritional components were measured: calcium, iron, protein, vitamin A, and vitamin C.

The given data for the 5 nutritional components on 737 observations are imported and cleaned in the following part :

```r
##Problem 1

#Importing and cleaning the given data
Nutrient_data <- read.csv("C://Users//HP//Desktop//AP//ProblemSet_2//nutrient_data.csv")
which(is.na(Nutrient_data))
```

```
## integer(0)
```

```r
Nutrient_data <- Nutrient_data[,-1]

#Rechecking given means
xbar <- sapply(Nutrient_data, mean)
xbar
```

```
##   calcium      iron   protein         a         c
## 624.04925  11.12990  65.80344 839.63535  78.92845
```

(a) Now we are going to conduct **Hotelling T-square** test to check whether women meet the recommended nutritional intake guideline or not.

Test :

$$T^2 = n(\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0)$$

$$F = \frac{(n-p)}{p(n-1)} * T^2 \sim F_{p,n-p}$$

In the following part, we compute the required quantities manually without using any package :

```r
#Performing Hotelling T2 test manually
S <- cov(Nutrient_data)
mu0 <- c(1000, 15, 60, 800, 75)
p <- ncol(Nutrient_data)
n <- nrow(Nutrient_data)
alpha <- 0.01
```

```
HT2 <- n * t(xbar - mu0) %*% solve(S) %*% (xbar - mu0)
Test.Statistic <- HT2 * (n-p)/ (p*(n-1))
Test.Statistic
```

```
##          [,1]
## [1,] 349.7968
```

```
Critical.Value <- qf(1-alpha, p, n-p)
Critical.Value
```

```
## [1] 3.042279
```

Now we perform the test by using ICSNP package.

```
#Performing Hotelling T2 test using ICSNP
library(ICSNP)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: ICS
```

```
test <- HotellingsT2(Nutrient_data, mu=mu0, test='f')
test
```

```
##
##  Hotelling's one sample T2-test
##
## data:  Nutrient_data
## T.2 = 349.8, df1 = 5, df2 = 732, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to c(1000,15,60,800,75)
```

Conclusion : Here we can see that the value of the test statistic exceeds the critical value. Also we can see that the p-value for the test is significantly small. For these reasons, we are going to **reject the null** hypothesis at 0.01 level of significance. The null hypothesis was the mean vector of the intake is equal to the mu0 (recommended intake of the nutrients).

(b) As women fail to meet the guideline, we need to identify the nutrients for which they could not meet the guideline.

Simultaneous CI :

$$\bar{X}_j \pm \sqrt{\frac{p(n-1)}{n-p}F_{p,n-p,\alpha}}\sqrt{\frac{s_j^2}{n}}$$

Bonferoni CI :

$$\bar{X}_j \pm t_{n-1,\frac{\alpha}{2p}}\sqrt{\frac{s_j^2}{n}}$$

For this we need to calculate simultaneous and Bonferoni's Confidence Intervals for the mean intake.

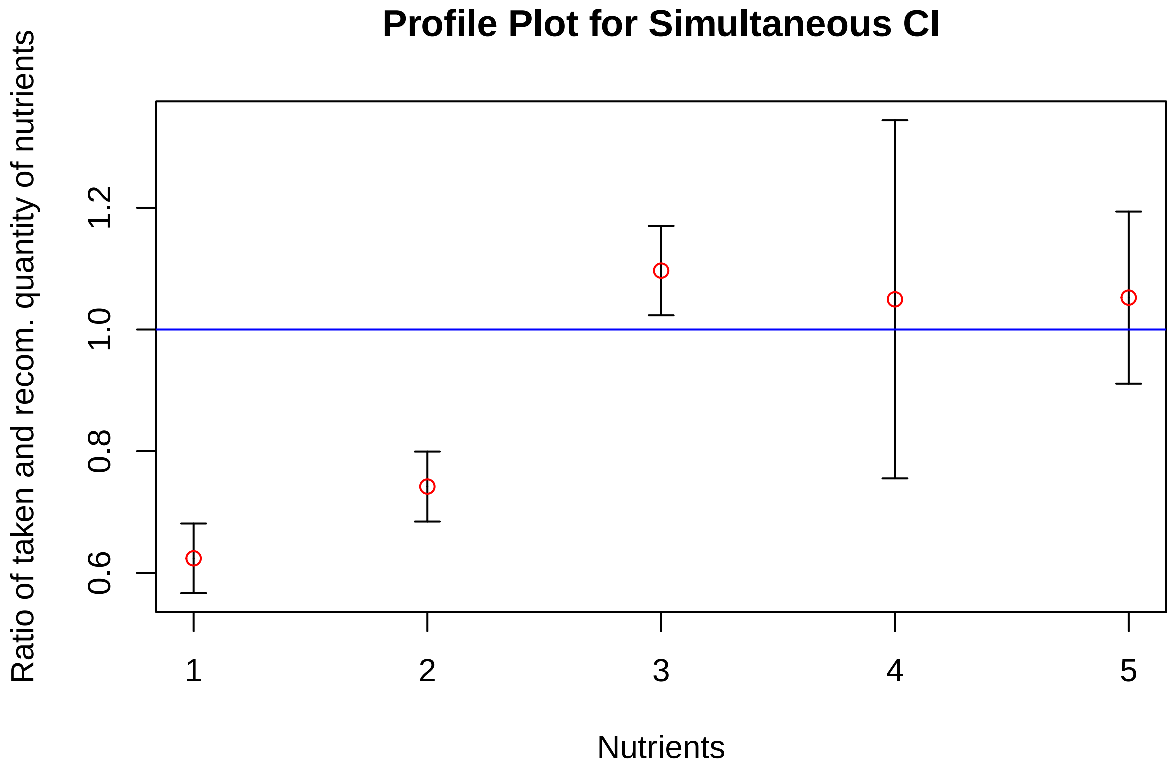
```
#Calculating Simultaneous Interval
MargSC <- sqrt(p * (n-1) * qf(1-alpha, p, (n-p)) / (n-p)) * sqrt(diag(S) / n)
MargSC <- as.matrix(MargSC)
SCI <- data.frame('lower.bound' = as.vector(xbar - MargSC),
                  'upper bound' = xbar + MargSC)
SCI
```

```
##          lower.bound upper.bound
## calcium   566.81868   681.27983
## iron       10.26784    11.99196
## protein    61.39879    70.20809
## a         604.31263  1074.95806
## c          68.32654    89.53035
```

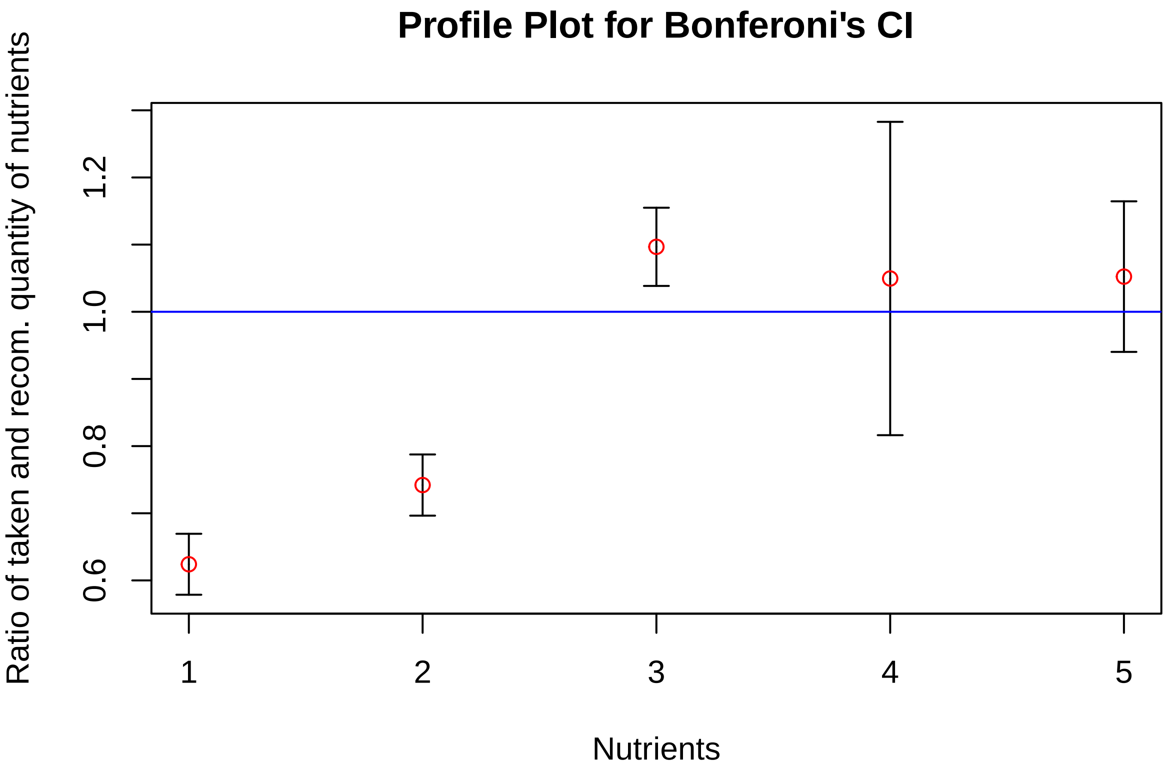
```
#Calculating Bonferoni Interval
MargBI <- qt(1-(alpha / (2*p)), n-1) * sqrt(diag(S) / n)
MargBI <- as.matrix(MargBI)
BI <- data.frame('lower bound' = as.vector(xbar - MargBI),
                 'upper bound' = xbar + MargBI)
BI
```

```
##          lower.bound upper.bound
## calcium   578.66450   669.43401
## iron       10.44627    11.81353
## protein    62.31048    69.29640
## a         653.02072  1026.24998
## c          70.52097    87.33593
```

(c) Now we need to generate profile plot for both types of CI.



And now we need to do the same for Bonferoni's CI.



`Conclusion` : From two profile plots, we can easily detect that

- vitamin A and C contain the value 1
- protein does not contain 1 but it is very near 1.
- Calcium and iron are significantly far from 1.

So, the **main reasons** for women not meeting the intake guidelines are **Calcium and Iron.**

Problem No 2

A shoe company evaluates new shoe models based on five criteria: style, comfort, stability, cushioning, and durability. Each of the first four criteria is evaluated on a scale of 1 to 20 and the durability criterion is evaluated on a scale of 1 to 10. The company is considering phasing out an existing shoe model (Model 2), replacing it with a new prototype (Model 1).

The data are from 25 observation on these 5 variables for Model 1 and Model 2. The given data are imported and cleaned in the following step :

```r
#Problem 2
#loading the data
shoe <- read.csv("C://Users//HP//Desktop//AP//ProblemSet_2//shoe.csv", header = FALSE)
which(is.na(shoe))
```

```
##  [1] 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187
## [20] 188 189 190 191 192 193 194 195 196
```

```r
shoe <- shoe[c(3:28),]
names(shoe) <- shoe[1,]
row.names(shoe) <- shoe[,1]
Model1 <- data.matrix(shoe[c(-1),c(2:6)])
Model2 <- data.matrix(shoe[c(-1),c(8:12)])
which(is.na(Model1))
```

```
## integer(0)
```

```r
which(is.na(Model2))
```

```
## integer(0)
```

(a) Now we are going to conduct **Hotelling T-square** test for *multivariate paired sample test for mean* to check if there are significant difference between two models or not.

Test :

$$Y_i = X_{1i} - X_{2i}; S_Y = \text{covariance matrix of } Y$$

$$T^2 = n\bar{Y}'S_Y^{-1}\bar{Y}$$

$$F = \frac{n-p}{p(n-1)}T^2$$

Reject H_0 at level α if $F > F_{p,n-p,\alpha}$.

In the following part, we compute the required quantities manually without using any package :

```r
#(a)
#calculation of yi's
Y <- Model1 - Model2
Ybar <- apply(Y, 2, mean)
Ybar
```

```
##      Style    Comfort  Stability    Cushion Durability
##       1.52       0.76       1.16       1.28       2.04
```

```
#Performing paired Hotelling's T-square Test
S <- cov(Y)
p <- ncol(Y)
n <- nrow(Y)
alpha <- 0.05

HT2 <- n * t(Ybar) %*% solve(S) %*% (Ybar)
Test.Statistic <- HT2 * (n-p)/ (p*(n-1))
Test.Statistic
```

```
##          [,1]
## [1,] 3.736954
```

```
Critical.Value <- qf(1-alpha, p, n-p)
Critical.Value
```

```
## [1] 2.71089
```

Here we perform the test using ICSNP :

```
#Test using ICSNP
library(ICSNP)
HotellingsT2(Y, test='f')
```

```
##
##  Hotelling's one sample T2-test
##
## data:  Y
## T.2 = 3.737, df1 = 5, df2 = 20, p-value = 0.01497
## alternative hypothesis: true location is not equal to c(0,0,0,0,0)
```

Conclusion : Here we can see that the value of the test statistic exceeds the critical value. Also we can see that the p-value for the test is significantly small. For these reasons, we are going to **reject the null** hypothesis at 0.05 level of significance. The null hypothesis was that the two mean vectors of two models are equal.

There is a significant difference between two models at 5% level of significance.

(b) As there is a significant difference between two models, we need to detect the individual criteria for which there is a significant difference. We need to find both of the simultaneous and Bonferoni's CI for this purpose.

Simultaneous CI :

$$\bar{X}_j \pm \sqrt{\frac{p(n-1)}{n-p}F_{p,n-p,\alpha}}\sqrt{\frac{s_j^2}{n}}$$

Bonferoni CI :

$$\bar{X}_j \pm t_{n-1,\frac{\alpha}{2p}}\sqrt{\frac{s_j^2}{n}}$$

```
#(b)
#Calculating Simultaneous Interval
MargSC <- sqrt(p * (n-1) * qf(1-alpha, p, (n-p)) / (n-p)) * sqrt(diag(S) / n)
MargSC <- as.matrix(MargSC)
```

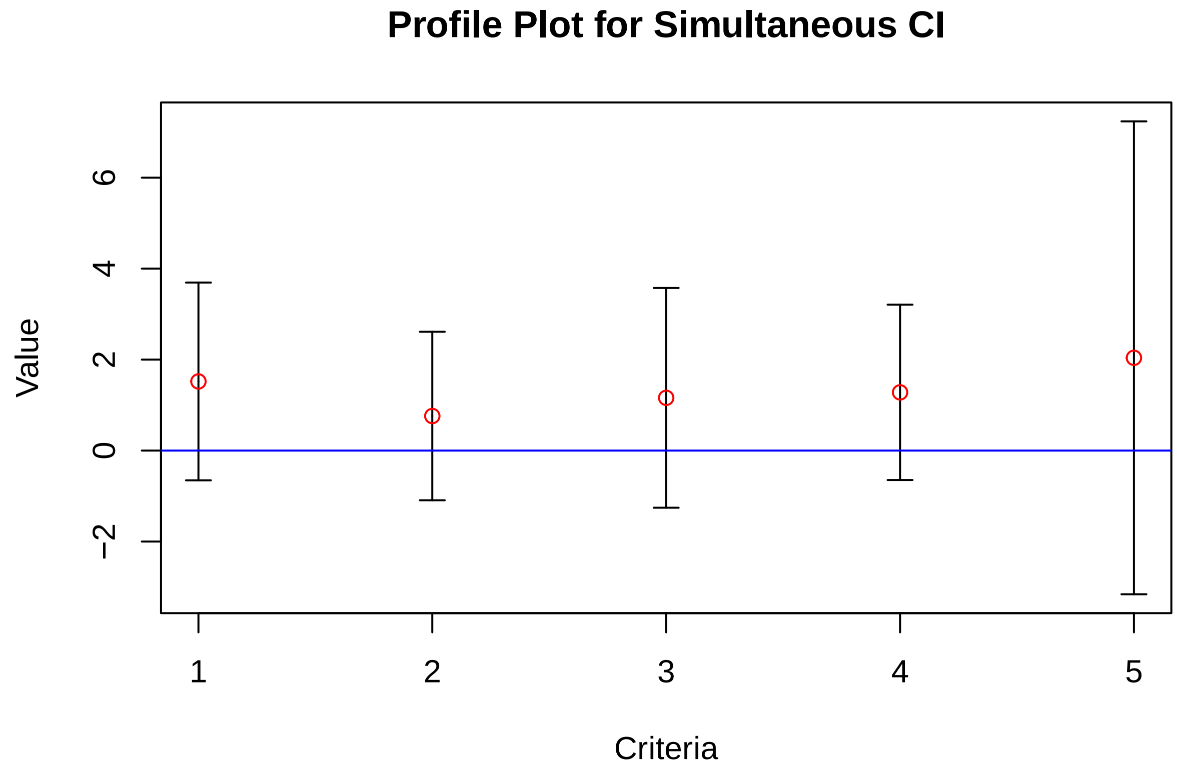
```r
SCI <- data.frame('lower.bound' = as.vector(Ybar - MargSC),
                  'upper bound' = Ybar + MargSC)
SCI
```

```
##              lower.bound upper.bound
## Style         -0.6533510    3.693351
## Comfort       -1.0922695    2.612269
## Stability     -1.2562310    3.576231
## Cushion       -0.6474344    3.207434
## Durability    -3.1587005    7.238701
```

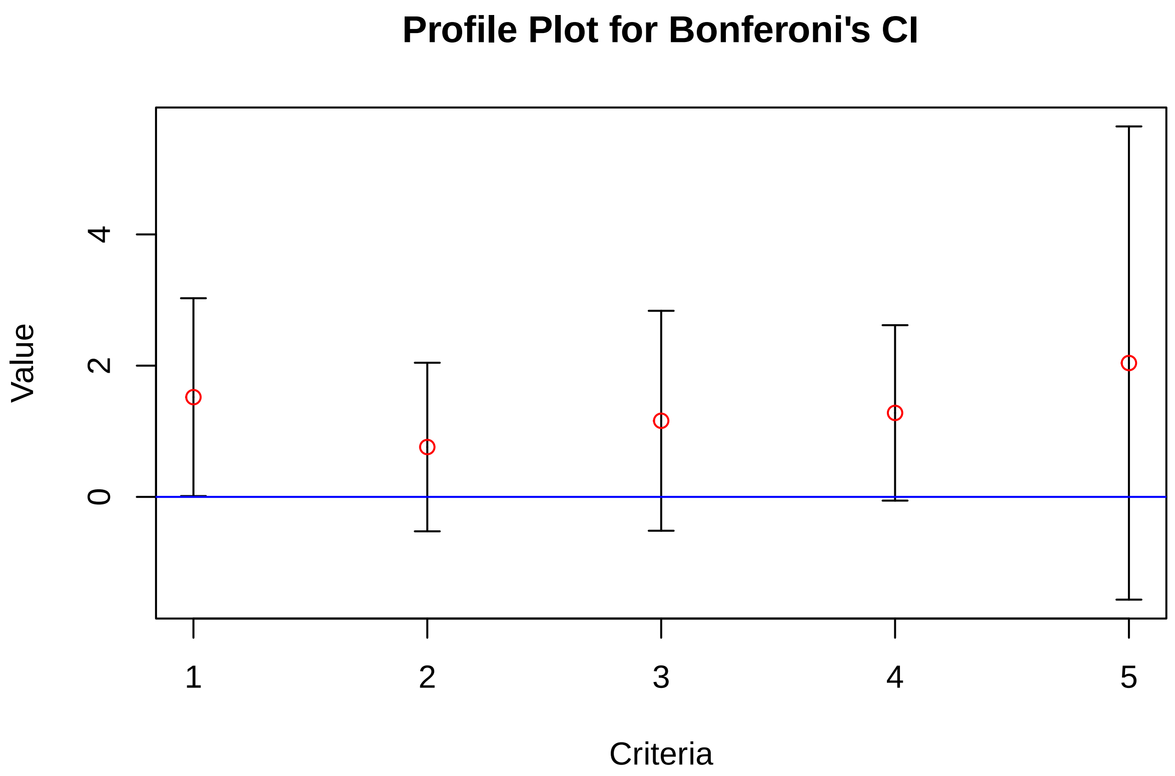
```r
#Calculating Bonferoni Interval
MargBI <- qt(1-(alpha / (2*p)), n-1) * sqrt(diag(S) / n)
MargBI <- as.matrix(MargBI)
BI <- data.frame('lower bound' = as.vector(Ybar - MargBI),
                 'upper bound' = Ybar + MargBI)
BI
```

```
##              lower.bound upper.bound
## Style         0.01276358    3.027236
## Comfort      -0.52456381    2.044564
## Stability    -0.51567569    2.835676
## Cushion      -0.05669130    2.616691
## Durability   -1.56534073    5.645341
```

Now we need to generate profile plot for both types of CI.



And now we need to do the same for Bonferoni's CI.



`Conclusion` : From two profile plots, we can easily detect that

- All the criteria contain the value 0
- The range of CI for ‘Durability’ criteria is very large.

So, the **main reason** for the difference between Model 2 and Model 1 is **Durability.**

Problem No 3

A certain disease is characterized by fever, low blood pressure, and body aches. A pharmaceutical company is working on a new drug to treat this type of disease and wants to determine whether the drug is effective.

They take a random sample of 20 people with this type of disease and 18 with a placebo. The given data are imported and cleaned in the following step :

```r
##Problem 3
#Importing and cleaning the given data
Drug_data <- read.csv("C://Users//HP//Desktop//AP//ProblemSet_2//drug.csv")
which(is.na(Drug_data))
```

```
##  [1] 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88
```

```r
for (i in 1:ncol(Drug_data))
{
  Drug_data[,i] = as.numeric(Drug_data[,i])
}
```

```
## Warning: NAs introduced by coercion
## Warning: NAs introduced by coercion
## Warning: NAs introduced by coercion
## Warning: NAs introduced by coercion
## Warning: NAs introduced by coercion
## Warning: NAs introduced by coercion
```

```r
data <- Drug_data[c(-1,-2),]
names(data) <- Drug_data[2,]
row.names(data) <- c(1:20)
Drug <- data.frame(data[,c(1:3)])
Placebo <- data.frame(data[,c(5:7)])
Placebo <- na.omit(Placebo)
which(is.na(Drug))
```

```
## integer(0)
```

```r
which(is.na(Placebo))
```

```
## integer(0)
```

To determine whether the drug is effective at reducing the three symptoms, we need to perform *Multivariate two sample test for mean.*

(a) For the case of **equal** population covariance matrices for the two populations :

$$\bar{X_i} = \frac{1}{n_i}\sum X_{ij};\ S_i = \frac{1}{n_i - 1}\sum(X_{ij} - \bar{X_i})(X_{ij} - \bar{X_i})'$$

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

$$T^2 = (\bar{x_1} - \bar{x_2})^T[S_P(\frac{1}{n_1} + \frac{1}{n_2})]^{-1}(\bar{x_1} - \bar{x_2})$$

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)}T^2 \sim F_{p, n_1 + n_2 - p - 1}$$

Reject H_0 at level α if $F > F_{p,n_1+n_2-p-1,\alpha}$.

```
#(a)Population covariance matrix is equal
alpha <- 0.05
p <- ncol(Drug)
n1 <- nrow(Drug)
n2 <- nrow(Placebo)
x1bar <- apply(Drug, 2, mean)
x2bar <- apply(Placebo, 2, mean)
S1 <- cov(Drug)
S2 <- cov(Placebo)
Sp <- ((n1 - 1) * S1 + (n2 - 1) * S2)/(n1 + n2 - 1)

HT2 <- t(x1bar - x2bar) %*% solve(Sp * ((1/n1) + (1/n2))) %*% (x1bar - x2bar)
test_statistic <- (n1 + n2 - p - 1)/(p * (n1 + n2 - 2)) * HT2
test_statistic
```

```
##         [,1]
## [1,] 14.5073
```

```
critical_value <- qf(1-alpha, p, (n1 + n2 - p - 1))
critical_value
```

```
## [1] 2.882604
```

Now we can clearly see that test statistic exceeds the critical value. Therefore there is **enough evidence to reject** the null hypothesis at 5% level of significance. Here, the null hypothesis was *Two mean vectors (corresponding to disease and placebo) are equal.*

(b) For the case of **unequal** population covariance matrices for the two populations :

$$T^2 = (\bar{x_1} - \bar{x_2})'(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2)^{-1}(\bar{x_1} - \bar{x_2})$$

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)}T^2 \sim F_{p,\nu}$$

$$\frac{1}{\nu} = \sum_{i=1}^{2} \frac{1}{n_i - 1}[\frac{(\bar{x_1} - \bar{x_2})'S_T^{-1}(\frac{1}{n_i}S_i)S_T^{-1}(\bar{x_1} - \bar{x_2})}{T^2}]^2$$

where $S_T = \frac{1}{n_1}S_1 + \frac{1}{n_2}S_2$

Reject H_0 at level α if $F > F_{p,\nu,\alpha}$.

```
#(b)Population covariance matrix is not equal
St <- (S1 * (1/n1) + S2 * (1/n2))
HT2UN <- t(x1bar - x2bar) %*% solve(St) %*% (x1bar - x2bar)
test_statisticUN <- (n1 + n2 - p - 1)/(p * (n1 + n2 - 2)) * HT2UN
test_statisticUN
```

```
##          [,1]
## [1,] 13.97625
```

```
A <- t(x1bar - x2bar) %*% solve(St) %*% ((1/n1)*S1) %*% solve(St) %*% (x1bar - x2bar)
B <- t(x1bar - x2bar) %*% solve(St) %*% ((1/n2)*S2) %*% solve(St) %*% (x1bar - x2bar)
v <- (1/(n1-1))*(A/HT2UN)^2 + (1/(n2-1))*(B/HT2UN)^2
```

```
criticalUN <- qf(1-alpha, p, 1/v)
criticalUN
```

```
## [1] 2.877026
```

Here also, we can clearly see that test statistic exceeds the critical value. Therefore there is **enough evidence to reject** the null hypothesis at 5% level of significance. Here, the null hypothesis was *Two mean vectors (corresponding to disease and placebo) are equal.*

Problem No 4

A new type of corn seed has been developed by a team of agronomists who want to determine whether there was a significant difference between the types of soils that they are planted in **(loam, sandy, salty, clay)** based on the yield of the crop, amount of water required and amount of herbicide needed. Eight fields of each type were chosen for the analysis.

The given data are imported and cleaned in the following step :

```
##Problem 3
#Importing and cleaning the given data
soil_data <- read.csv("C://Users//HP//Desktop//AP//ProblemSet_2//soil.csv")
which(is.na(soil_data))
```

```
## integer(0)
```

```
names(soil_data)[1] <- 'SoilType'
```

(a) To determine whether there is a significant difference in the effects of the types of soil on the three variables, we need to perform MANOVA on the given data.

```
#performing MANOVA
model <- manova(cbind(yield, water, herbicide) ~ SoilType, data = soil_data)
summary(model, test = 'Pillai')
```

```
##           Df Pillai approx F num Df den Df  Pr(>F)
## SoilType   3 0.5345   2.0234      9     84 0.04641 *
## Residuals 28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model, test = 'Wilks')
```

```
##           Df   Wilks approx F num Df den Df  Pr(>F)
## SoilType   3 0.48941    2.405      9 63.428 0.02047 *
## Residuals 28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model, test = 'Roy')
```

```
##           Df     Roy approx F num Df den Df    Pr(>F)
## SoilType   3 0.94364   8.8073      3     28 0.0002844 ***
## Residuals 28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model, test = 'Hotelling-Lawley')
```

```
##           Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
## SoilType   3          0.99464    2.726      9     74 0.008399 **
## Residuals 28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results of the 4 tests, we can see that there is a **significant** difference in the **effects** of the **types of soil** on the three variables at 5% level of significance as the p-values are smaller than 0.05.

(b) As significant difference is present among the 4 levels of the factor, we are interested to perform more tests to answer different questions :

(i) First we need to perform MANOVA to check whether there is significant difference between the clay and salty groups.

```
#performing tests only on clay and salty
data1 <- soil_data[which(soil_data$SoilType == 'clay' | soil_data$SoilType == 'salty'),]
md1 <- manova(cbind(yield, water, herbicide) ~ SoilType, data = data1)
summary(md1, test = 'Pillai')
```

```
##           Df   Pillai approx F num Df den Df Pr(>F)
## SoilType   1 0.041866  0.17478      3     12 0.9114
## Residuals 14
```

```
summary(md1, test = 'Wilks')
```

```
##           Df   Wilks approx F num Df den Df Pr(>F)
## SoilType   1 0.95813  0.17478      3     12 0.9114
## Residuals 14
```

```
summary(md1, test = 'Roy')
```

```
##           Df      Roy approx F num Df den Df Pr(>F)
## SoilType   1 0.043696  0.17478      3     12 0.9114
## Residuals 14
```

```
summary(md1, test = 'Hotelling-Lawley')
```

```
##           Df Hotelling-Lawley approx F num Df den Df Pr(>F)
## SoilType   1         0.043696  0.17478      3     12 0.9114
## Residuals 14
```

From the results of 4 tests, we can find that the p-values are very large. So there is **not enough evidence to reject** the null hypothesis (mean vectors corresponding to clay and salty are equal) at 5% level of significance.

(ii) Now we need to perform MANOVA to check whether there is significant difference between the loam and sandy groups.

```
#performing tests only on loam and sandy
data2 <- soil_data[which(soil_data$SoilType == 'loam' | soil_data$SoilType == 'sandy'),]
md2 <- manova(cbind(yield, water, herbicide) ~ SoilType, data = data2)
summary(md2)
```

```
##           Df  Pillai approx F num Df den Df  Pr(>F)
## SoilType   1 0.45579   3.3501      3     12 0.05554 .
## Residuals 14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(md2, test = 'Wilks')
```

```
##           Df   Wilks approx F num Df den Df  Pr(>F)
## SoilType   1 0.54421   3.3501      3     12 0.05554 .
## Residuals 14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(md2, test = 'Roy')
```

```
##           Df     Roy approx F num Df den Df  Pr(>F)
## SoilType   1 0.83753   3.3501      3     12 0.05554 .
## Residuals 14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(md2, test = 'Hotelling-Lawley')
```

```
##           Df Hotelling-Lawley approx F num Df den Df  Pr(>F)
## SoilType   1          0.83753   3.3501      3     12 0.05554 .
## Residuals 14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results of 4 tests, we can find that the p-values are larger that 0.05. So there is **not enough evidence to reject** the null hypothesis (mean vectors corresponding to loam and sandy are equal) at 5% level of significance.

(iii) Now we need to perform MANOVA to check whether there is significant difference between **the clay and salty** & **the loam and sandy**.

```
#performing tests for clay and salty VS loam and sandy
data3 <- soil_data
data3[which(soil_data$SoilType == 'clay' | soil_data$SoilType == 'salty'), 1] <- 'Group1'
data3[which(soil_data$SoilType == 'loam' | soil_data$SoilType == 'sandy'), 1] <- 'Group2'
md3 <- manova(cbind(yield, water, herbicide) ~ SoilType, data = data3)
summary(md3)
```

```
##           Df  Pillai approx F num Df den Df   Pr(>F)
## SoilType   1 0.34056   4.8201      3     28 0.007893 **
## Residuals 30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(md3, test = 'Wilks')
```

```
##           Df   Wilks approx F num Df den Df   Pr(>F)
## SoilType   1 0.65944   4.8201      3     28 0.007893 **
## Residuals 30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(md3, test = 'Roy')
```

```
##           Df     Roy approx F num Df den Df   Pr(>F)
## SoilType   1 0.51644   4.8201      3     28 0.007893 **
## Residuals 30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(md3, test = 'Hotelling-Lawley')
```

```
##           Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
## SoilType   1          0.51644   4.8201      3     28 0.007893 **
## Residuals 30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results of 4 tests, we can find that the p-values are very small. So there is **enough evidence to reject** the null hypothesis (mean vectors corresponding to (clay and salty) & (loam and sandy) are equal) at 5% level of significance.

As there is significant difference, we need to obtain the simultaneous and Bonferoni's CIs.

Here we make two groups namely Group1 (for loam and sandy) & Group 2 (for clay and salty). After that we find simultaneous and Bonferoni's CI using similar method as for two sample Multivariate data.

```
##finding 95% simultaneous and bonferoni CI for individual means
Group1 <- data.matrix(data3[1:16,c(-1)])
Group2 <- data.matrix(data3[17:32,c(-1)])
Y <- Group1 - Group2
Ybar <- apply(Y, 2, mean)
S <- cov(Y)
p <- ncol(Y)
n <- nrow(Y)
alpha <- 0.01

#Calculating Simultaneous Interval
MargSC <- sqrt(p * (n-1) * qf(1-alpha, p, (n-p)) / (n-p)) * sqrt(diag(S) / n)
MargSC <- as.matrix(MargSC)
SCI <- data.frame('lower.bound' = as.vector(Ybar - MargSC),
                  'upper bound' = Ybar + MargSC)
SCI
```

```
##           lower.bound upper.bound
## yield       -8.264298    26.40180
## water      -16.019515    13.23201
## herbicide   -2.422880     5.21038
```

```
#Calculating Bonferoni Interval
MargBI <- qt(1-(alpha / (2*p)), n-1) * sqrt(diag(S) / n)
MargBI <- as.matrix(MargBI)
BI <- data.frame('lower bound' = as.vector(Ybar - MargBI),
                 'upper bound' = Ybar + MargBI)
BI
```

```
##           lower.bound upper.bound
## yield       -4.478331   22.615831
## water      -12.824885   10.037385
## herbicide   -1.589233    4.376733
```

From both CIs, we can see that every variables contain 1 but range of yield and water is pretty high.

So **yield and water** are the reasons for the significant difference between two groups.

Problem No 5

A study was conducted to see the impact of social-economic class (rich, middle, poor) and gender (male, female) on kindness and optimism using a sample of 24 people based on the data.

The given data is imported and cleaned in the following part :

```
##Problem5

#Importing and cleaning data
study <- read.csv("C://Users//HP//Desktop//AP//ProblemSet_2//study.csv")
which(is.na(study))
```

```
## integer(0)
```

For the significance of the differences among the main effects of the two factors as well as their interactions on the two variables, we need to perform MANOVA two way layout test.

```
#performing two way MANOVA
model <- manova(cbind(kindness, optimism) ~ gender + economic + gender*economic, data = study)

#test results
summary(model, test = 'Pillai')
```

```
##                 Df  Pillai approx F num Df den Df   Pr(>F)   
## gender           1 0.41175   5.9496      2     17 0.010997 * 
## economic         2 0.51728   3.1399      4     36 0.025881 * 
## gender:economic  2 0.70379   4.8866      4     36 0.002985 **
## Residuals       18                                           
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model, test = 'Wilks')
```

```
##                 Df   Wilks approx F num Df den Df   Pr(>F)   
## gender           1 0.58825   5.9496      2     17 0.010997 * 
## economic         2 0.50412   3.4716      4     34 0.017562 * 
## gender:economic  2 0.38703   5.1630      4     34 0.002325 **
## Residuals       18                                           
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model, test = 'Roy')
```

```
##                 Df     Roy approx F num Df den Df   Pr(>F)   
## gender           1 0.69995   5.9496      2     17 0.010997 * 
## economic         2 0.89368   8.0431      2     18 0.003193 **
## gender:economic  2 1.14396  10.2957      2     18 0.001045 **
## Residuals       18                                           
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model, test = 'Hotelling-Lawley')
```

```
##                 Df Hotelling-Lawley approx F num Df den Df   Pr(>F)   
## gender           1          0.69995   5.9496      2     17 0.010997 * 
## economic         2          0.94119   3.7648      4     32 0.012790 * 
## gender:economic  2          1.34909   5.3964      4     32 0.001948 **
## Residuals       18                                                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results of 4 tests, we can see that p-value is very small for every factors and their interaction. So we can conclude that among gender, economic factors, and their interaction have significant effect on the two variables.

From the p-values, the interaction effect has more significant effect than the two factors.