Assignment:2

732A97 Multivariate Statistical Methods,Linkoping University

25 December 2019

Question 1: Test of outliers

Consider again the data set from the T1-9.dat file, National track records for women. In the first assignment we studied different distance measures between an observation and the sample average vector. The most common multivariate residual is the Mahalanobis distance and we computed this distance for all observations. a) The Mahalanobis distance is approximately chi~square distributed, if the data comes from a multivariate normal distribution and the number of observations is large. Use this chi~square approximation for testing each observation at the 0.1% significance level and conclude which countries can be regarded as outliers. Should you use a multiple testing correction procedure? Compare the results with and without one. Why is (or maybe is not) 0.1% a sensible significance level for this task?
b) One outlier is North Korea. This country is not an outlier with the Euclidean distance. Try to explain these seemingly contradictory results.

```
## The outlier countries are for 0.1% are KORN PNG SAM

## The outlier countries are for 5% are COK KORN MEX PNG SAM
```

The Outlier are reduced when 0.1% is used compared to 5%. So, 0.1% is sensible significance value for this task.
b) In terms of equation, the euclidean distance is same as the Mahalanobis distance provided the covaiance matrix is identity matrix. Therefore, the Mahalanobis distance takes covaiance terms into consideration while finding outliers. This explains why North KOrea is outlier in Mahalanobis distance but not the euclidean distance.

Question 2: Test, confidence region and confidence intervals for a mean vector

Look at the bird data in file T5-12.dat and solve Exercise 5:20 of Johnson, Wichern. Do not use any extra R package or built{in test but code all required matrix calculations. You MAY NOT use loops!
(a) Find and sketch the 95% confidence ellipse for the population means and Suppose it is known that $\mu_1 = 190mm$ *and* $\mu_2 = 275mm$ *for male hook-billed kites. Are these plausible values for the mean tail length and mean wing length for the female birds? Explain.*

```
## The calculated T-square value is 5.543134

## The Actual T-square value is 6.578471
```

The calculated T^2 is more than actual T^2, so the null hypothesis cannot be rejected. Therefore the mean vector [190,275] are plausible for the female bird.The T^2 can be calculated by the formula

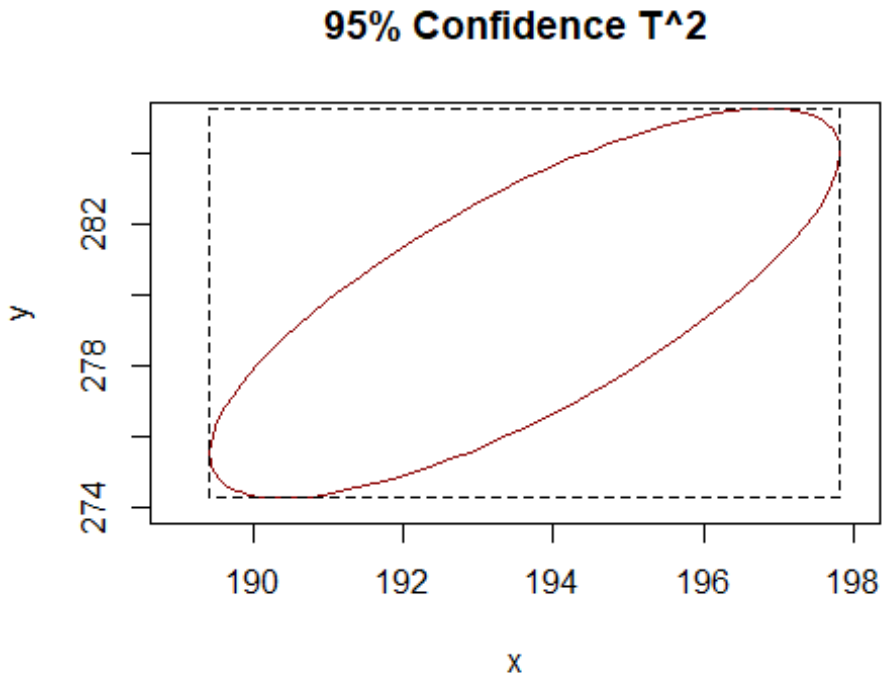
$$T^2 < \frac{(n-1)p}{(n-p)} F_{n,n-p}$$

(b) Construct the simultaneous 95% -intervals for and and the 95% Bonferroni intervals for and Compare the two sets of intervals.What advantage, if any, do the -intervals have over the Bonferroni intervals?

Simultaneous Intervals

```
## The simultaneous confidence intervals of tail length are 189.4217 197.8227
```

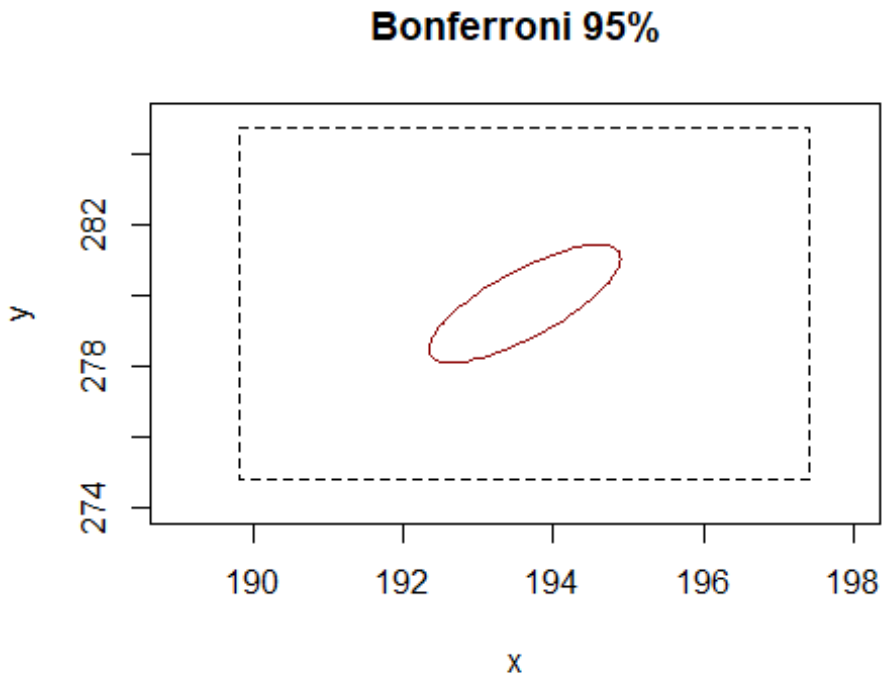
```
## The simultaneous confidence intervals of wing length are 274.2564 285.2992
```



Bonferroni Interval

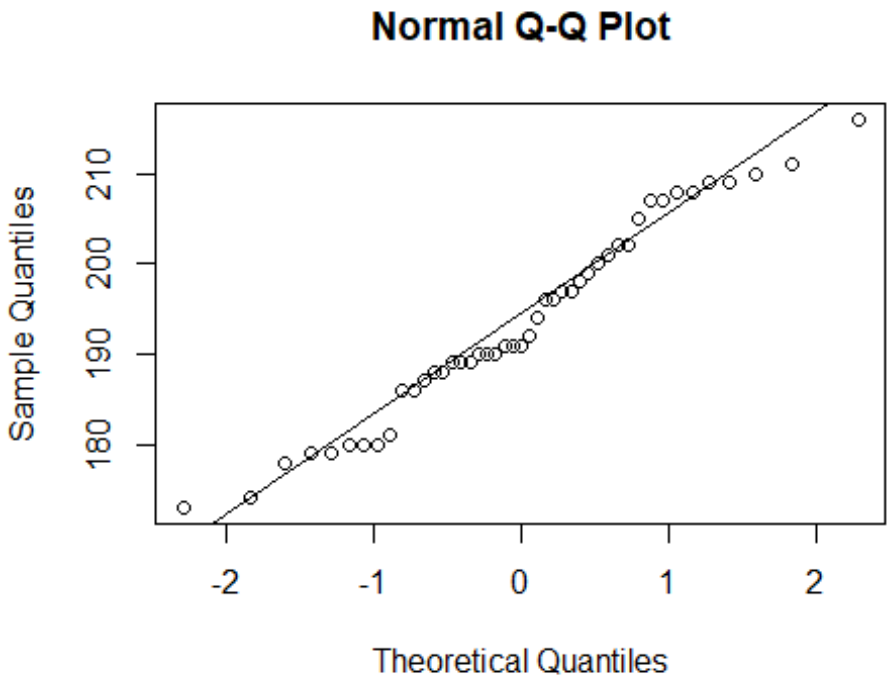
```
## The Bonferroni confidence intervals of tail length are 189.8216 197.4229
```

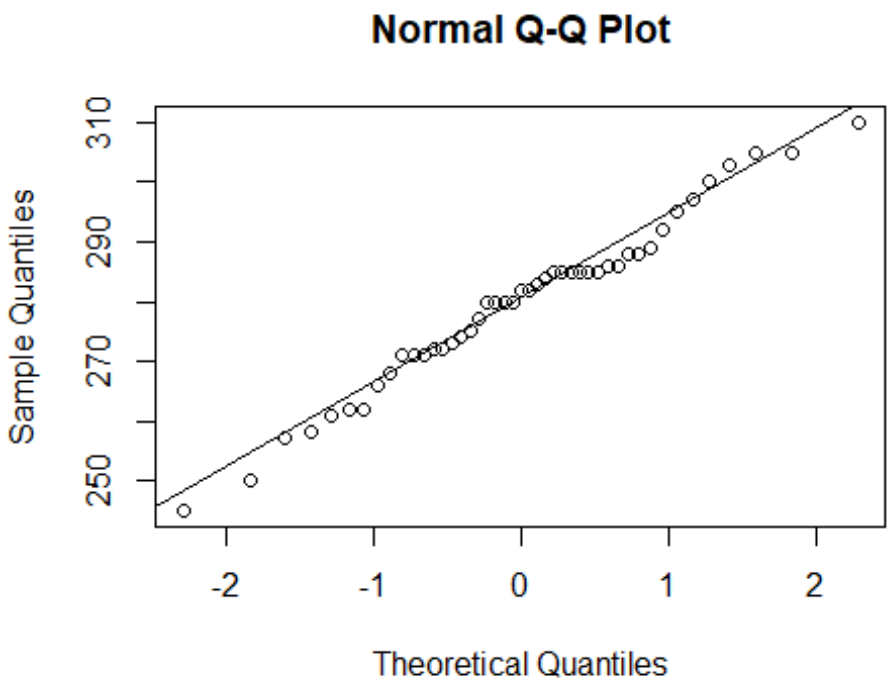
```
## The Bonferroni confidence intervals of wing length are 274.7819 284.7736
```

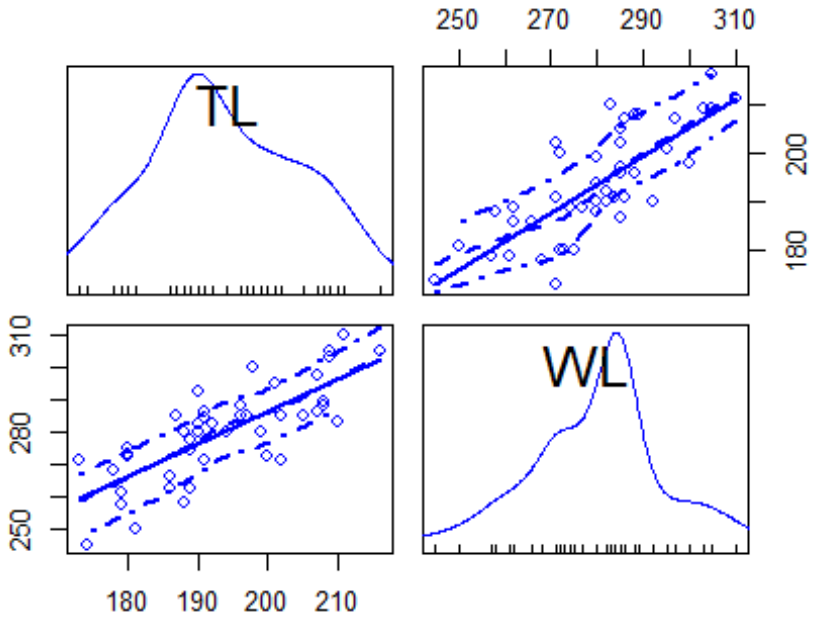


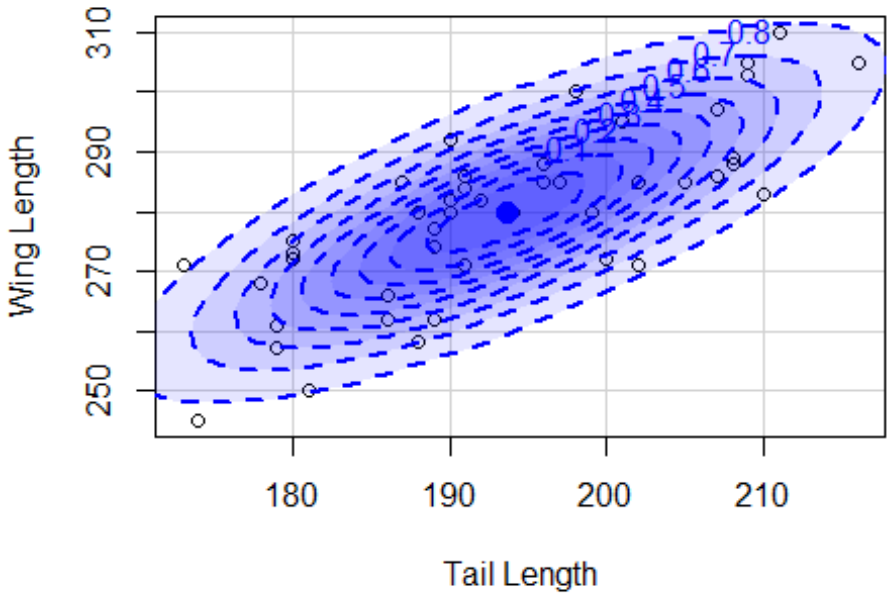
For this data, the bonferroni Intervals is a subset of Simultaneous Intervals. So, If the mean lie between bonferroni Intervals then it will also come between Simultaneous Intervals. So Bonferroni Intervals since it is narrowed can be used.

(c) Is the bivariate normal distribution a viable population model? Explain with reference to Q~Q plots and a scatter diagram.


Normal Q-Q Plot
Sample Quantiles
180
190
200
210
-2
-1
0
1
2
Theoretical Quantiles


Normal Q-Q Plot
Sample Quantiles
250
270
290
310
-2
-1
0
1
2
Theoretical Quantiles





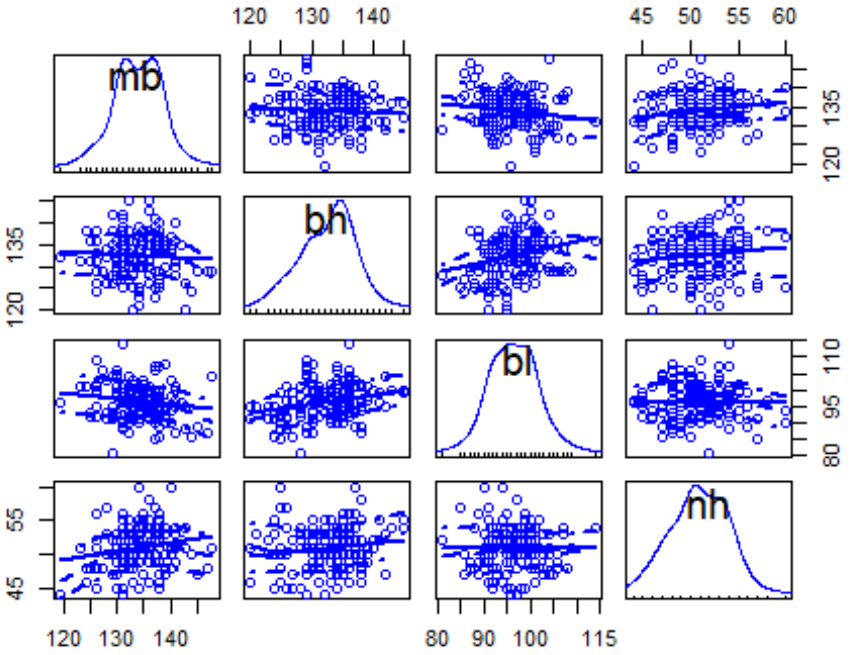
From the graphs, we can conclude that the data is normally distributed and it is a viable distribution for this data.

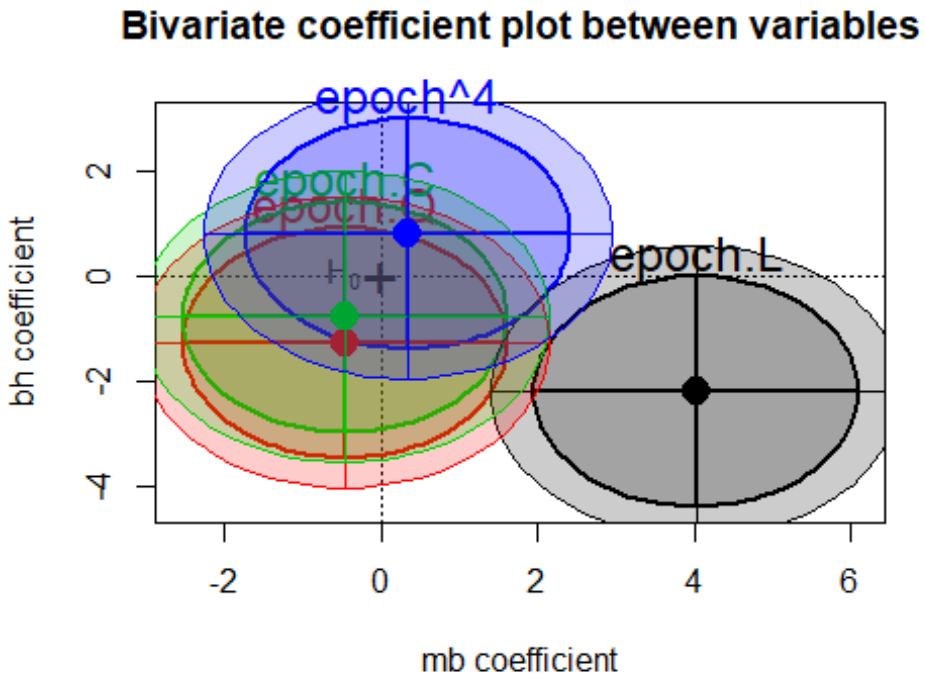
Question 3: Comparison of mean vectors (oneway MANOVA)

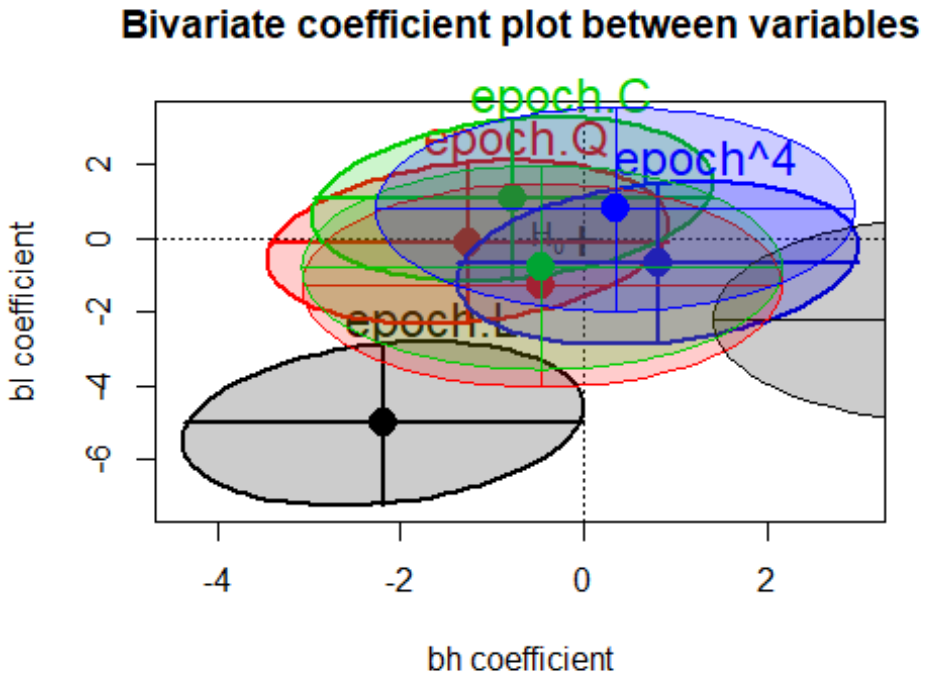
We will look at a data set on Egyptian skull measurements (published in 1905 and now in heplots R package as the object Skulls). Here observations are made from five epochs and on each object the maximum breadth (mb), basibregmatic height (bh), basialiveolar length (bl) and nasal height (nh) were measured.

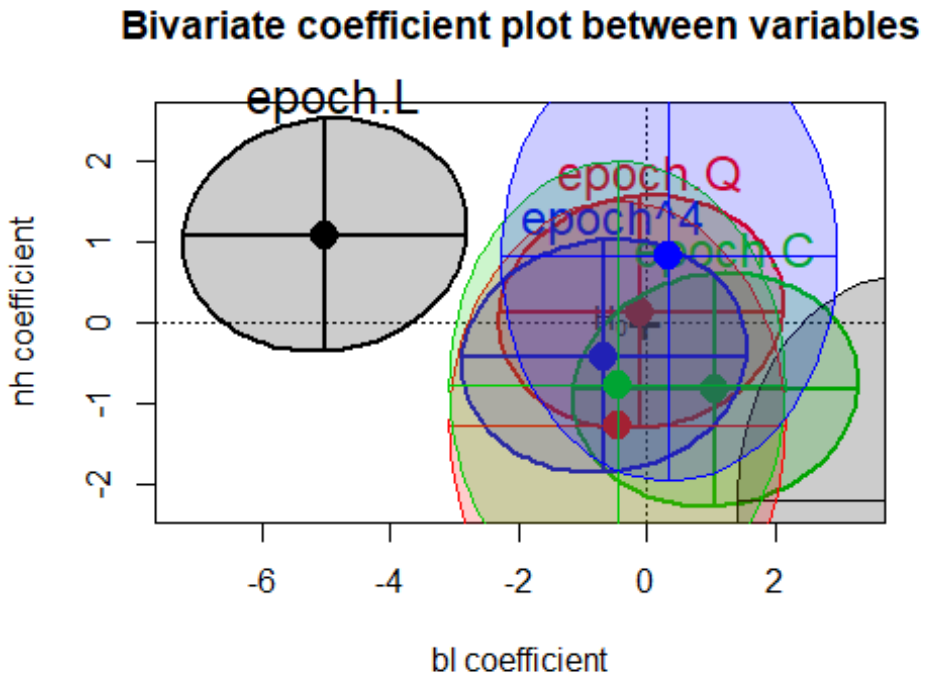
a) Explore the data first and present plots that you find informative.

```
## Call:
##    manova(cbind(mb, bh, bl, nh) ~ as.factor(epoch), data = skull)
##
## Terms:
##                 as.factor(epoch) Residuals
## mb                       502.827  3061.067
## bh                       229.907  3405.267
## bl                       803.293  3505.967
## nh                        61.200  1472.133
## Deg. of Freedom                4       145
##
## Residual standard errors: 4.59465 4.846091 4.917223 3.186321
## Estimated effects are balanced

## The Mean value for different variable are 133.9733 132.5467 96.46 50.93333

## The standard Deviation for different variables are 4.89068 4.939346
5.377844 3.207932
```


mb
bh
bl
nh
120 130 140
45 50 55 60
120 135
120 135
80 95 110
45 55
120 130 140
80 90 100 115


Bivariate coefficient plot between variables
epoch^4
epoch.C
epoch.S
epoch.L
bh coefficient
mb coefficient
2
0
-2
-4
-2
0
2
4
6





b) Now we are interested whether there are differences between the epochs. Do the mean vectors differ? Study this question and justify your conclusions.

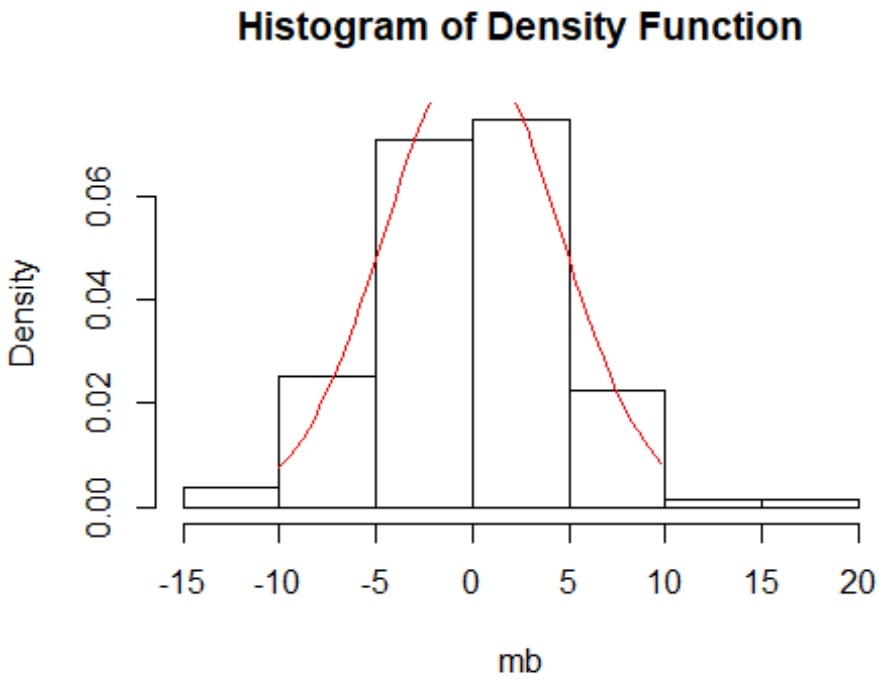
```
##                   Df  Pillai approx F num Df den Df    Pr(>F)
## as.factor(epoch)   4 0.35331    3.512     16    580 4.675e-06 ***
## Residuals        145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

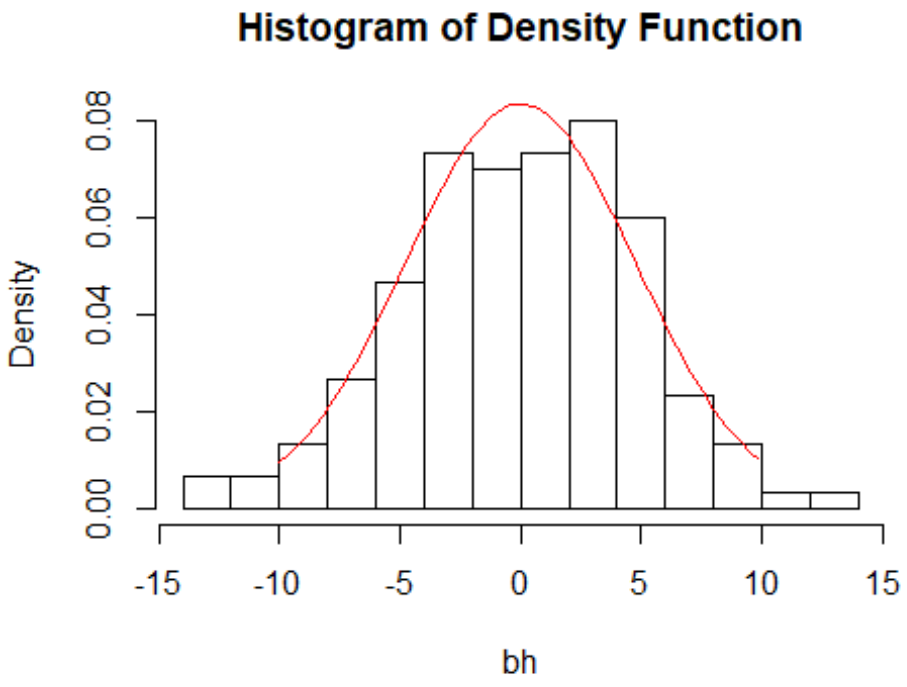
The pvalue is less than the significance value, Null hypothesis cannot be rejected. so the mean vectors does not differ for these parameters.

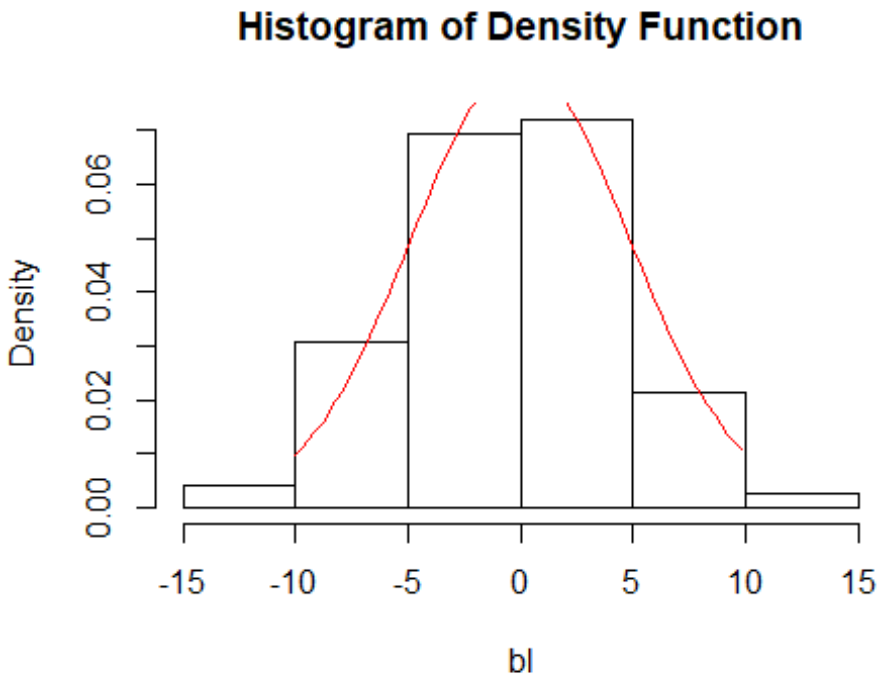
c) If the means differ between epochs compute and report simultaneous confidence intervals.Inspect the residuals whether they have mean 0 and if they deviate from normality (graphically).

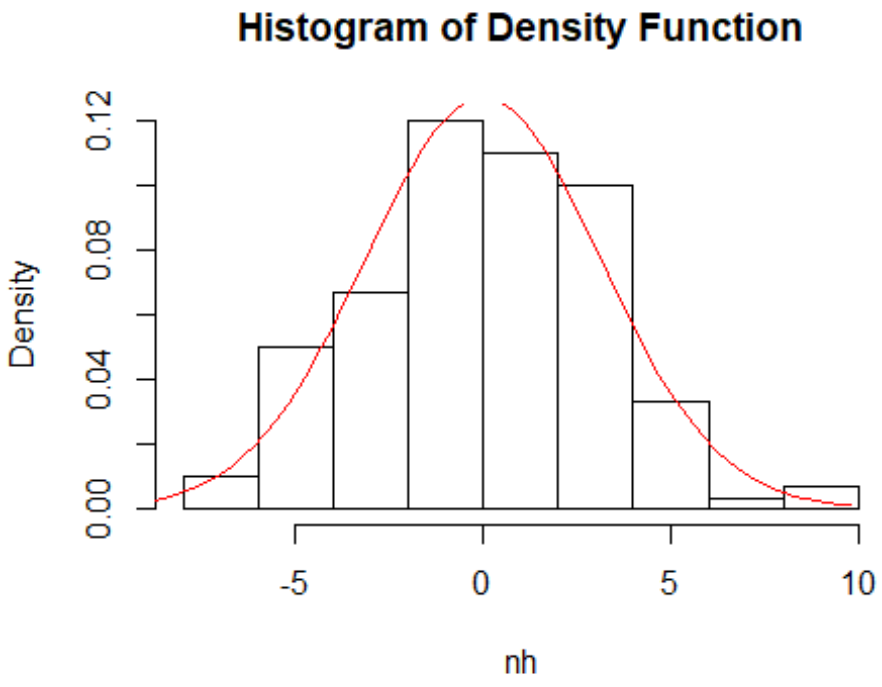
```
##  Response 1 :
##                          Df  Sum Sq Mean Sq F value    Pr(>F)
## as.factor(skull$epoch)   4  502.83 125.707  5.9546 0.0001826 ***
## Residuals              145 3061.07  21.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response 2 :
##                          Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(skull$epoch)   4  229.9  57.477  2.4474 0.04897 *
## Residuals              145 3405.3  23.485
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response 3 :
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(skull$epoch)   4  803.3 200.823  8.3057 4.636e-06 ***
## Residuals              145 3506.0  24.179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response 4 :
##                          Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(skull$epoch)   4   61.2  15.300   1.507 0.2032
## Residuals              145 1472.1  10.153

## The simultaneous Intervals for the variables are

##         left    right
## mb 132.71470 135.2320
## bh 131.27551 133.8178
## bl  95.07600  97.8440
## nh  50.10776  51.7589
```

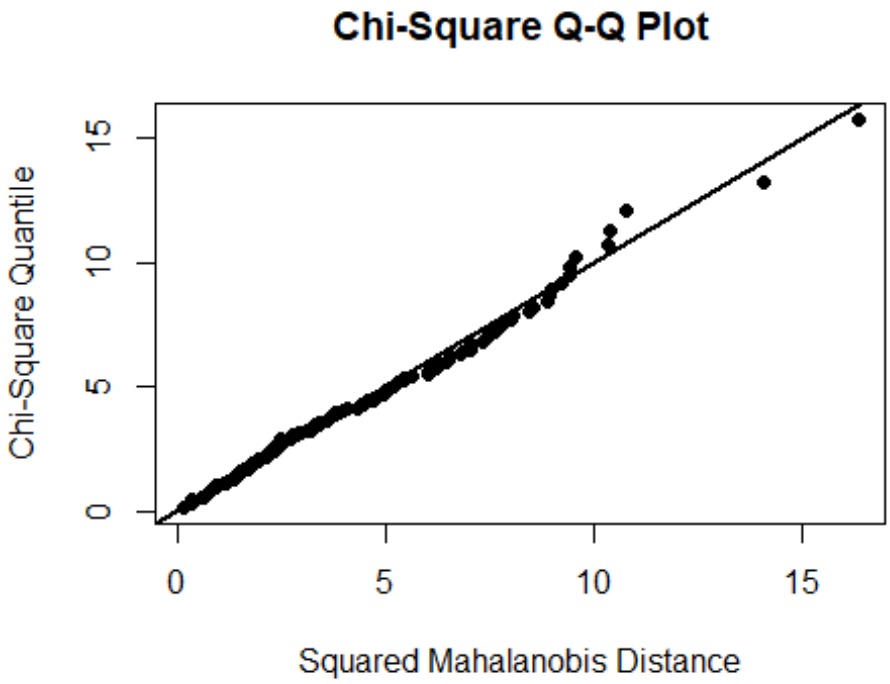
The pvalue for bh,bl,mb the pvalue is less than the significance value, Null hypothesis cannot be rejected. so the mean vectors does not differ for these parameters. The pvalue for nh is more than the significance value, therefore the null hypothesis is rejected.


Histogram of Density Function
Density
0.00
0.02
0.04
0.06
-15
-10
-5
0
5
10
15
20
mb


Histogram of Density Function
Density
0.00
0.02
0.04
0.06
0.08
-15
-10
-5
0
5
10
15
bh


Histogram of Density Function
Density
0.00
0.02
0.04
0.06
-15
-10
-5
0
5
10
15
bl


Histogram of Density Function
Density
0.00
0.04
0.08
0.12
-5
0
5
10
nh




```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  19.2570567874466 0.505177925496235    YES
## 2 Mardia Kurtosis 0.267418576192474 0.789146898351176    YES
## 3             MVN              <NA>              <NA>    YES
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    mb        0.9933    0.7116    YES
## 2 Shapiro-Wilk    bh        0.9936    0.7493    YES
## 3 Shapiro-Wilk    bl        0.9956    0.9347    YES
## 4 Shapiro-Wilk    nh        0.9933    0.7189    YES
##
## $Descriptives
##      n          Mean  Std.Dev      Median        Min       Max      25th
## mb 150 -4.107617e-17 4.532557  0.16666667 -12.366667 15.633333 -2.500000
## bh 150  2.177714e-16 4.780600  0.20000000 -12.600000 12.300000 -3.200000
## bl 150 -8.890458e-17 4.850771 -0.03333333 -12.500000 14.833333 -3.166667
## nh 150 -2.568746e-17 3.143261  0.03333333  -7.366667  9.433333 -2.166667
##        75th        Skew    Kurtosis
## mb 2.633333  0.06539367  0.37846150
## bh 3.375000 -0.14739003 -0.14947035
## bl 3.500000  0.08996086 -0.05326603
## nh 2.433333  0.07413299 -0.16714612
```

There is no skewness in the graphs and therefore residuals remains in normality and the $\mu = 0$. From the mvn, we got a table which contains the normality test result showing that $\mu = 0$ for all the variable.

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(heplots)
library(car)
library(robustbase)
library(investr)
library(ggplot2)
library(ellipse)
library(DescTools)
library(matlib)
library(MVN)
sf <- read.csv(file.choose(),header = F)
sf1 <- read.csv(file.choose(),header = F,sep = '',as.is = T)
skull <- Skulls

colnames(sf) <-
c("Countries","X100M","X200M","X400M","X800M","X1500M","X3000M","Marathon")
centered=scale(x = as.matrix(sf[,2:8]), center = TRUE, scale = FALSE)
covariance_mat=cov(sf[,2:8])
mahalnob<-centered%*%solve(covariance_mat)%*%t(centered)
diagonal<-diag(mahalnob)
names(diagonal)<-sf[,1]
sf$distance <- diagonal
extrems=order(diagonal, decreasing=TRUE)[1:5]
x <- as.vector(sf$distance)
sf["p_value"] <- pchisq(x, df = 7)
cat("The outlier countries are for 0.1%
are",as.character(sf[which(sf["p_value"] > 0.999),"Countries"]),"\n")
cat("The outlier countries are for 5%
are",as.character(sf[which(sf["p_value"] > 0.95),"Countries"]),"\n")
pvalue <- sf$p_value
sf["p_value_adjusted"] <- p.adjust(pvalue,method = c("holm", "hochberg",
"hommel", "bonferroni", "BH", "BY", "fdr", "none"), n = length(pvalue))
p<-2
n<-length(sf1[,1])
f<-qf(0.05, df1 = p, df2 = n-p , lower.tail = F)
x1 <- mean(sf1[,1])
x2 <- mean(sf1[,2])
u1 <- c(x1,x2)
u2 <- c(190,275)
s12 <- (sum(((sf1$V1 - x1)*(sf1$V2-x2)))/(nrow(sf1)-1))
s11 <- (sum(((sf1$V1 - x1)^2))/(nrow(sf1)-1))
s22 <- (sum(((sf1$V2 - x2)^2))/(nrow(sf1)-1))
S <- matrix(c(s11,s12,s12,s22),nrow = 2)
```

```
S_inverse <- inv(S)
T_square_calculated <- nrow(sf1)*((t(u1-u2))%*%S_inverse%*%(u1-u2))
cat("The calculated T-square value is",T_square_calculated,"\n")
T_square_actual <- (((n-1)*p)/(n-p))*f
cat("The Actual T-square value is",T_square_actual,"\n")
lambda<-eigen(S)
colnames(sf1)<-c("TL","WL")
left<-c()
right<-c()
left[1]<-u1[1]-sqrt((p*(n-1)/(n-p))*f*(S[1,1]/n))
right[1]<-u1[1]+sqrt((p*(n-1)/(n-p))*f*(S[1,1]/n))
left[2]<-u1[2]-sqrt((p*(n-1)/(n-p))*f*(S[2,2]/n))
right[2]<-u1[2]+sqrt((p*(n-1)/(n-p))*f*(S[2,2]/n))
intervals<-t(rbind(left,right))
rownames(intervals)<-c("TL","WL")
cat("The simultaneous confidence intervals of tail length
are",intervals[1,],"\n")
cat("The simultaneous confidence intervals of wing length
are",intervals[2,],"\n")

plot(ellipse(S,centre=u1,t=sqrt(((n-1)*p/(n*(n-p)))*qf(0.95,p,n-
p))),col="dark red",type="l",xlim=c(189,198),ylim=c(274,285),main="95%
Confidence T^2")
rect(xleft = 189.4217,ybottom = 274.2564,xright = 197.8227,ytop =
285.2992,lty=2)

colnames(sf1)<-c("TL","WL")
p<-2
n<-length(sf1[,1])
tval<-qt(0.05/(2*p),df=n-1)

right1<-u1[1] - (tval*sqrt((S[1,1]/n)))
left1<-u1[1] + (tval*sqrt((S[1,1]/n)))
right2<-u1[2]- (tval*sqrt((S[2,2]/n)))
left2<-u1[2] + (tval*sqrt((S[2,2]/n)))
intervals<-matrix(c(left1,left2,right1,right2),nrow=2)
rownames(intervals)<-c("TL","WL")
colnames(intervals)<-c("left","right")
cat("The Bonferroni confidence intervals of tail length
are",intervals[1,],"\n")
cat("The Bonferroni confidence intervals of wing length
are",intervals[2,],"\n")
plot(ellipse(S,centre=u1,t=tval*(0.05/2*p)),col="dark
red",type="l",xlim=c(189,198),ylim=c(274,285),main="Bonferroni 95%")
rect(xleft = 189.8216,ybottom = 274.7819,xright = 197.4229,ytop =
284.7736,lty=2)
qqnorm(sf1$TL)
qqline(sf1$TL, datax = FALSE, distribution = qnorm,
       probs = c(0.25, 0.75))
qqnorm(sf1$WL)
```

```r
qqline(sf1$WL, datax = FALSE, distribution = qnorm,
       probs = c(0.25, 0.90))
scatterplotMatrix(sf1)
dataEllipse(sf1$TL, sf1$WL, levels=0.1*1:9,
    ellipse.label=0.1*1:9, lty=2, fill=TRUE, fill.alpha=0.1,xlab = "Tail
Length",ylab = "Wing Length")
epoch.manova<-manova(cbind(mb,bh,bl,nh)~as.factor(epoch),data=skull)
mean <- colMeans(skull[,-1])
print(epoch.manova)
cat("The Mean value for different variable are",as.matrix(mean),"\n")
cat("The standard Deviation for different variables are",sapply(skull[,-
1],sd),"\n")
scatterplotMatrix(skull[,-1])
mod <- lm(cbind(mb,bh,bl,nh)~epoch, data=skull)
for (i in 1:3){
coefplot(mod, variables = i:(i+1),lwd=2, main="Bivariate coefficient plot
between variables", fill=TRUE)
coefplot(mod, add=TRUE, Scheffe=TRUE, fill=TRUE)
}
summary(epoch.manova)
summary(aov(cbind(skull$mb,skull$bh,skull$bl,skull$nh)~as.factor(skull$epoch)
))
#Simultaneous CI
p<-4
n<-length(skull[,1])
m<-colMeans(skull[,-1])
s<-cov(as.matrix(skull[,-1]))
f<-qf(0.05, df1 = p, df2 = n-p , lower.tail = F)
left<-c()
right<-c()
 for(i in 1:4){

  left[i]<-m[i]-sqrt((p*(n-1)/(n-p))*f*(s[i,i]/n))
  right[i]<-m[i]+sqrt((p*(n-1)/(n-p))*f*(s[i,i]/n))

  }
 intervals<-t(rbind(left,right))
 rownames(intervals)<-c("mb","bh","bl","nh")
 cat("The simultaneous Intervals for the variables are","\n")
 intervals
 r <- as.data.frame(epoch.manova$residuals)
 z <- colnames(r[,1:4])
 for(i in 1:4){
  hist(r[,i], freq = FALSE,xlab = z[i],main = "Histogram of Density
Function")
  x <- seq(-10, 10, 0.3)
  y <- with(skull, dnorm(x, mean(r[,i]), sd(r[,i])))
  lines(x, y, col = "red")
 }
 mvn(r,multivariatePlot = "qq")
```