

Pass. Congratulation!

Multivariate Statistical Methods

Assignment 2. Inference about mean vectors

*Ahmet Hakan Akdeve(ahmak554), Jooyoung Lee(joole336), Weng Hang Wong(wonwo535),
Zhixuan Duan(zhidu838)*

2019 12 01

Question 1.

A)

Outlier detection using mahalanobis distances.

Table 1: Outliers

| | Country | d2values | pvalues |
|----|---------|-----------|-----------|
| 1 | ARG | 6.650003 | 0.0099156 |
| 5 | BER | 7.630567 | 0.0057387 |
| 11 | COK | 19.834001 | 0.0000084 |
| 13 | CZE | 10.901456 | 0.0009609 |
| 20 | GRE | 9.540322 | 0.0020101 |
| 29 | KEN | 7.625466 | 0.0057550 |
| 30 | KORS | 8.034918 | 0.0045884 |
| 31 | KORN | 26.167141 | 0.0000003 |
| 32 | LUX | 11.108846 | 0.0008592 |
| 33 | MAS | 8.237149 | 0.0041042 |
| 34 | MRI | 6.664985 | 0.0098325 |
| 35 | MEX | 14.230932 | 0.0001617 |
| 39 | NOR | 6.888063 | 0.0086773 |
| 40 | PNG | 30.507248 | 0.0000000 |
| 41 | PHI | 9.065884 | 0.0026042 |
| 46 | SAM | 35.014063 | 0.0000000 |
| 47 | SIN | 8.016395 | 0.0046356 |
| 51 | TPE | 10.183996 | 0.0014166 |
| 52 | THA | 7.815219 | 0.0051808 |
| 53 | TUR | 8.045368 | 0.0045620 |
| 54 | USA | 9.155697 | 0.0024794 |

The table shows 9 countries that could be assumed as outliers. Most countries that were classified as outliers with different distance functions from previous lab can be found on this list such as “COK”.

In your code, the degree of freedom of the chi squared distribution is set to one, however, Result 4-7 in the text book states that it should be the number of columns.

Table 2: Outliers Adjusted p-values

| | Country | d2values | pvalues | adj_pvalues |
|----|---------|----------|-----------|-------------|
| 11 | COK | 19.83400 | 0.0000084 | 0.0004561 |
| 31 | KORN | 26.16714 | 0.0000003 | 0.0000169 |
| 35 | MEX | 14.23093 | 0.0001617 | 0.0087313 |
| 40 | PNG | 30.50725 | 0.0000000 | 0.0000018 |
| 46 | SAM | 35.01406 | 0.0000000 | 0.0000002 |

The table above shows the result of adjusted p-values and the countries that still remain as outliers. The p-values are adjusted for a multiple-testing correction using Bonferroni correction. Multiple testing correction adjusts the individual p-value for each observation to keep the overall error rate to less than or equal to the user-specified p-cutoff value. Bonferroni correction takes each value and multiply by the number of observations. If corrected p-value is still below cutoff than the observation is significant.

B)

Euclidean distance assumes data to be Gaussian and treat each feature equally. Mahalanobis distance measures the correlation between variable and this can lead to different conclusions about the features.

We do not need multiple testing correction to control the error rate. We are only performing one hypothesis testing on each data point. Think about it this way: it does not matter if you have a 10-million-row or 1000-row data set, if these data perfectly normal, you will still have about 0.1% of the data points identified as outliers using this test (because of Result 4-7). So you have perfect control on how much data you want to remove by controlling alpha. If you have a 10-million-row data, for example, you can always say "I don't think there are 10000 (that is 0.1%) outliers; I think only 100 of them are", and go ahead to set alpha = 0.001%. The moment you set your new alpha, you know your false discovery rate is about 0.001% if the rest of your data were normal enough. So there is no need to force yourself to divide your alpha by 10-million: divide or not, you still have full control.

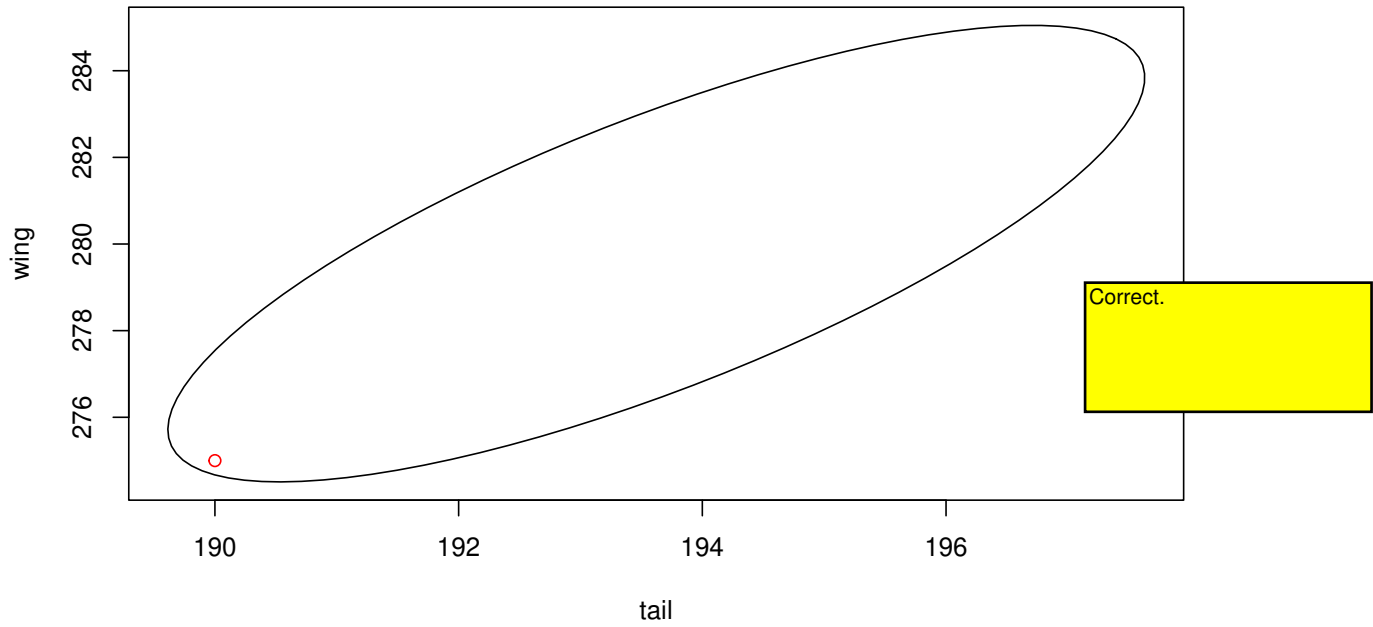
On the other hand, if you t-test each column separately without Bonferroni correction, by setting alpha = 0.1%, you cannot say that about 0.1% of data points will be removed anymore: in other words, in this case, you lost control on your false discovery rates. This is why univariate t-testings on multiple columns independently require Bonferroni correction.

QUESTION 2

A)

Table 3: Test value and the critical value

| | Value |
|------------|----------|
| Test value | 5.543130 |
| Crit value | 6.578471 |



In the table above presents the results of the calculated test statistics and the critical value. Since the test-statistic is lower than the critical F-value means that the new mean values is in the confidence region and is therefore plausible values. That conclusion is also confirmed by the ellipse above which is representing the confidence region. The highlighted red circle is the tested mean vector and it is within the confidence region which should mean that that values are plausible.

B)

Table 4: T2 confidence intervals

| | Lower | Upper |
|------|----------|----------|
| Tail | 189.4217 | 197.8227 |
| Wing | 274.2564 | 285.2992 |

Correct.

Table 5: Bonferroni intervals

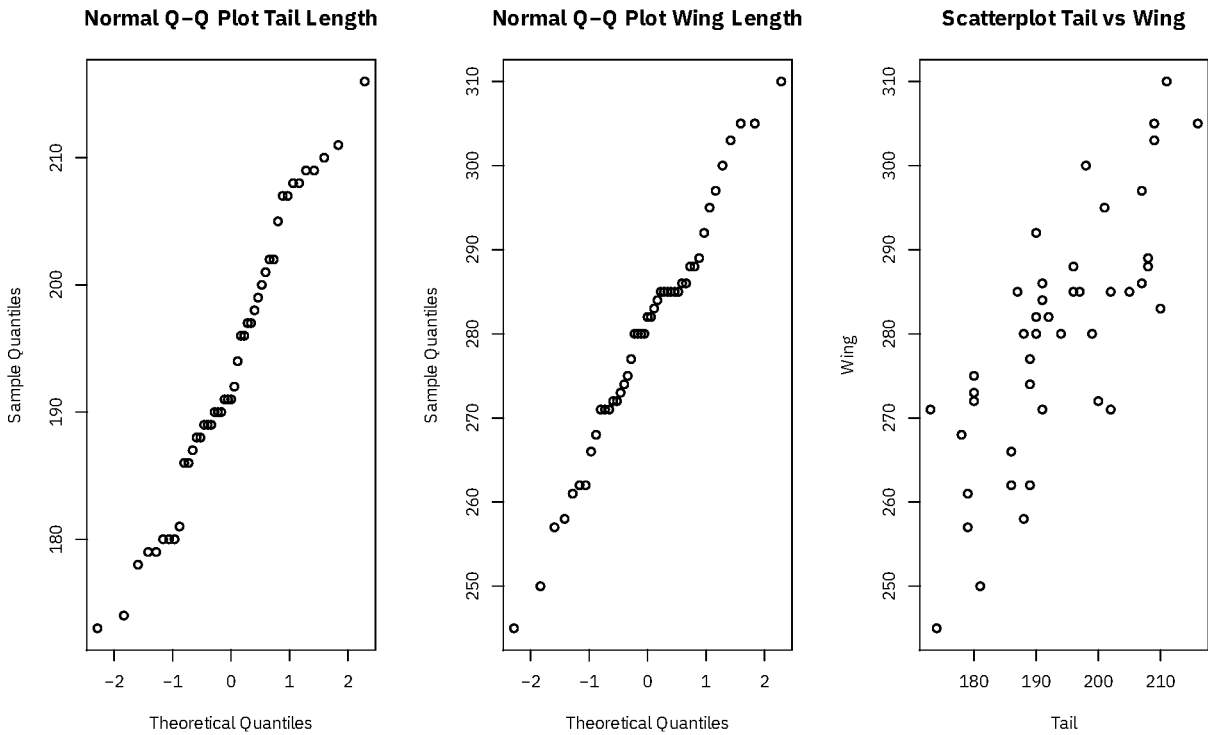
| | Lower | Upper |
|------|----------|----------|
| Tail | 189.8216 | 197.4229 |
| Wing | 274.7819 | 284.7736 |

Correct.

T^2 intervals are larger than Bonferroni intervals. T^2 -intervals takes the correlation between measured variables into account. #If only interested in the component means, the Bonferonni intervals provide more precise estimates. The difference does not depend on mean vector or covariance matrix. It depends on the critical value which obtains the length of the intervals.

The difference does depend on the mean and covariance. But the ratio between their lengths does not.

C)

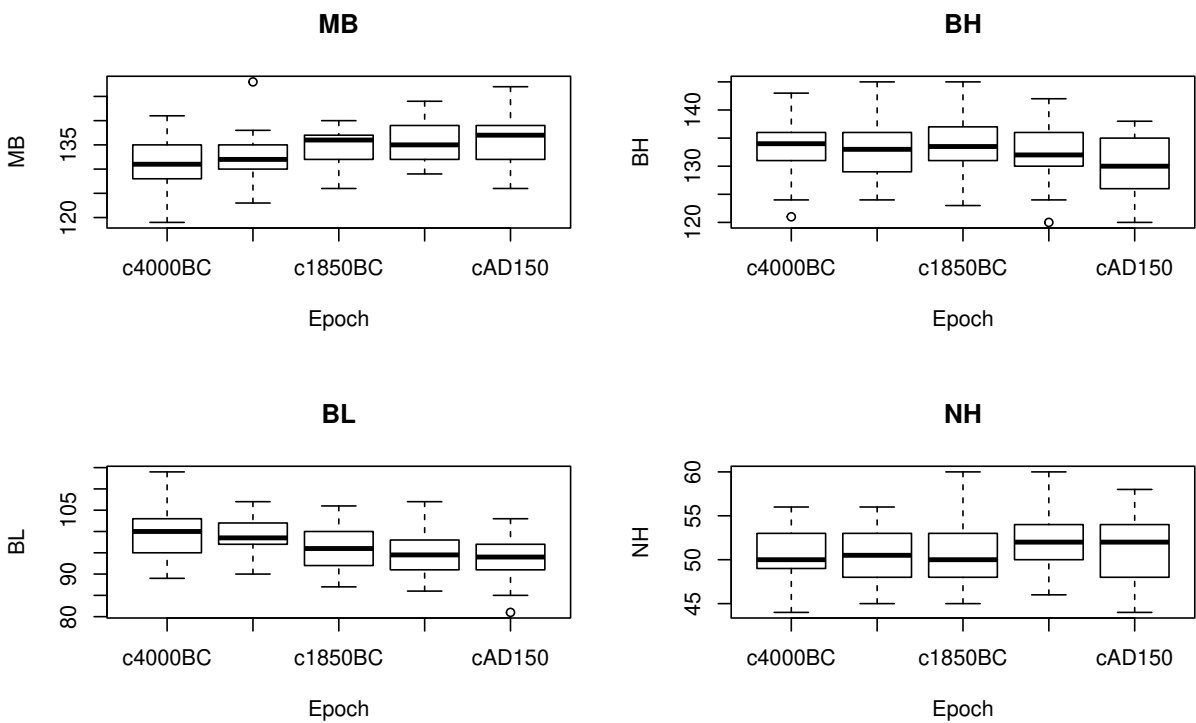


From the plots above, normality assumption seems not to be viable in this case. One possible solution could be transform the variables by a appropriate way.

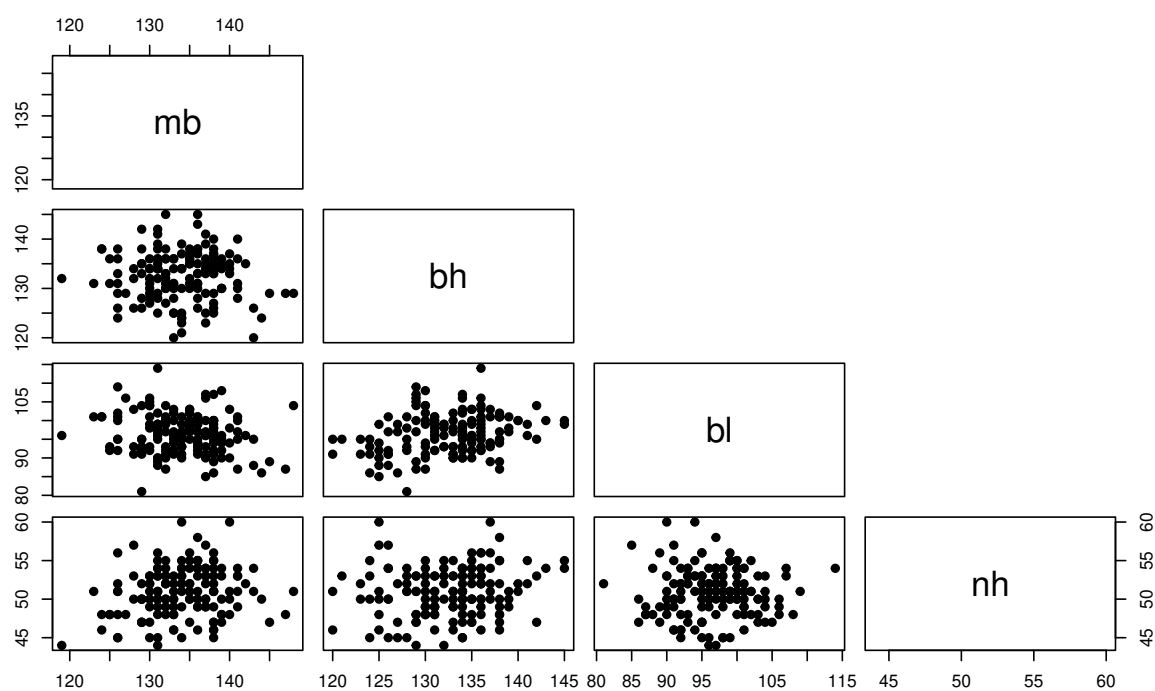
In fact, if you see this plot in a real world data set, this is usually considered "normal enough" in a sense that few people would blame you for assuming normality.

QUESTION 3

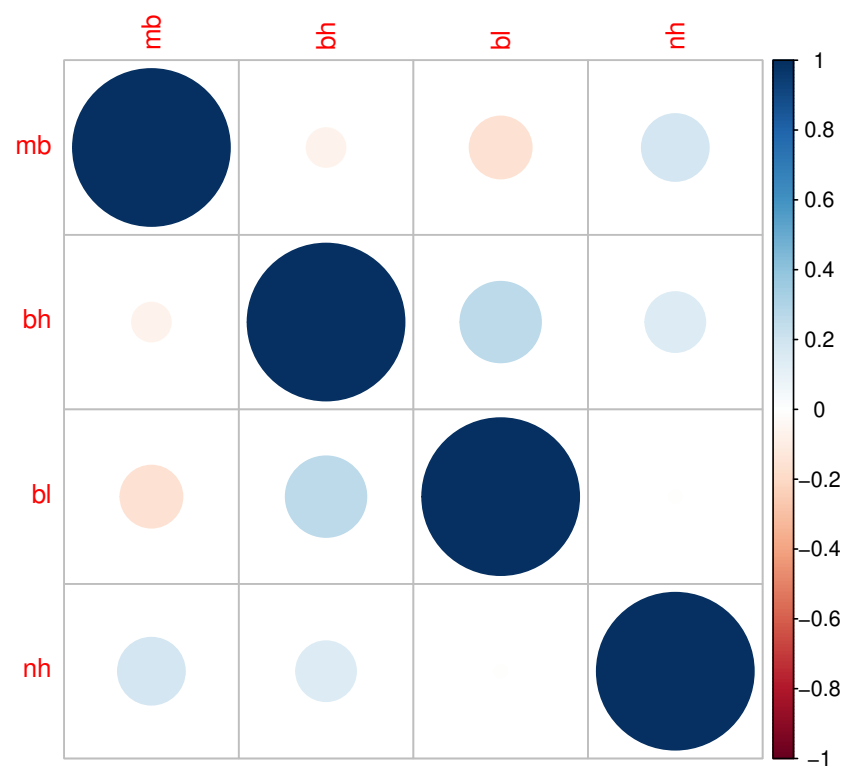
A)



In the figures above, changes for each variable over the different epochs are presented. One can see that there are differences within each variable across the different epochs.



The matrixplot scatterplot above presents the relation between the variables without considering the epochs. No patterns can be detected clearly which should be an indication of uncorrelated variables.



Also the correlation plot above shows that there are low correlations between each pair of variables.

B)

```
##           One Way Multivariate Analysis of Variance
##
## Method : Wilks
## The Value of Test Statistic = 0.6635858
## F value = 3.901 , df1 = 16 , df2 = 434.4548 , p-value: 7.01e-07
##
##           Descriptive Statistics
## $c4000BC
##           mb           bh           bl           nh
## Means 131.366667 133.600000 99.166667 50.533333
## Sd      5.129249   4.469051  5.884423  2.763473
##
## $c3300BC
##           mb           bh           bl           nh
## Means 132.366667 132.700000 99.066667 50.233333
## Sd      4.810071   4.647209  4.346488  2.955805
##
## $c1850BC
##           mb           bh           bl           nh
## Means 134.466667 133.800000 96.033333 50.566667
## Sd      3.481313   4.978575  4.552251  3.549486
##
## $c200BC
##           mb           bh           bl           nh
## Means 135.500000 132.300000 94.533333 51.966667
## Sd      3.919448   5.133729  4.591847  2.822121
##
## $cAD150
##           mb           bh           bl           nh
## Means 136.166667 130.333333 93.500000 51.366667
## Sd      5.350368   4.971181  5.056576  3.718392
##
##
##           Detection important variable(s)
## [1] " Confidence Intervals are calculated only in the Roy Method."
```

Correct!

MANOVA with Wilk’s method was applied. A p-value that is almost zero is obtained. Thus, there are differences between the mean vectors. There have been changes over time for the variables.

C)

```
## $`Epoch1-Epoch2`
##
##
##           Lower           Upper
## ---  -
## mb    -4.905276    2.905276
## bh    -3.218991    5.018991
## bl    -4.079451    4.279451
## nh    -2.408251    3.008251
```

Congratulation! You are among just a few groups that does this correctly! :)

```

##
## $`Epoch1-Epoch3`
##
##
##           Lower           Upper
## ---  -
## mb    -7.005276    0.8052757
## bh    -4.318991    3.9189912
## bl    -1.046117    7.3127839
## nh    -2.741584    2.6749174
##
## $`Epoch1-Epoch4`
##
##
##           Lower           Upper
## ---  -
## mb    -8.0386090   -0.2280576
## bh    -2.8189912    5.4189912
## bl     0.4538827    8.8127839
## nh    -4.1415841    1.2749174
##
## $`Epoch1-Epoch5`
##
##
##           Lower           Upper
## ---  -
## mb    -8.7052757   -0.8947243
## bh    -0.8523246    7.3856579
## bl     1.4872161    9.8461173
## nh    -3.5415841    1.8749174
##
## $`Epoch2-Epoch3`
##
##
##           Lower           Upper
## ---  -
## mb    -6.005276    1.805276
## bh    -5.218991    3.018991
## bl    -1.146117    7.212784
## nh    -3.041584    2.374917
##
## $`Epoch2-Epoch4`
##
##
##           Lower           Upper
## ---  -
## mb    -7.0386090    0.7719424
## bh    -3.7189912    4.5189912
## bl     0.3538827    8.7127839
## nh    -4.4415841    0.9749174
##
## $`Epoch2-Epoch5`
##
##

```



```

##           Lower           Upper
## ---- -
## mb      -7.705276    0.1052757
## bh      -1.752325    6.4856579
## bl       1.387216    9.7461173
## nh      -3.841584    1.5749174
##
## $`Epoch3-Epoch4`
##
##
##           Lower           Upper
## ---- -
## mb      -4.938609    2.871942
## bh      -2.618991    5.618991
## bl      -2.679451    5.679451
## nh      -4.108251    1.308251
##
## $`Epoch3-Epoch5`
##
##
##           Lower           Upper
## ---- -
## mb      -5.6052757    2.205276
## bh      -0.6523246    7.585658
## bl      -1.6461173    6.712784
## nh      -3.5082508    1.908251
##
## $`Epoch4-Epoch5`
##
##
##           Lower           Upper
## ---- -
## mb      -4.571942    3.238609
## bh      -2.152325    6.085658
## bl      -3.146117    5.212784
## nh      -2.108251    3.308251

```

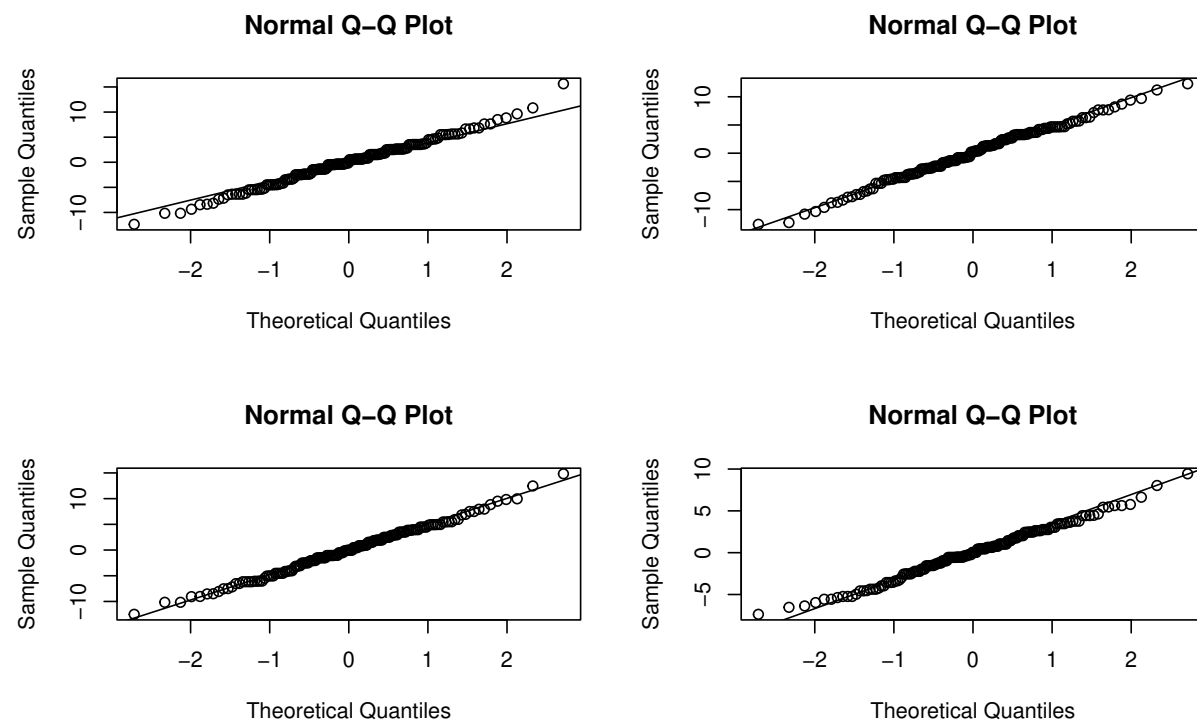
The tables above shows the simultaneous confidence intervals. If the intervals go through 0, this means that there are NO significant differences between the specified epochs for the specific variable.

Now we check if mean of residuals are zero.

| | Residual mean |
|----|---------------|
| mb | -4.11e-17 |
| bh | 2.18e-16 |
| bl | -8.89e-17 |
| nh | -2.57e-17 |

The table above shows values extremely close to zero which means that they can be considered as zeros.

Checking if the residuals are normally distributed by studying the following plots:



The plot shows theoretical quantiles versus residual values' quantiles of the variable. More observations deviated from the line means less likelihood of being normally distributed. The second and the third variables' residuals are mostly aligned on qq lines - residuals of variable 2 and 3 are normally distributed. Observations of residuals on the first and the fourth variables generally follow qq lines. However, compare to the second and the third variable, residual observations are deviated from qq lines, especially as they move away from the mean 0.

APPENDIX

```
##Setup

knitr::opts_chunk$set(echo = TRUE)
library(knitr)
library(heplots)
library(corrplot)
library(MVTests)
knitr::opts_chunk$set(fig.width=8, fig.height=5)

#####-----Q1.A-----#####

my_data <- read.delim(file="C:\\Users\\Suat\\Desktop\\Master_courses\\732A97_Multivariate\\Materials\\T:
colnames(my_data)<-c("Country", "100m", "200m", "400m", "800m", "1500m", "3000m", "Marathon")

data<-my_data

charac <- data.frame(variable = as.numeric(), mean=as.numeric(), std_dev=as.numeric(), stringsAsFactors
for (i in 2:length(data)) {
  mean_v <- mean(data[[i]])
  std_dev_v <- sd(data[[i]])
  charac <- rbind(charac, data.frame(variable = colnames(data)[i], mean=mean_v, std_dev=std_dev_v))
}
mean_vect <- as.vector(charac[,2])
data_mat <- as.matrix(data[, -1])
mean_mat <- matrix(0, ncol=7, nrow=nrow(data_mat))
for (i in 1:nrow(data_mat)) {
  mean_mat[i,] <- mean_vect
}
mean_correct_mat <- data_mat - mean_mat
cov_mat<-cov(data[, -1])
cor_mat<-cor(data[, -1])
d2values<-vector()
for(i in 1:nrow(data)){
  d2values[i]<-t(mean_correct_mat[i,])%*%solve(cov_mat)%*%mean_correct_mat[i,]
}
frame_mah<-data.frame("Country"=my_data$Country,d2values)
frame_mah$pvalues<-1-pchisq(frame_mah$d2values,df=1)
kable(frame_mah[frame_mah$pvalues<0.01,],caption="Outliers") #Outliers

frame_mah$adj_pvalues<-p.adjust(frame_mah$pvalues,method = "bonferroni")
kable(frame_mah)

#####-----Q2.A-----#####

bird_data<-read.table("C:\\Users\\Suat\\Desktop\\Master_courses\\732A97_Multivariate\\Materials\\T5-12.1
colnames(bird_data)<-c("tail", "wing")

the_means<-apply(bird_data,2,mean)
one_matrix<-matrix(1,nrow=dim(bird_data)[1],ncol = 1)

#Cov matrix
```

```

M_mean<-matrix(data=1,nrow=nrow(bird_data))%%cbind(mean(bird_data$tail),mean(bird_data$wing))
diff_matrix<-as.matrix(bird_data-M_mean)
covs<-((nrow(bird_data)-1)^-1)*t(diff_matrix)%%diff_matrix

#Confidence interval
mu1<-190
mu2<-275
n<-nrow(bird_data)
tmp_vector<-c(the_means[1]-mu1,the_means[2]-mu2)
theres<-matrix(c(n*t(tmp_vector)%%solve(covs)%%tmp_vector,2*(n-1)/(n-2)*qf(0.95,ncol(bird_data),n-2))
colnames(theres)<-"Value";rownames(theres)<-c("Test value", "Crit value")
kable(theres,caption="Test value and the critical value")


tmp<-ellipse::ellipse(covs/n,centre = the_means,level = 0.95)
plot(tmp,type = "l")
points(x=190,y=275,col="red")

#####-----Q2.B-----#####

mu1_vector<-c(1,0)
n_col<-ncol(bird_data)
n<-nrow(bird_data)
crit_value<-sqrt((n_col*(n-1))*qf(0.95,n_col,n-n_col)/(n-n_col))
t2_res<-matrix(NA,ncol = 2,nrow = 2);colnames(t2_res)<-c("Lower","Upper");rownames(t2_res)<-c("Tail","W:
#mu1
t2_res[1,1]<-t(mu1_vector)%%the_means-crit_value*sqrt((t(mu1_vector)%%covs%%mu1_vector)/n)
t2_res[1,2]<-t(mu1_vector)%%the_means+crit_value*sqrt((t(mu1_vector)%%covs%%mu1_vector)/n)
#mu2
mu2_vector<-c(0,1)
t2_res[2,1]<-t(mu2_vector)%%the_means-crit_value*sqrt((t(mu2_vector)%%covs%%mu2_vector)/n)
t2_res[2,2]<-t(mu2_vector)%%the_means+crit_value*sqrt((t(mu2_vector)%%covs%%mu2_vector)/n)
kable(t2_res,caption="T2 confidence intervals")

bon_res<-matrix(NA,ncol = 2,nrow = 2);colnames(bon_res)<-c("Lower","Upper");rownames(bon_res)<-c("Tail",
crit_value_bon<-qt(1-(0.05/(2*n_col)),n-1)

bon_res[1,1]<-t(mu1_vector)%%the_means-crit_value_bon*sqrt((t(mu1_vector)%%covs%%mu1_vector)/n)
bon_res[1,2]<-t(mu1_vector)%%the_means+crit_value_bon*sqrt((t(mu1_vector)%%covs%%mu1_vector)/n)
#mu2
mu2_vector<-c(0,1)
bon_res[2,1]<-t(mu2_vector)%%the_means-crit_value_bon*sqrt((t(mu2_vector)%%covs%%mu2_vector)/n)
bon_res[2,2]<-t(mu2_vector)%%the_means+crit_value_bon*sqrt((t(mu2_vector)%%covs%%mu2_vector)/n)
kable(bon_res,caption="Bonferroni intervals")

#####-----Q2.C-----#####

par(mfrow=c(1,3),pch=1)
qqnorm(bird_data$tail,main = "Normal Q-Q Plot Tail Length")

```

```

qqnorm(bird_data$wing,main = "Normal Q-Q Plot Wing Length")
plot(bird_data$tail,bird_data$wing,xlab = "Tail",ylab="Wing",main="Scatterplot Tail vs Wing")

#####-----Q3.A-----#####

skulls<-Skulls

par(mfrow=c(2,2))
boxplot(skulls$mb~skulls$epoch,xlab = "Epoch",ylab="MB",main="MB")
boxplot(skulls$bh~skulls$epoch,xlab = "Epoch",ylab="BH",main="BH")
boxplot(skulls$bl~skulls$epoch,xlab = "Epoch",ylab="BL",main="BL")
boxplot(skulls$nh~skulls$epoch,xlab = "Epoch",ylab="NH",main="NH")

pairs(skulls[,-1],pch=19,upper.panel = NULL)

corrplot(cor(skulls[,-1]))

#####-----Q3.B-----#####

my_manova<-Manova(data=skulls[,-1],group = skulls[,1])
summary(my_manova)

#####-----Q3.C-----#####

ep<-5
p<-ncol(skulls[,-1])
split1<-skulls[skulls$epoch=="c4000BC",]
split2<-skulls[skulls$epoch=="c3300BC",]
split3<-skulls[skulls$epoch=="c1850BC",]
split4<-skulls[skulls$epoch=="c200BC",]
split5<-skulls[skulls$epoch=="cAD150",]
n1<-30;n2<-30;n3<-30;n4<-30;n5<-30
n<-150

means1<-apply(split1[,-1],2,mean)
means2<-apply(split2[,-1],2,mean)
means3<-apply(split3[,-1],2,mean)
means4<-apply(split4[,-1],2,mean)
means5<-apply(split5[,-1],2,mean)

S1<-cov(split1[,-1])
S2<-cov(split2[,-1])
S3<-cov(split3[,-1])
S4<-cov(split4[,-1])
S5<-cov(split5[,-1])

critical<-qt(1-0.05/(p*ep*(ep-1)),df=n-ep)

W <- (n1-1)*S1+(n2-1)*S2+(n3-1)*S3+(n4-1)*S4+(n5-1)*S5

CI12<-matrix(NA,ncol=2,nrow=4);colnames(CI12)<-c("Lower","Upper");rownames(CI12)<-c("mb","bh","bl","nh")
for(i in 1:p){

```

```

    CI12[i,1]<-(means1[i]-means2[i])-critical*sqrt(W[i,i]/(n-ep)*(1/n1+1/n2))
    CI12[i,2]<-(means1[i]-means2[i])+critical*sqrt(W[i,i]/(n-ep)*(1/n1+1/n2))
  }
  CI13<-matrix(NA,ncol=2,nrow=4);colnames(CI13)<-c("Lower","Upper");rownames(CI13)<-c("mb","bh","bl","nh")
  for(i in 1:p){
    CI13[i,1]<-(means1[i]-means3[i])-critical*sqrt(W[i,i]/(n-ep)*(1/n1+1/n3))
    CI13[i,2]<-(means1[i]-means3[i])+critical*sqrt(W[i,i]/(n-ep)*(1/n1+1/n3))
  }

  CI14<-matrix(NA,ncol=2,nrow=4);colnames(CI14)<-c("Lower","Upper");rownames(CI14)<-c("mb","bh","bl","nh")
  for(i in 1:p){
    CI14[i,1]<-(means1[i]-means4[i])-critical*sqrt(W[i,i]/(n-ep)*(1/n1+1/n4))
    CI14[i,2]<-(means1[i]-means4[i])+critical*sqrt(W[i,i]/(n-ep)*(1/n1+1/n4))
  }

  CI15<-matrix(NA,ncol=2,nrow=4);colnames(CI15)<-c("Lower","Upper");rownames(CI15)<-c("mb","bh","bl","nh")
  for(i in 1:p){
    CI15[i,1]<-(means1[i]-means5[i])-critical*sqrt(W[i,i]/(n-ep)*(1/n1+1/n5))
    CI15[i,2]<-(means1[i]-means5[i])+critical*sqrt(W[i,i]/(n-ep)*(1/n1+1/n5))
  }

  CI23<-matrix(NA,ncol=2,nrow=4);colnames(CI23)<-c("Lower","Upper");rownames(CI23)<-c("mb","bh","bl","nh")
  for(i in 1:p){
    CI23[i,1]<-(means2[i]-means3[i])-critical*sqrt(W[i,i]/(n-ep)*(1/n2+1/n3))
    CI23[i,2]<-(means2[i]-means3[i])+critical*sqrt(W[i,i]/(n-ep)*(1/n2+1/n3))
  }

  CI24<-matrix(NA,ncol=2,nrow=4);colnames(CI24)<-c("Lower","Upper");rownames(CI24)<-c("mb","bh","bl","nh")
  for(i in 1:p){
    CI24[i,1]<-(means2[i]-means4[i])-critical*sqrt(W[i,i]/(n-ep)*(1/n2+1/n4))
    CI24[i,2]<-(means2[i]-means4[i])+critical*sqrt(W[i,i]/(n-ep)*(1/n2+1/n4))
  }

  CI25<-matrix(NA,ncol=2,nrow=4);colnames(CI25)<-c("Lower","Upper");rownames(CI25)<-c("mb","bh","bl","nh")
  for(i in 1:p){
    CI25[i,1]<-(means2[i]-means5[i])-critical*sqrt(W[i,i]/(n-ep)*(1/n2+1/n5))
    CI25[i,2]<-(means2[i]-means5[i])+critical*sqrt(W[i,i]/(n-ep)*(1/n2+1/n5))
  }

  CI34<-matrix(NA,ncol=2,nrow=4);colnames(CI34)<-c("Lower","Upper");rownames(CI34)<-c("mb","bh","bl","nh")
  for(i in 1:p){
    CI34[i,1]<-(means3[i]-means4[i])-critical*sqrt(W[i,i]/(n-ep)*(1/n3+1/n4))
    CI34[i,2]<-(means3[i]-means4[i])+critical*sqrt(W[i,i]/(n-ep)*(1/n3+1/n4))
  }

  CI35<-matrix(NA,ncol=2,nrow=4);colnames(CI35)<-c("Lower","Upper");rownames(CI35)<-c("mb","bh","bl","nh")
  for(i in 1:p){
    CI35[i,1]<-(means3[i]-means5[i])-critical*sqrt(W[i,i]/(n-ep)*(1/n3+1/n5))
    CI35[i,2]<-(means3[i]-means5[i])+critical*sqrt(W[i,i]/(n-ep)*(1/n3+1/n5))
  }

  CI45<-matrix(NA,ncol=2,nrow=4);colnames(CI45)<-c("Lower","Upper");rownames(CI45)<-c("mb","bh","bl","nh")
  for(i in 1:p){
    CI45[i,1]<-(means4[i]-means5[i])-critical*sqrt(W[i,i]/(n-ep)*(1/n4+1/n5))
    CI45[i,2]<-(means4[i]-means5[i])+critical*sqrt(W[i,i]/(n-ep)*(1/n4+1/n5))
  }

my_list<-list("Epoch1-Epoch2"=CI12,"Epoch1-Epoch3"=CI13,"Epoch1-Epoch4"=CI14,"Epoch1-Epoch5"=CI15,"Epoch2-Epoch3"=CI23,"Epoch2-Epoch4"=CI24,"Epoch2-Epoch5"=CI25,"Epoch3-Epoch4"=CI34,"Epoch3-Epoch5"=CI35,"Epoch4-Epoch5"=CI45)

```

```

lapply(X = my_list, FUN = function(i) {
  kable(x = i, caption = names(i),booktabs="TRUE")

tmp1 <- manova(cbind(mb, bh, bl, nh)~epoch, data= Skulls)
res <- tmp1$residuals

res_mean <- c()
for (i in 1:4) {
  meanval <- mean(res[,i])
  res_mean <- append(res_mean, meanval)
}
res_mean<-as.matrix(res_mean)
rownames(res_mean)<-c("mb","bh","bl","nh");colnames(res_mean)<- "Residual mean"
res_mean[1:4]<-format(res_mean[1:4],digits=3)
kable(res_mean)

par(mfrow=c(2,2))
for (i in 1:4) {
  qqnorm(res[,i])
  qqline(res[,i])
}

```