Lab2 - Inference about mean vectors, MANOVA

Andreas C Charitos [andch552]

1/22/2021

Contents

Problem 1

Data Overview

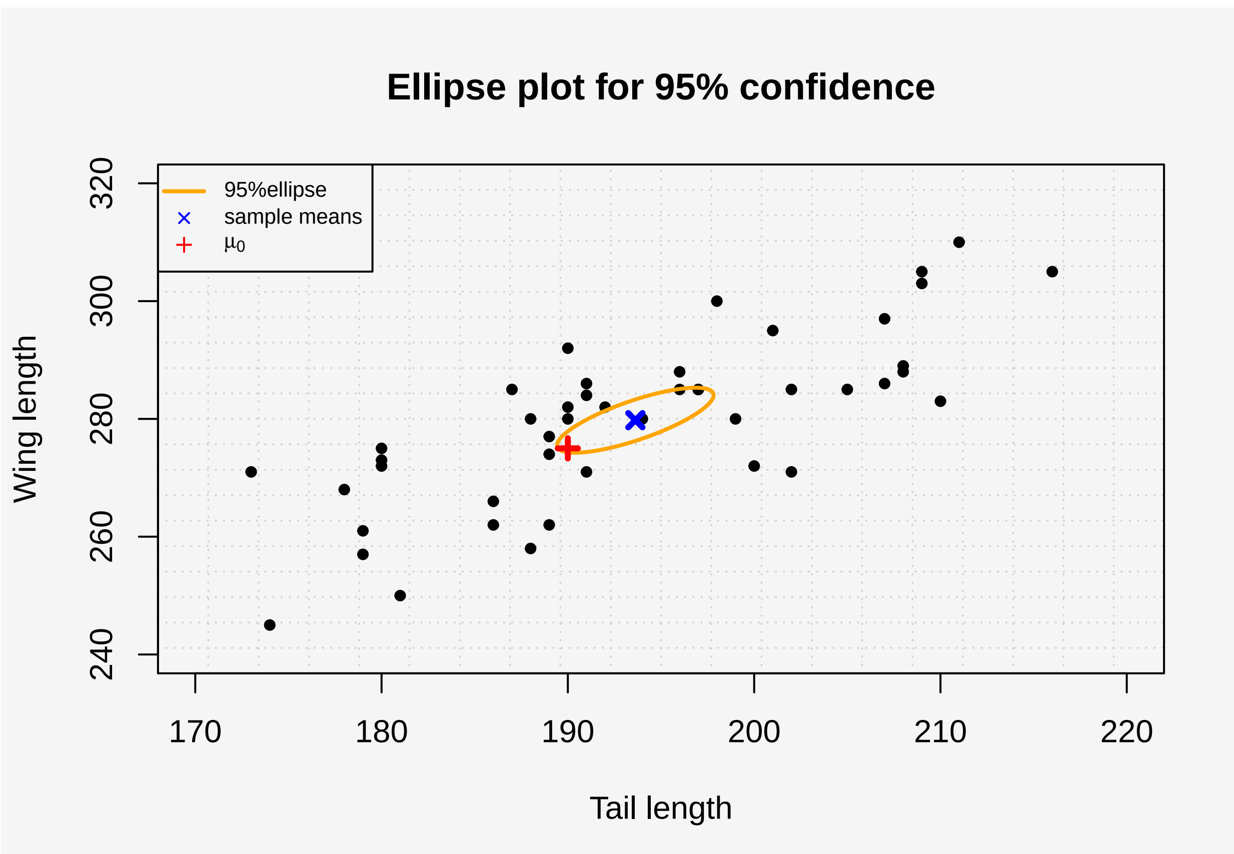
Table of the first 3 lines of the bird data

Tail Length	Wing Length
191	284
197	285
208	288
180	273
180	275
188	280

a)

FInd and sketch the 95% confidence ellipse for the polulation means μ_1 and μ_2. Suppose it is known that $\mu_1 = 190\ mm$ and $\mu_2 = 275\ mm$ for male hook-biled kites. Are these plausible values for the mean tail length and mean wing length for the female birds? Explain.

Confidence Ellipse



In the above plot above we can see a visualization of the 95% ellipse confidence, the sample means (blue cross) and the μ_0 (red cross). As it is evident from the plot the μ_0 lie inside the ellipse thus the vector contains the values of the means.

- Because T^2 is smaller than the citical value of T^2 at $\alpha = 5\%$ we can not reject the null hypotheses and we can conclude that the population means of the male birds are plausible means for the female birds.

b)

Construct the simultaneous 95% $T^2 - intervals$ for μ_1 and μ_2 and the 95% Bonferroni intervals for μ_1 and μ_2. Compare the two sets of intervals. What advantage, if any, do the $T^2 - intervals$ have over the Bonferroni intervals?

T^2 *Simultaneous intervals*

Table 2: Simultaneous Intervals Table

	Tail Length	Wing Length
lower band	189.42	274.26
upper band	197.82	285.30

Bonferroni intervals

Table 3: Bonferroni Intervals Table

	Tail Length	Wing Length
lower band	190.32	275.44
upper band	196.92	284.12

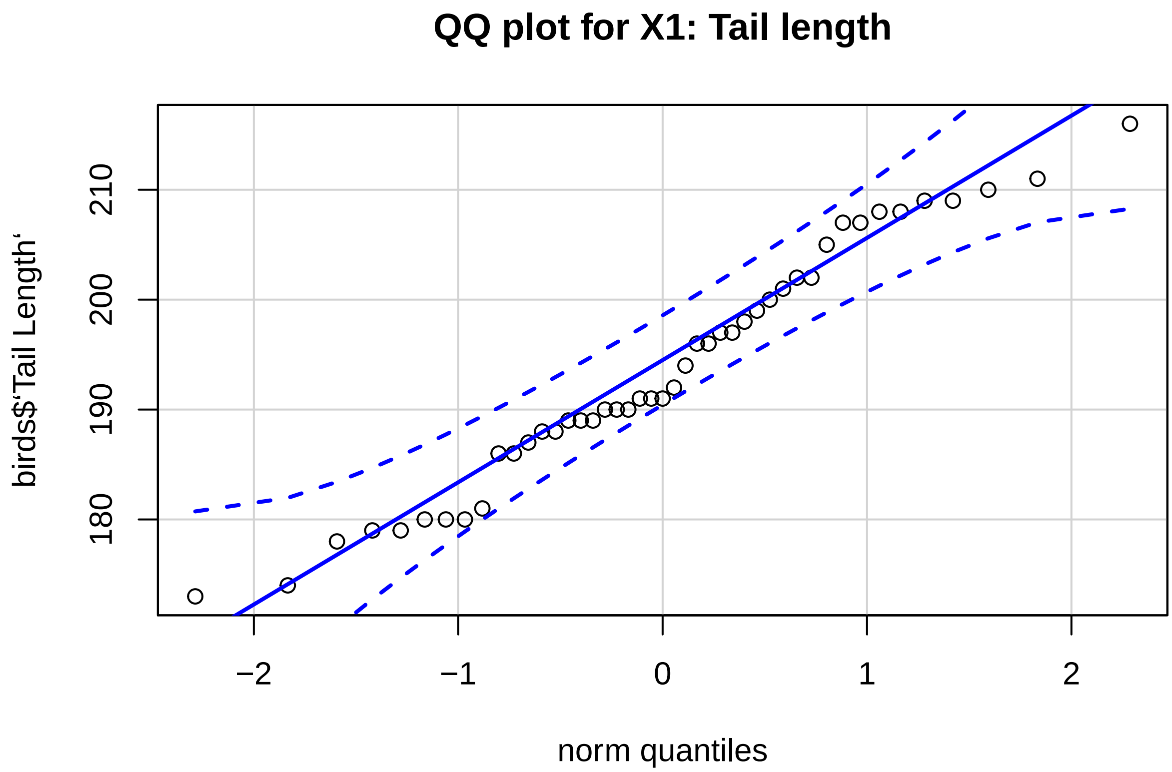
As it is evident, $T^2 - simultaneous$ CI is slightly wider than Bonferroni Intervals. Bonferroni method provides shorter intervals when m = p. Because they are easy to apply and provide the relatively short confidence intervals needed for inference.

According to the book: "The simultaneous confidence intervals(T^2) are ideal for "data snooping." The confidence coefficient $1 - \alpha$ remains unchanged for any choice of $\mathbf{a}$, so linear combinations of the components μ_i that merit inspection based upon an examination of the data can be estimated.

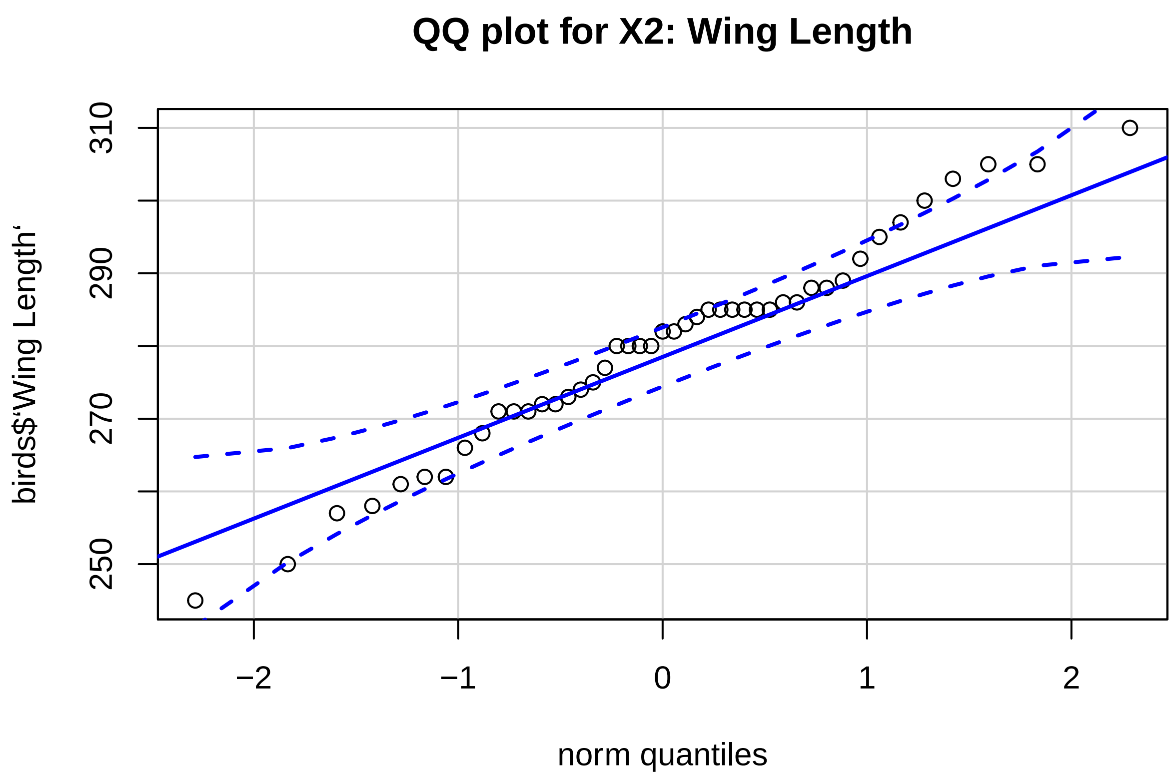
c)

Is the bivariate normal distribution a viable population model? Explain with reference to Q-Q plots and scatter digram

Q-Q plots

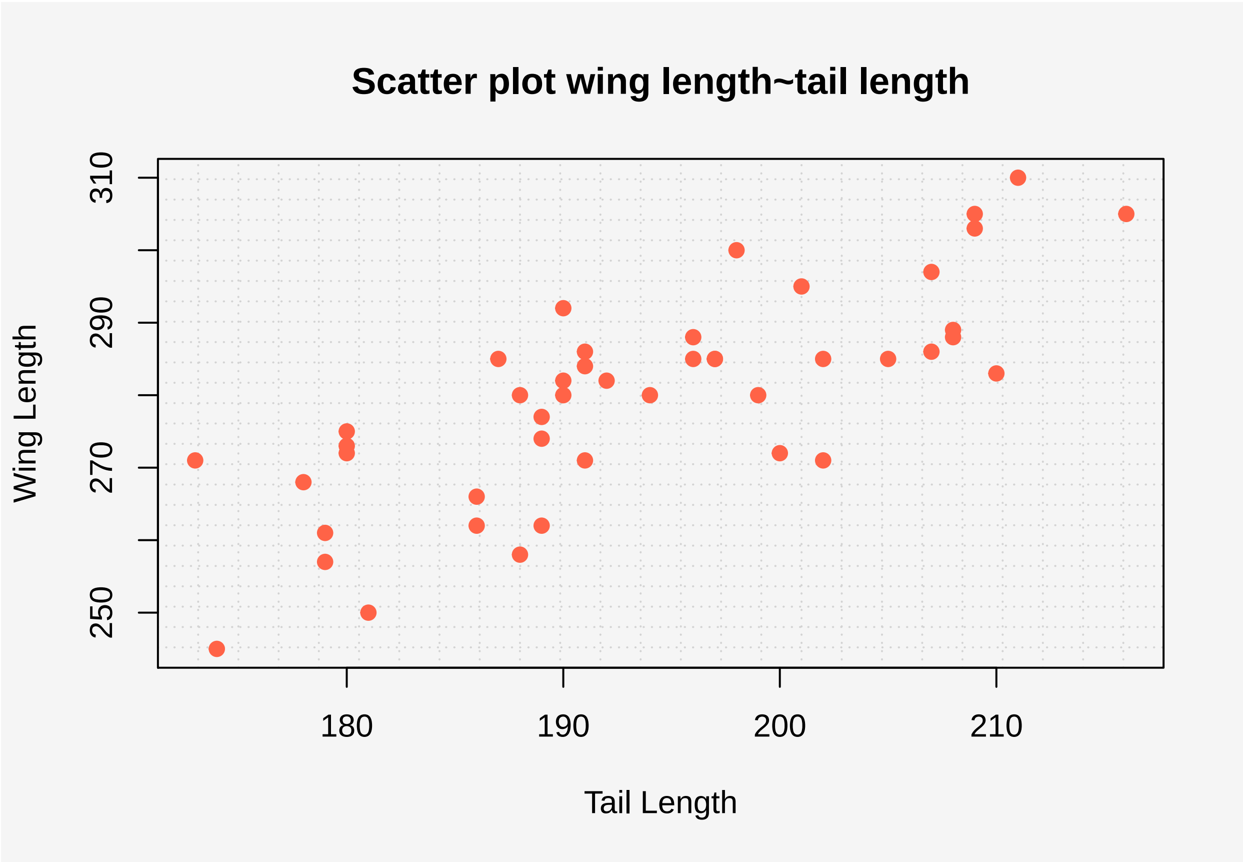


QQ plot for X1: Tail length
birds$`Tail Length`
180
190
200
210
−2
−1
0
1
2
norm quantiles



QQ plot for X2: Wing Length
birds$`Wing Length`
250
270
290
310
−2
−1
0
1
2
norm quantiles

Scatter plot



The Q-Q plot for both variables (x_1 & x_2) illustrate a linear trend. The scatterplot of two variables also indicates a linear relationship between these two features. These linear trends can lead us to this conclusion that the population can be considered as normal.

Problem 2

Data Overview

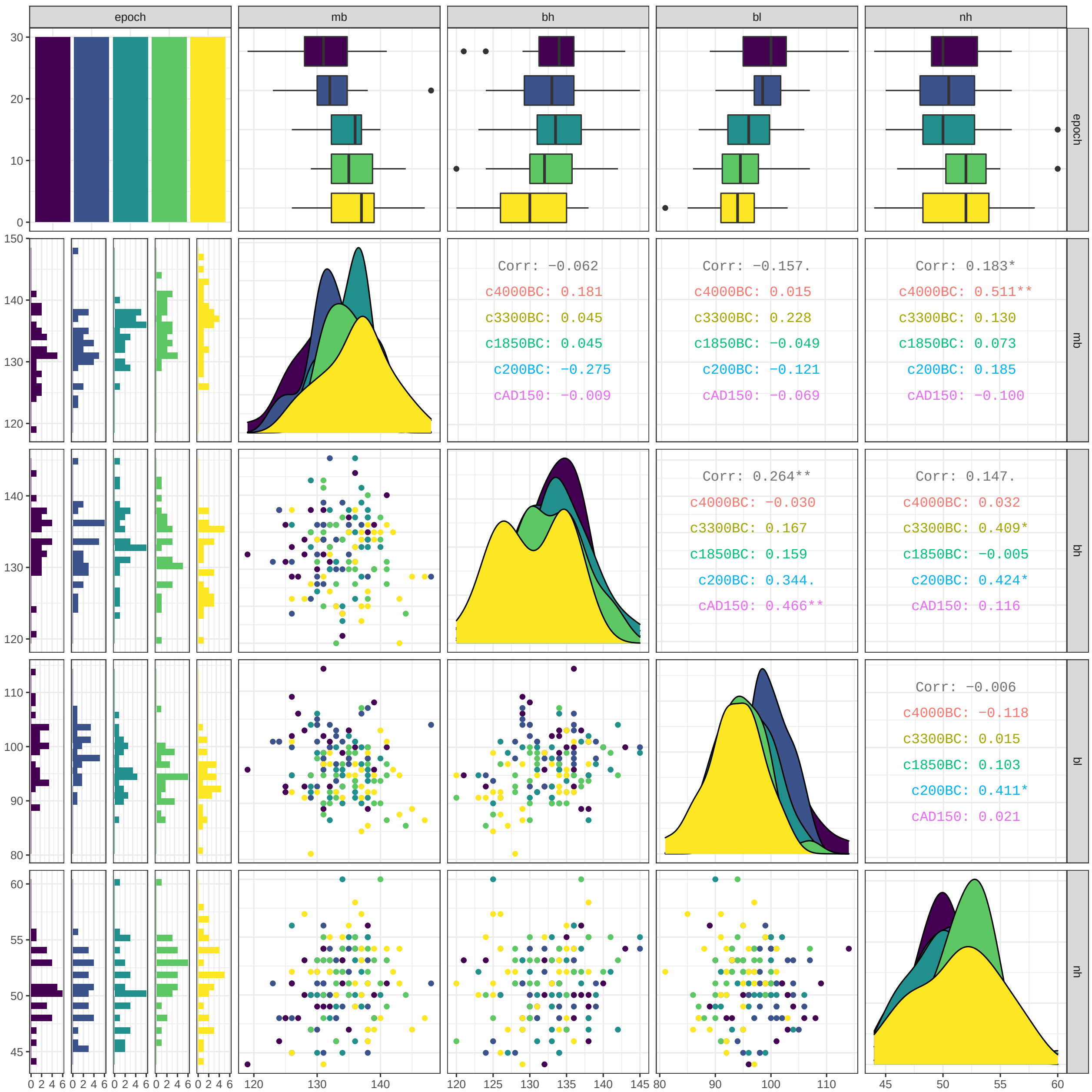
epoch	mb	bh	bl	nh
c4000BC	131	138	89	49
c4000BC	125	131	92	48
c4000BC	131	132	99	50
c4000BC	119	132	96	44

a)

Plot the data using graphs that you find informative. Justify your choice.

Pair plots

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Looking at the different distributions of the variables we can see that all of them are distributed around a mean. The variables mb and bl seem to be symmetrical while the variables bh and nh seem not. Looking at the scatterplots and the corresponding correlations we can observe that apart from bh with mb and nh and bl all of them have a slight correlation.

b)

Test (at 5% significance level) if the mean vectors differ for different epochs.

MANOVA

```
## [1] "================== Result ========================"
```

```
## Call:
##    manova(cbind(mb, bh, bl, nh) ~ epoch, Skulls)
##
## Terms:
##                     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
```

```
## [1] "================== Summary ======================="
```

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## 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 mean vectors do defer except nasal height. All other means are different between the epochs with a significant level of 5%. Moreoever, according to the p-value given from *Summary* we should reject the hypothesis that all in-group means (per-epoch means) are equal.

c)

Construct the 95% simultaneous confidence intervals for the components of the mean vectors.

Confidence intervals

	mb	bh	bl	nh
(epoch1 - epoch2)	(-4.905 2.905)	(-7.005 0.805)	(-8.039 -0.228)	(-8.705 -0.895)
(epoch1 - epoch3)	(-3.219 5.019)	(-4.319 3.919)	(-2.819 5.419)	(-0.852 7.386)
(epoch1 - epoch4)	(-4.079 4.279)	(-1.046 7.313)	(0.454 8.813)	(1.487 9.846)
(epoch1 - epoch5)	(-2.408 3.008)	(-2.742 2.675)	(-4.142 1.275)	(-3.542 1.875)

Appendix

```
## ----message=FALSE,echo=FALSE-----------------------------------------------
# Import libraries ---------------------------------------------------------
# import libraries ---------------------------------------------------------
library(car)
library(latticeExtra)
library(heplots)
library(GGally)
library(knitr)
library(CARS)


## ---- echo=FALSE-------------------------------------------------------------
# a) -----------------------------------------------------------------------

birds=read.table("T5-12.DAT")
colnames(birds)=c("Tail Length","Wing Length")
kable(head(birds))




## ----echo=F------------------------------------------------------------------

x1=birds$`Tail Length` ; x2=birds$`Wing Length`
n=dim(birds)[1] ; p=dim(birds)[2]

a=0.05
mu0=c(190,275)
xbar=colMeans(birds)
S=cov(birds)

dist=xbar-mu0
# crit_val=sqrt(p*(n-1)/(n*(n-p)))*qf(1-a,p,n-p)
crit_val=sqrt(p*(n-1)/(n*(n-p))*qf(1-a,p,n-p))


angles=seq(0,2*pi,length.out=200)
# eigen values and eigen vectors of covariance matrix
eigVal <-eigen(S)$values
eigVec <- eigen(S)$vectors
eigScl <- eigVec%*%diag(sqrt(eigVal))
xMat <- rbind(xbar[1] + eigScl[1,]**crit_val, xbar[1]- eigScl[1,]*crit_val)
yMat <- rbind(xbar[2] + eigScl[2,]**crit_val, xbar[2]- eigScl[2,]*crit_val)
ellBase <- cbind(sqrt(eigVal[1])*crit_val*cos(angles), sqrt(eigVal[2])* crit_val*sin(angles))
ellRot <- eigVec%*%t(ellBase)




## ---- echo=F-----------------------------------------------------------------
par(bg='whitesmoke')
plot(birds$`Tail Length`,
     birds$`Wing Length` ,
```

```r
     xlim = c(170,220),ylim=c(240,320),
     xlab='Tail length',
     ylab='Wing length',
     main='Ellipse plot for 95% confidence',
     panel.first = grid(20,20),pch=20)
lines( (ellRot+xbar)[1,],(ellRot+xbar)[2,],
       asp=1,type="l",lwd=2,col="orange")
points(xbar[1],xbar[2],pch=4,col="blue",lwd=3)
points(mu0[1],mu0[2],pch=3,col="red",lwd=3)
legend('topleft',legend=c('95%ellipse','sample means',expression(mu[0])),
          col=c('orange','blue', 'red'), pch=c(NA,4,3),
          lwd=c(2,NA, NA), cex=0.7)



## ---- echo=F-------------------------------------------------------------------
# Simultaneous Intervals

f <- sqrt(((n-1)*p/(n-p))*qf(1-a, p, n-p))
sim_low <- round((t(xbar) - f * sqrt(diag(S)/n)),2)
sim_up  <- round((t(xbar) + f * sqrt(diag(S)/n)),2)
sim_interval=rbind(sim_low, sim_up)
rownames(sim_interval)=c("lower band", "upper band")
kable(sim_interval, caption = "Simultaneous Intervals Table")



## ---- echo=F-------------------------------------------------------------------
#Bonferroni Intervals
t <- qt((1-a/(2)), df = (n-1))
bon_low <- round((t(xbar) - t * sqrt(diag(S)/n)),2)
bon_up  <- round((t(xbar) + t * sqrt(diag(S)/n)),2)
bon_interval=rbind(bon_low, bon_up)
rownames(bon_interval) <- c("lower band", "upper band")

kable(bon_interval, caption = "Bonferroni Intervals Table")



## ---- echo=F-------------------------------------------------------------------
# c) ----------------------------------------------------------------------

# qqnorm(birds[,1], main = "Q-Q plot for x1",
#        col="purple", pch=19, panel.first=grid(25, 25))
# qqline(birds[,1], col="orange", lwd=2)

qqPlot(birds$`Tail Length`, main ="QQ plot for X1: Tail length",id=F)



## ---- echo=F-------------------------------------------------------------------
# qqnorm(birds[,2], main = "Q-Q plot for x2",
#        col="mediumaquamarine", pch =19,
```

```r
#        panel.first=grid(25, 25))
# qqline(birds[,2], col="mediumslateblue", lwd=2)

qqPlot(birds$`Wing Length`, main="QQ plot for X2: Wing Length",id=F )



## ---- echo=F------------------------------------------------------------------
par(bg='whitesmoke')
plot(birds[,1], birds[,2],
     xlab=colnames(birds)[1], ylab = colnames(birds)[2],
     col="tomato",pch=19, panel.first = grid(25,25),
     main="Scatter plot wing length~tail length")



## ---- echo=F------------------------------------------------------------------
# a) ---------------------------------------------------------------------
kable(head(Skulls, n=4))



## ---- echo=F, messenge=F, fig.width=12, fig.height=12------------------------

ggpairs(Skulls, mapping = aes(color = epoch)) + theme_bw()


## ---- echo=F------------------------------------------------------------------
res = manova(cbind(mb,bh,bl,nh)~epoch, Skulls)
print("=================== Result ========================")
res
cat("\n")
cat("\n")
print("=================== Summary =======================")
summary(res)


## ---- echo=F------------------------------------------------------------------

# c) ---------------------------------------------------------------------
w_mb =sum(res$residuals[,1]^2)
w_bh=sum(res$residuals[,2]^2)
w_bl=sum(res$residuals[,3]^2)
w_nh=sum(res$residuals[,4]^2)
w=c(w_mb, w_bh, w_bl, w_nh)
epoch =as.character(unique(Skulls$epoch))
g =length(unique(Skulls$epoch))
p =ncol(Skulls)
n = 150
a = 0.05
C = -qt(a/((p-1)*g*(g-1)), (n-g))* sqrt(2*w/(30*(n-g)))
C_mat =matrix(c(1,1,1,1),4)%*%C
```

```
#Calculating the mean values of the samples and the differences between them.
xbar=matrix(0, nrow = 5, ncol = 4, dimnames =list(epoch,names(Skulls[,-1])))
dist=matrix(0, 4,4)

for(i in 2:p){
        for(j in 1:g){
                xbar[j,(i-1)] =mean(Skulls[which(Skulls$epoch==epoch[j]),i])
        }
        for(k in 1:4) {dist[k, i-1] <- xbar[1, i-1]-xbar[k+1, i-1]
        }
}

SI_lower = dist-C_mat
SI_upper = dist+C_mat
e1 =round(t(SI_lower),3)
e2 =round(t(SI_upper),3)
interval =matrix(0, 4,4)
for(i in 1:4){
        for(j in 1:4) {
                interval[i,j] =paste("(",e1[i,j],e2[i,j],")" ,sep = " ")
                }
        }
colnames(interval) =names(Skulls[,-1])
rownames(interval) =c("(epoch1 - epoch2)", "(epoch1 - epoch3)","(epoch1 - epoch4)", "(epoch1 - epoch5)")
knitr::kable(interval)


## ----code=readLines(knitr::purl("/home/quartermaine/Courses/Multivariate-Statistical-Methods/labs/Ass
## NA
```