Computer Lab 2

Thomas Zhang

2015-11-24

Assignment 1

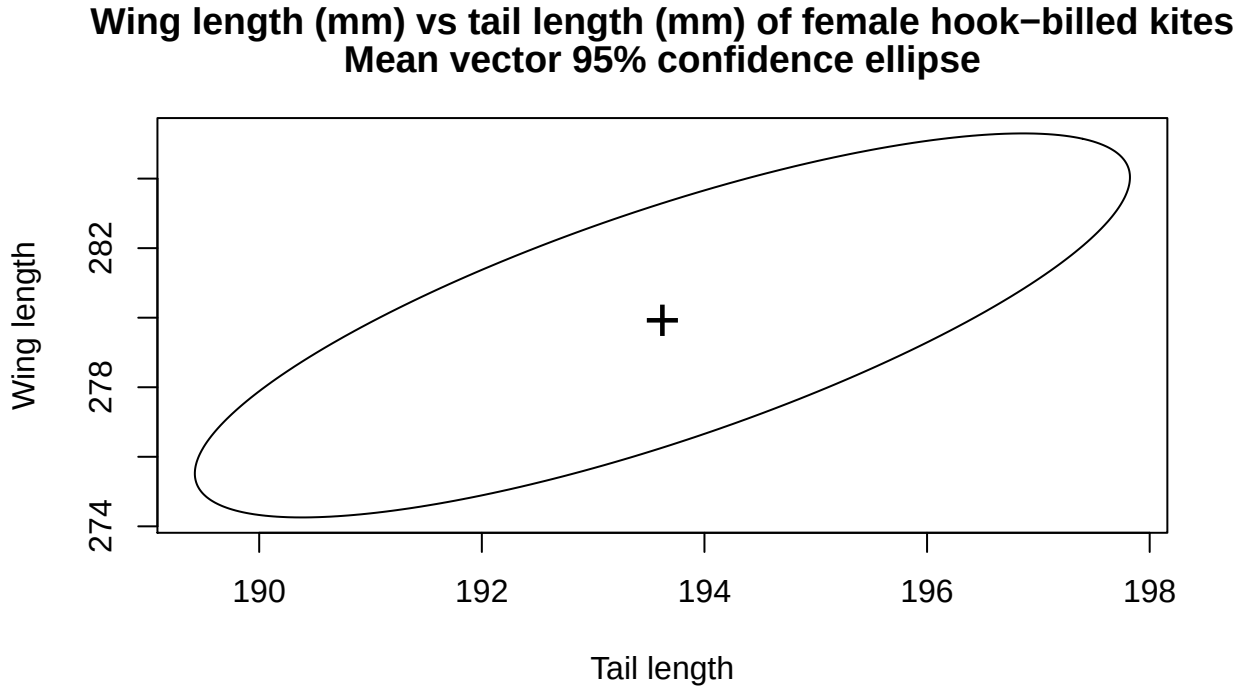
We calculate the chi-square test statistic for seven degrees of freedom at 0.1% significance level and compare them to the squared Mahalanobis distances of countries from the sample mean vector in order to determine which countries can be considered outliers.

```
## [1] "KOR, N  is an outlier according to chisq test at 0.1% significance level"
## [1] "PNG  is an outlier according to chisq test at 0.1% significance level"
## [1] "SAM  is an outlier according to chisq test at 0.1% significance level"
```

North Korea probably does not perform very extremely (i.e. different from the sample mean) in any one distance record, but their records must not exhibit the same covariance with each other as that of other nations. One interpretation is that their training regimen produced vastly different results from that of most countries. Another interpretation could be that the records are fabricated by somebody without knowledge of running record covariances.

Assignment 2

Based on the data given, we plot the 95% confidence ellipse for the tail length and wing length population means for female hook-billed kites.



According to wikipedia, "Most accipitrids exhibit sexual dimorphism in size, unusually for birds, it is the females that are larger than the males".

We can easily see that our mean vector 95% confidence ellipse barely touches the population means for male hook-billed kites in both variables, so our data may reflect sexual dimorphism in hook-billed kites.

Let us calculate simultaneous confidence intervals and Bonferroni confidence intervals for the population means and compare them.

```
## [1] "Simultaneous confidence interval for Tail length: ( 189.42 , 197.82 )"
```

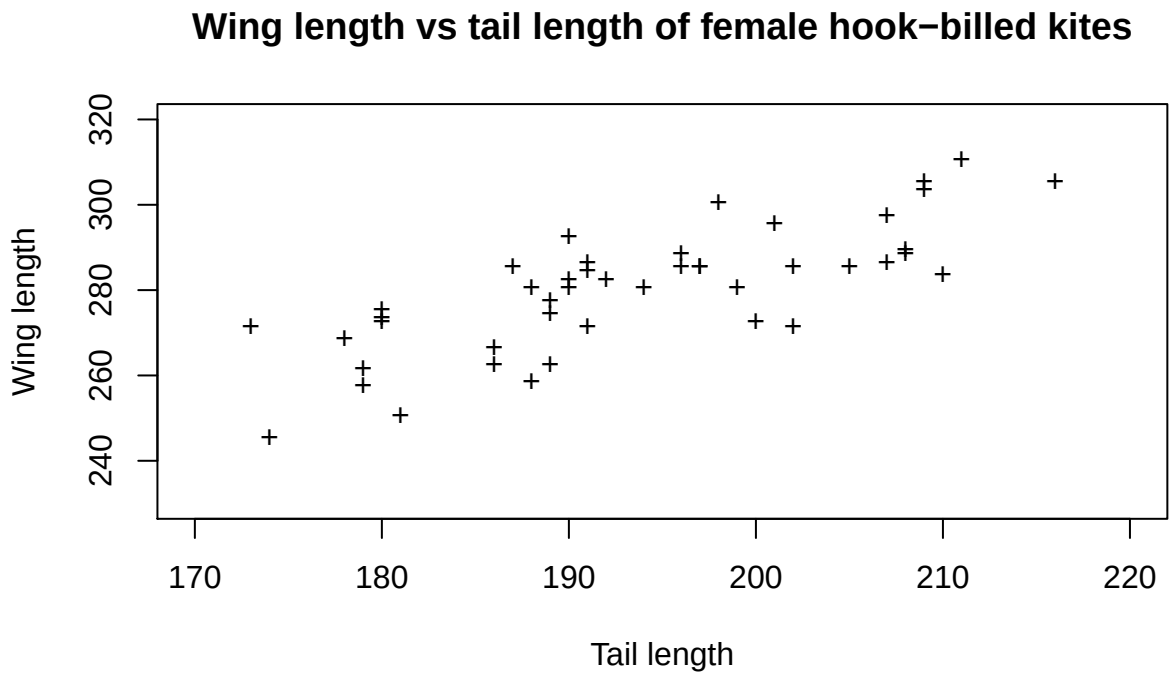
```
## [1] "Simultaneous confidence interval for Wing length: ( 274.26 , 285.3 )"
```

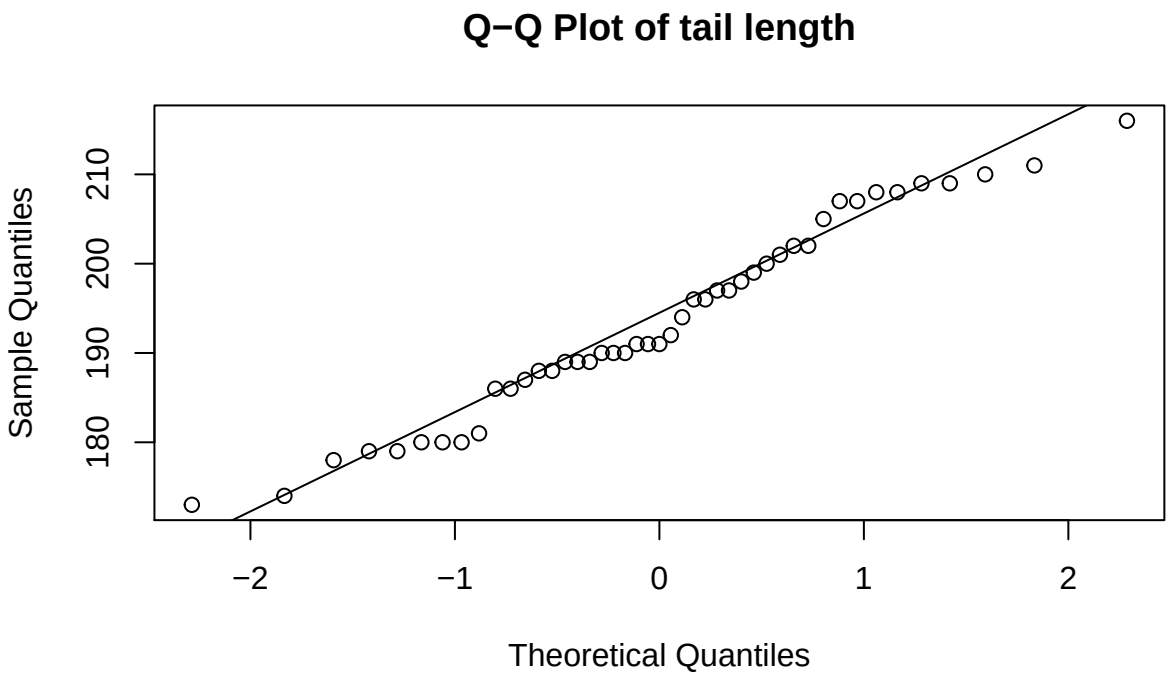
```
## [1] "Bonferroni confidence interval for Tail length: ( 189.82 , 197.42 )"
```

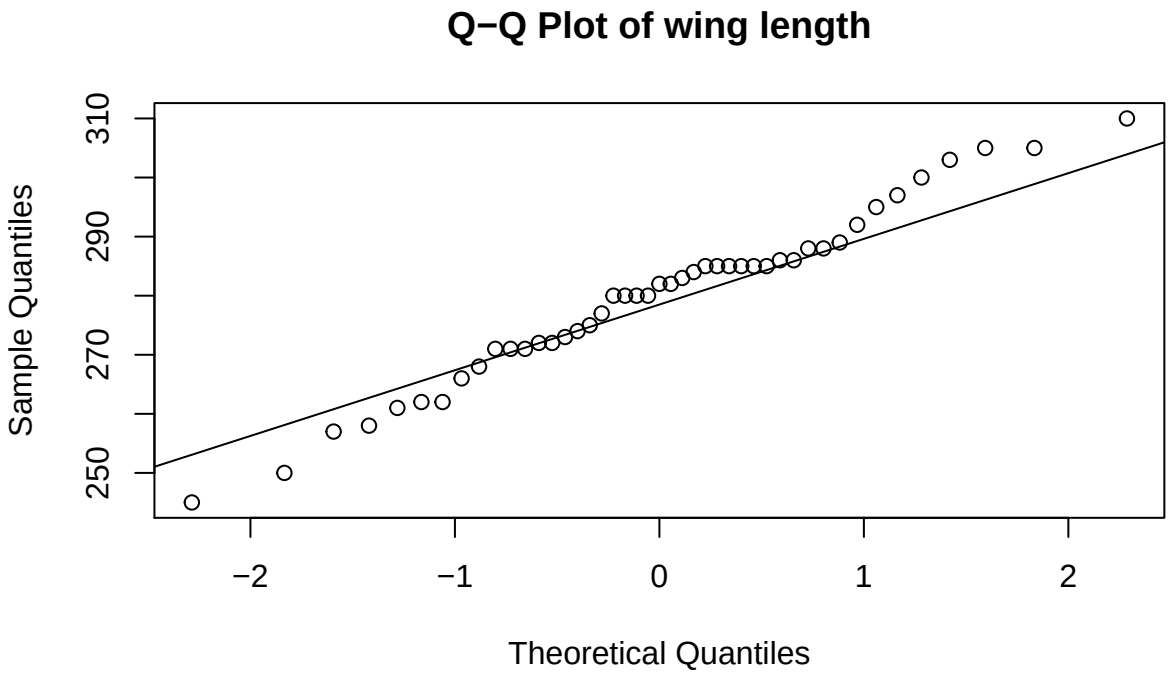
```
## [1] "Bonferroni confidence interval for Wing length: ( 274.78 , 284.77 )"
```

It looks as if the Bonferroni intervals are a little smaller than the simultaneous confidence intervals, and therefore more precise. The downside about using Bonferroni intervals is maybe that we lose the ability to easily make mean confidence intervals around the differences of mean vector components, like we could using simultaneous confidence intervals.

Now we consider whether the data can be described by a bivariate normal distribution by looking at the scatter-plot and Q-Q plots of the data variables. In the Q-Q plots the straight line indicate the theoretical quantile distribution of normally distributed data with the line passing through the first and third quartile sample data points.







The scatter-plot as well as the Q-Q plots suggest that this data is well described by a bivariate normal distribution, and thus it is a viable population model. One can easily imagine an oval shape in the scatter-plot encapsulating all the data points. The Q-Q plot indicate that the data variables are each approximately normally distributed.

Assignment 3

We are tasked to perform one-way MANOVA analysis on four skull data variables given for three different time periods (populations). We test whether the treatments are all zero at 95% significance level.

```
## [1] "Wilks Lambda: 0.83"
```

```
## Test statistic value for three groups and one or more variables for
##  this Wilk's Lambda: 2.05
```

```
## F-statistic for above situation at 5% significance level: 1.99
```

We see that the test statistic is greater than the F-statistic at the 5% significance level, so we reject the null hypothesis of all treatments are zero vectors. In order to see which variables changes between which time periods, we find the simultaneous confidence intervals of the treatment differences by vector component.

```
## [1] "Simultaneous 95% confidence intervals for treatment differences"
```

```
## [1] "between periods 3 and 1 for variable 1 : ( -0.342 , 6.54 )"
## [1] "between periods 3 and 2 for variable 1 : ( -1.34 , 5.54 )"
## [1] "between periods 2 and 1 for variable 1 : ( -2.44 , 4.44 )"
## [1] "between periods 3 and 1 for variable 2 : ( -3.37 , 3.77 )"
## [1] "between periods 3 and 2 for variable 2 : ( -2.47 , 4.67 )"
## [1] "between periods 2 and 1 for variable 2 : ( -4.47 , 2.67 )"
## [1] "between periods 3 and 1 for variable 3 : ( -6.91 , 0.647 )"
## [1] "between periods 3 and 2 for variable 3 : ( -6.81 , 0.747 )"
## [1] "between periods 2 and 1 for variable 3 : ( -3.88 , 3.68 )"
## [1] "between periods 3 and 1 for variable 4 : ( -2.33 , 2.39 )"
## [1] "between periods 3 and 2 for variable 4 : ( -2.03 , 2.69 )"
## [1] "between periods 2 and 1 for variable 4 : ( -2.66 , 2.06 )"
```

We see that the greatest treatment differences probably happen between periods 1 and 3. This is no surprise, as these time periods are furthest apart in time. we also notice that variable one and three (breath and length of skull) are the ones which change their means the most over time.

This data is appropriate for MANOVA, since I believe that the data is multivariate normally distributed since we are talking about data representative of an actual human population, and this type of physiological attributes which the variables consists of are usually normally distributed. It is not certain that the data is randomly sampled, since presumably the more well-off citizens of Ancient Egypt were more well preserved after death and thus stand a higher chance of being included in the data. I do not know whether the populations followed the same covariance matrix.

Appendix - R-Code

```
library(outliers)
library(car)
www = "http://www.ida.liu.se/~732A37/T1-9.dat"
www2 = "http://www.ida.liu.se/~732A37/T5-12.dat"
www3 = "http://www.ida.liu.se/~732A37/T6-13.dat"
data <- read.delim(www, header = FALSE, sep="\t")
data2 <- read.delim(www2, header = FALSE, sep="")
```

```
data3 <- read.delim(www3, header = FALSE, sep="")
colnames(data) <- c("NAT","100m(s)","200m(s)","400m(s)",
                    "800m(min)","1500m(min)","3000m(min)","Mara(min)")
colnames(data2) <- c("Tail length","Wing length")
colnames(data3) <- c("MaxBreath","BasHeight","BasLength","NasHeight","Time period")
# head(data)
# head(data2)
# head(data3)

nations <- as.character(data[,1])
data <- data[,-1]
col_mu <- colMeans(data)
covmatr <- var(data)

centerdata <- data - matrix(rep(col_mu, dim(data)[1]),ncol=7,byrow=TRUE)
centerdata <- as.matrix(centerdata)
covarianceadjusteddists <-c()

for(i in 1:dim(data)[1]){
  covarianceadjusteddists <- c( covarianceadjusteddists,
                                sqrt(centerdata[i,] %*% solve(covmatr)
                                     %*% centerdata[i,]))
}

for(i in 1:dim(data)[1]){
  if(covarianceadjusteddists[i]^2 > qchisq(0.999,7)){
    print(paste(nations[i]," is an outlier according to chisq test at 0.1% significance level"))
  }
}

alpha <- 0.95
col_mu2 <- colMeans(data2)
vardata2 <-var(data2)
lambda<-eigen(vardata2)
n <- dim(data2)[1]
p <- 2
aa <- c(1,0)
ab <- c(0,1)
quadraa <- aa %*% vardata2 %*% aa
quadrab <- ab %*% vardata2 %*% ab
csquared <- p*(n - 1) /(n * (n- p)) * qf(alpha,p,n-p)
evs <- sqrt(lambda$values * csquared)
evecs <- lambda$vectors

a <- evs[1]
b <- evs[2]
x0 <- col_mu2[1]
y0 <- col_mu2[2]
alpha <- atan(evecs[ , 1][2] / evecs[ , 1][1])
theta <- seq(0, 2 * pi, length=(1000))

x <- x0 + a * cos(theta) * cos(alpha) - b * sin(theta) * sin(alpha)
y <- y0 + a * cos(theta) * sin(alpha) + b * sin(theta) * cos(alpha)
```

```
plot(x, y, type = "l",main=c("Wing length (mm) vs tail length (mm) of female hook-billed kites",
                             "Mean vector 95% confidence ellipse"),
     xlab=names(data2)[1],ylab=names(data2)[2]
     )
points(col_mu2[1],col_mu2[2],cex=2,pch="+")


simuldista <- sqrt(csquared * quadraa)
simuldistb <- sqrt(csquared * quadrab)
t_val <- -qt(0.05/(2*p),n-1)
bonfdista <- t_val * sqrt(quadraa/n)
bonfdistb <- t_val * sqrt(quadrab / n)

paste("Simultaneous confidence interval for Tail length: (",round(col_mu2[1] - simuldista,2),
      ",",round(col_mu2[1] + simuldista,2),")")
paste("Simultaneous confidence interval for Wing length: (",round(col_mu2[2] - simuldistb,2),
      ",",round(col_mu2[2] + simuldistb,2),")")
paste("Bonferroni confidence interval for Tail length: (",round(col_mu2[1] - bonfdista,2),
      ",",round(col_mu2[1] + bonfdista,2),")")
paste("Bonferroni confidence interval for Wing length: (",round(col_mu2[2] - bonfdistb,2),
      ",",round(col_mu2[2] + bonfdistb,2),")")
plot(data2[,1],data2[,2],main="Wing length vs tail length of female hook-billed kites",
     xlab=names(data2)[1],ylab=names(data2)[2],pch="+",ylim=c(230,320),xlim=c(170,220))
qqnorm(data2[,1],main="Q-Q Plot of tail length")
qqline(data2[,1])
qqnorm(data2[,2],main="Q-Q Plot of wing length")
qqline(data2[,2])

old <- data3[1:30,1:4]
oldmu <- colMeans(old)
mid <- data3[31:60,1:4]
midmu <- colMeans(mid)
new <- data3[61:90,1:4]
newmu <- colMeans(new)

totmu <- colMeans(data3[,1:4])

betwold<-tcrossprod((oldmu - totmu),(oldmu - totmu))
betwmid<-tcrossprod((midmu - totmu),(midmu - totmu))
betwnew<-tcrossprod((newmu - totmu),(newmu - totmu))

Sold <- var(old)
Smid <- var(mid)
Snew <- var(new)


W <- 29 * (Sold + Smid + Snew)
B <- 30 * (betwold + betwmid + betwnew)
WilkLambda <- det(W) / det(B + W)

weirdexpr <- ((90-4-2)/(4)) * (1-sqrt(WilkLambda)) / sqrt(WilkLambda)
limit <- qf(0.95,8,168)
```

```
paste("Wilks Lambda:",signif(WilkLambda,3))
cat("Test statistic value for three groups and one or more variables for\n",
      "this Wilk's Lambda:",signif(weirdexpr,3))
cat("F-statistic for above situation at 5% significance level:",
    signif(limit,3))
p <- 4
g <- 3
m <- p * g * (g - 1) / 2
xbarmatr <- rbind(oldmu,midmu,newmu)

tterm <- -qt(0.05/(2 * m),87)
wterms <- sqrt(diag(W)*(2/30)*(1/87))

diff31 <- xbarmatr[3,] - xbarmatr[1,]
diff32 <- xbarmatr[3,] - xbarmatr[2,]
diff21 <- xbarmatr[2,] - xbarmatr[1,]

simulconfints31 <-c()
simulconfints32 <-c()
simulconfints21 <-c()
for(i in 1:4){
  simulconfints31[c(i,i+4)] <- c(diff31[i]-tterm*wterms[i],
                                 diff31[i]+tterm*wterms[i])
  simulconfints32[c(i,i+4)] <- c(diff32[i]-tterm*wterms[i],
                                 diff32[i]+tterm*wterms[i])
  simulconfints21[c(i,i+4)] <- c(diff21[i]-tterm*wterms[i],
                                 diff21[i]+tterm*wterms[i])
}

  paste("Simultaneous 95% confidence intervals for treatment differences")

for(i in 1:4){

  print(paste("between periods 3 and 1 for variable",i,": (",
        signif(simulconfints31[i],3),",",signif(simulconfints31[i+4],3),")"))
  print(paste("between periods 3 and 2 for variable",i,": (",
              signif(simulconfints32[i],3),",",signif(simulconfints32[i+4],3),")"))
  print(paste("between periods 2 and 1 for variable",i,": (",
              signif(simulconfints21[i],3),",",signif(simulconfints21[i+4],3),")"))
}
## NA
```