Assignment 2: Inference about mean vectors

Group 12

Dávid Hrabovszki (davhr856), Laura Julia Melis (lauju103), Spyridon Dimitriadis (spydi472), Vasileia Kampouraki (vaska979)

25/11/2019

Question 1: Test of outliers.

Top 5 five countries ranked by Mahalanobis distance (as calculated in Lab1)

```
##      SAM      PNG     KORN      COK      MEX
## 35.01406 30.50725 26.16714 19.83400 14.23093
```

1.a.

Outlier testing with no multiple-testing correction procedure at significance level: 0.1%

From Chi-square we get that the critical value, for $\alpha = 0.001$ is 24.32189, so every distance that is greater than this critical value is considered as an outlier.

```
##      SAM      PNG     KORN
## 35.01406 30.50725 26.16714
```

Outlier testing with Bonferroni multiple-testing correction procedure (with alpha = 0.1% / 54)

From Chi-square with Bonferroni correction we get that the critical value, for $\alpha = 0.001/54$ is 33.83184 so, again, every distance that is greater than this critical value is considered as an outlier.

```
##      SAM
## 35.01406
```

Using 0.1% significance levels and no correction procedure, we define 3 outliers in our dataset (SAM, PNG, KORN). Using the Bonferroni multiple-testing correction procedure, we conclude that only 1 observation is an outlier (SAM) based on the Mahalanobis distances.

According to McDonald in Handbook of Biological Statistics, the Bonferroni correction is appropriate when we want to be very careful not to get any false positives during the tests.

(http://www.biostathandbook.com/multiplecomparisons.html)

The correction happens at the expense of finding many false negatives, i.e. not finding outliers which actually are present. The Bonferroni approach might be useful in medical problems, but in our case, we believe that it is too conservative when it comes to classifying outliers.

The 0.1% significance level is considered to be low, as most tests are conducted at 5% level. This means that we want to be very certain that the outliers we define actually exist. Setting this level depends highly on the nature of the problem at hand, but in this case we think that a significance level of 5% would be more justifiable. This approach would result in 5 outliers.

1.b.

KORN seems like an outlier based on Mahalanobis distance, but not on Euclidean. This is because the Mahalanobis distance removes redundant information from correlated variables.

(https://waterprogramming.wordpress.com/2018/07/23/multivariate-distances-mahalanobis-vs-euclidean/)

With Mahalanobis distance we also use the relationships (covariances) between the variables (and not only the marginal variances as Euclidean distance does).

In the case of North Korea, running results were extreme in runtypes which have little correlation to other variables and low where the correlation was large.

Question 2: Test, confidence region and confidence inter- vals for a mean vector.

2.a.

The $100(1-\alpha)\%$ confidence region for (μ_1, μ_2) of a p-dimensional distribution is the ellipse detemined by all $\boldsymbol{\mu}$ such that:

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \leq c^2 = \frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)$$

In our case, the 95% confidence ellipse for $\boldsymbol{\mu}$ consists of all values (μ_1, μ_2) satisfying:

$$45 \cdot [193.62 - \mu_1, 279.78 - \mu_2] \begin{bmatrix} 0.02 & -0.01 \\ -0.01 & 0.01 \end{bmatrix} \begin{bmatrix} 193.62 - \mu_1 \\ 279.78 - \mu_2 \end{bmatrix} \leq c^2$$

where $c^2 = \frac{2(45-1)}{45-2} \cdot F_{2,43}(0.05) = 2.047 \cdot 3.215 = 6.579535$.

If (λ_i, e_i) are the eigenvalue-eigenvector pairs od S, then the i-th axis of the confidence ellipse has half length $\sqrt{p(n-1) \cdot F_{p,n-p}(\alpha)/(n-p)} \sqrt{\frac{\lambda_i}{n}}$ along the e_i direction.

Then, the axes of the confidence ellipse are:

$$\pm \sqrt{\lambda_i} \sqrt{\frac{p(n-1)}{n(n-p)} \cdot F_{p,n-p}(\alpha)} \cdot \boldsymbol{e}_i = \pm \sqrt{\lambda_i} \sqrt{c^2} \cdot \boldsymbol{e}_i$$

When we calculate the values for our data, we obtain the following eigenvalues-eigenvectors:

$$\lambda_1 = 294.60898, \quad e_1' = [0.5754, 0.8179]$$
$$\lambda_2 = 34.62637, \quad e_2' = [-0.8179, 0.5754]$$

So, beginning at the center $\bar{x}' = [193.62, 279.78]$, the axes of the 95% confidence ellipse are:

$$\text{major axis:} \quad \begin{bmatrix} 0.5753739 \\ 0.8178905 \end{bmatrix}$$
$$\text{minor axis:} \quad \begin{bmatrix} -0.8178905 \\ 0.5753739 \end{bmatrix}$$

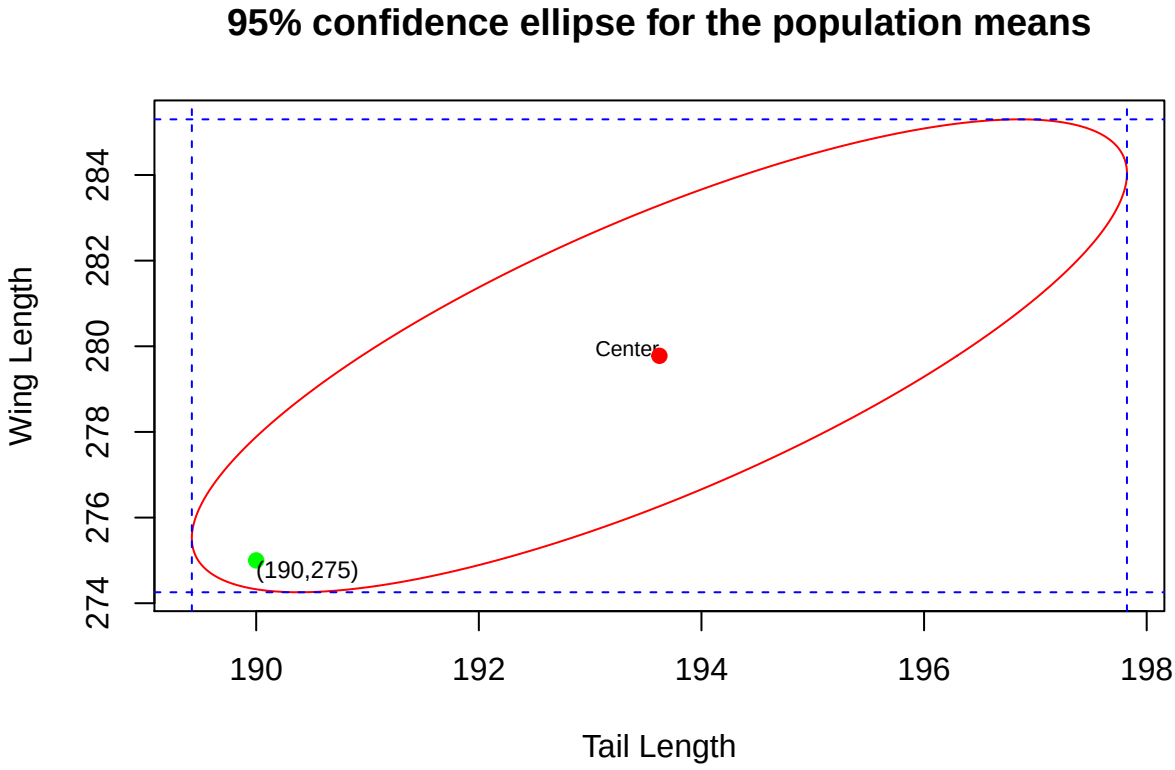
And the half length of each axis are:

$$\text{major axis half-length:} \quad + \sqrt{294.60898} \sqrt{0.1461883}$$

$$\text{minor axis half-length:} \quad + \sqrt{34.62637} \sqrt{0.1461883}$$

If we plot these results, we obtain the following graph [1]

[1] Sources: https://stackoverflow.com/questions/41820683/how-to-plot-ellipse-given-a-general-equation-in-r and https://stackoverflow.com/questions/15915625/plotting-an-ellipse-in-matlab-given-in-matrix-form



In the plot we can see the 95% confidence ellipse in red and also the confidence intervals in the dashed rectangle in blue. The green dot represents the population mean values for male hook-billed kites ($\mu' = [190, 275]$).

As we can observe, the green dot falls inside the ellipse, so we could conclude that these are in fact plausible values for the mean tail length and mean wing length for the female birds.

Also, we can confirm that $\mu' = [190, 275]$ is in the confidence region by computing the inequality explained above in this exercise:

$$45 \cdot [193.62 - 190, 279.78 - 275] \begin{bmatrix} 0.02 & -0.01 \\ -0.01 & 0.01 \end{bmatrix} \begin{bmatrix} 193.62 - 190 \\ 279.78 - 275 \end{bmatrix} \leq \frac{2(45-1)}{45-2} \cdot F_{2,45-2}(0.05)$$

$$5.54313 \leq 6.578471$$

2.b.

T^2 simultaneous confidence intervals:

$$a'\bar{x} - c\sqrt{\frac{a'Sa}{n}} \leq a'\mu \leq a'\bar{x} + c\sqrt{\frac{a'Sa}{n}}$$

where $c = \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} = \sqrt{6.578471} = 2.564853$.

So, the $100(1-\alpha)\%$ simultaneous 95% T^2-intervals are

$$\text{For } \mu_1 : \quad \bar{x}_1 \pm c\sqrt{\frac{s_{11}}{n}} = 193.62 \pm 2.565 \cdot 1.638$$

$$\Rightarrow \quad \boxed{189.4217 \leq \mu_1 \leq 197.8227}$$

$$\text{For } \mu_2: \quad \bar{x}_2 \pm c\sqrt{\frac{s_{22}}{n}} = 279.78 \pm 2.565 \cdot 2.153$$

$$\Rightarrow \quad \boxed{274.2564 \leq \mu_2 \leq 285.2992}$$

Bonferroni $100(1-\alpha)\%$ confidence intervals:

$$\bar{x}_i \pm t_{n-1}\Big(\frac{\alpha}{2m}\Big)\sqrt{\frac{s_{ii}}{n}}, \text{ for } i = 1, 2, \ldots, p$$

where m is the number of tests being carried out and $t_{n-1}\big(\frac{\alpha}{2p}\big) = t_{44}\big(\frac{0.05}{4}\big) = 2.32$.

So, the 95% Bonferroni intervals for the two population means are:

$$\text{For } \mu_1: \quad \bar{x}_1 \pm t_{44}\Big(\frac{0.05}{4}\Big)\sqrt{\frac{s_{11}}{n}} = 193.62 \pm 2.3207 \cdot 1.638$$

$$\Rightarrow \quad \boxed{189.8216 \leq \mu_1 \leq 197.4229}$$

$$\text{For } \mu_2: \quad \bar{x}_2 \pm t_{44}\Big(\frac{0.05}{4}\Big)\sqrt{\frac{s_{22}}{n}} = 279.78 \pm 2.3207 \cdot 1.638$$

$$\Rightarrow \quad \boxed{274.7819 \leq \mu_2 \leq 284.7736}$$

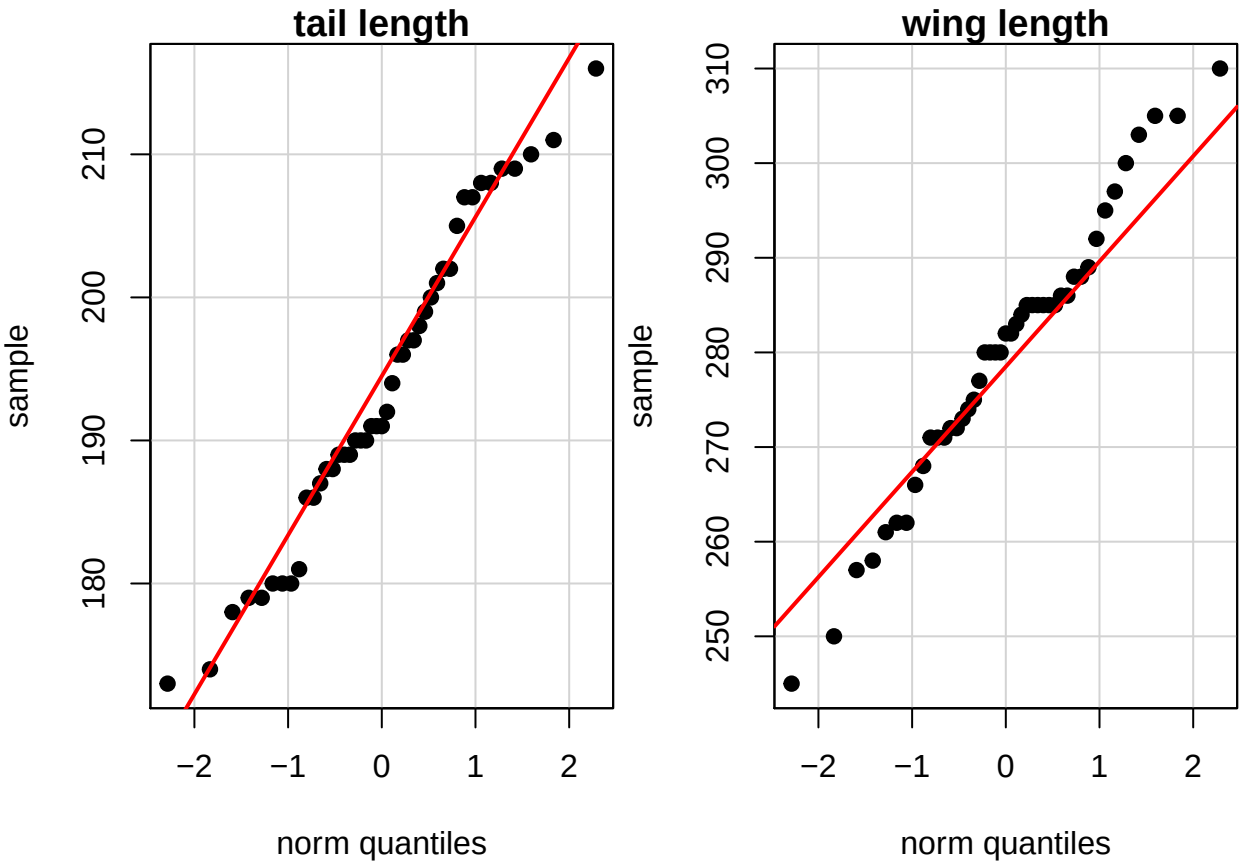
Comparison and advantage of T^2-intervals over Bonferroni.

The simultaneous T^2-intervals are a bit wider than the Bonferroni confidence interval in both variables, the Bonferroni interval falls within the T^2-interval. Bonferroni correction tries to ensure that the probability of declaring even one false positive is no more than, e.g., 5%. The Bonferroni correction declares as significant (rejects the null) any hypothesis where the p-value is $\leq 0.05/2m$.

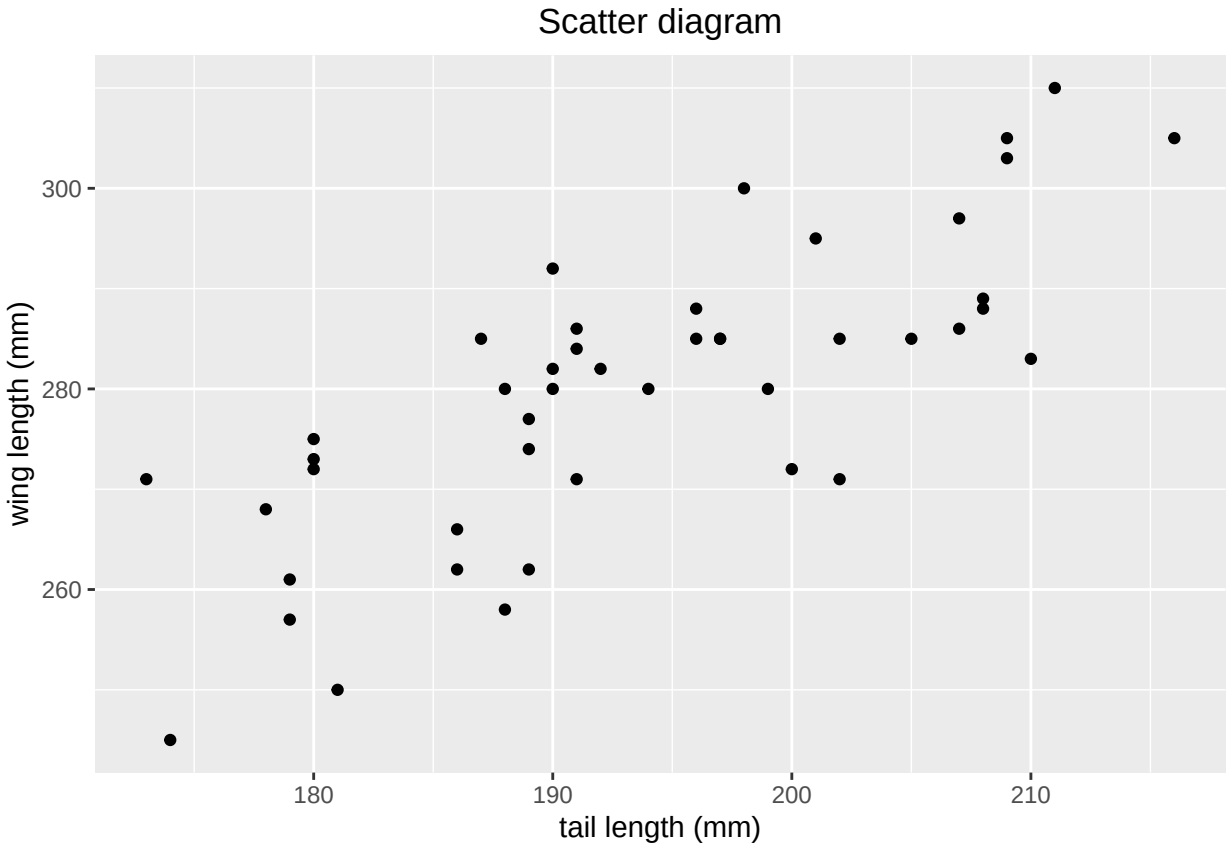
If we are interested only in the component means, the Bonferroni intervals provide more precise estimates than the T^2-intervals.

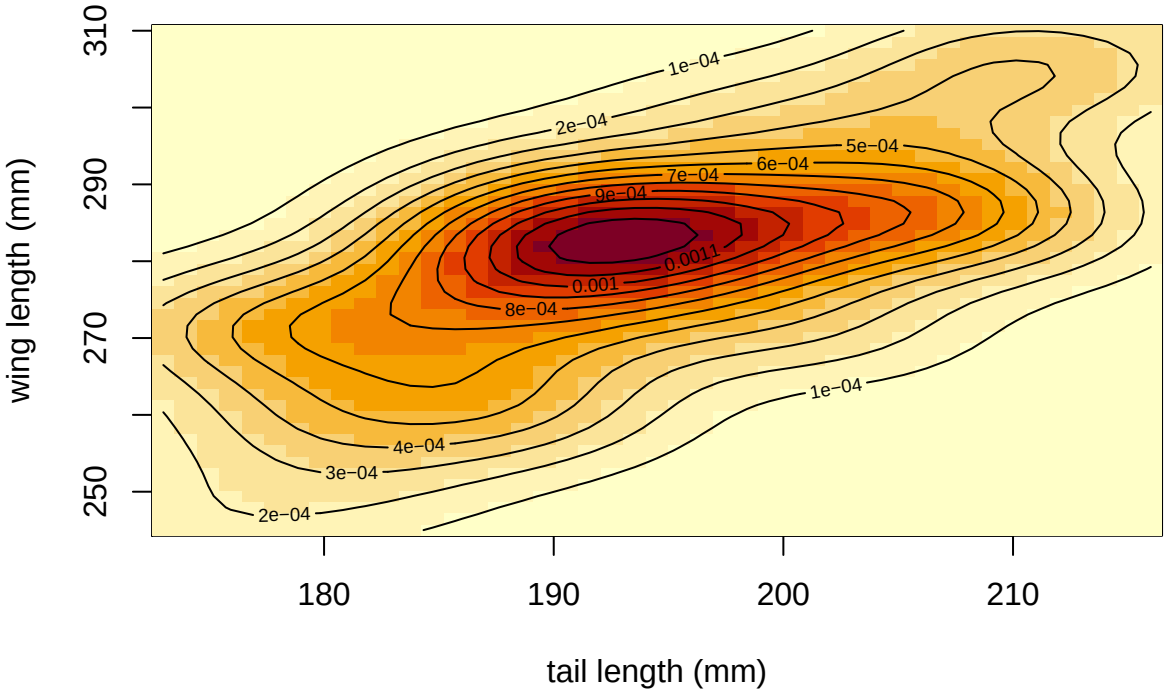
2.c.

We will explore if the data is plausibly coming from a bivariate Normal distribution analyzing Q-Q plots of each variable as well as the scatterplot of the observations and the sample contour:



We could say that the univariate normal distribution is not a viable model for each variable separately.





Question 3: Comparison of mean vectors (one–way 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.

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

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).

Tip: It might be helpful for you to read Exercise 6.24 of Johnson, Wichern. The function manova() can be useful for this question and the residuals can be found in the $res field.

Appendix

Exercise 2

```
# Importing and modifying data file
df = read.table("T5-12.dat")
colnames(df) <- c("Tail.length", "Wing.length")

# SUBQUESTION A
# A.0. Initial values
mu <- c(190,275)
xbar <- as.numeric(colMeans(df))
p <- 2
n <- nrow(df)
alpha <- 0.05
S <- cov(df)

# Axes of the confidence ellipse
c2 <- ((p*(n-1))/(n-p))*qf(1-alpha,df1=p,df2=n-p)
lambdas <- eigen(S)$values
e <-eigen(S)$vector

# A.1. Ellipse plot

# Sources 1: https://stackoverflow.com/questions/41820683/how-to-plot-ellipse-given-a-general-equation-
# Source 2: https://stackoverflow.com/questions/15915625/plotting-an-ellipse-in-matlab-given-in-matrix-

f_value <- qf(1-alpha,df1=p,df2=n-p)

# Points of the ellipse
theta <- seq(0, 2*pi, length = 1000)
r <- sqrt((n-1)*p/(n-p)*f_value/n)
v <- rbind(r*cos(theta), r*sin(theta))
z <- backsolve(chol(solve(S)), v)+xbar

# Confidence intervals (rectangle)
c <- sqrt(c2)
low1 <- xbar[1] - c*sqrt(S[1,1]/n)
upp1 <- xbar[1] + c*sqrt(S[1,1]/n)
low2 <- xbar[2] - c*sqrt(S[2,2]/n)
```

```r
upp2 <- xbar[2] + c*sqrt(S[2,2]/n)

# Plot of the ellipse
{plot(t(z), type="l", xlab="Tail Length", ylab="Wing Length",
      main="95% confidence ellipse for the population means", col="red")
text(xbar[1], xbar[2], "Center", cex=0.7, adj = c(1,0))
points(xbar[1], xbar[2], pch=19,col="red")
points(mu[1], mu[2], pch=19,col="green")
text(mu[1], mu[2], "(190,275)", cex=0.75, adj = c(0,1))
points(df$Tail.length, df$Wing.length, pch=20,col="black")
abline(v=low1, col="blue", lty=2)
abline(v=upp1, col="blue", lty=2)
abline(h=low2, col="blue", lty=2)
abline(h=upp2, col="blue", lty=2)
}

# A.2. Mu inside the ellipse
right_side<- n*(xbar-mu)%*%solve(S)%*%(xbar-mu)
result <- right_side <= c2 #TRUE

# SUBQUESTION B
# B.1. T2 simultaneous intervals
CI1 <- c(low1, upp1)
CI2 <- c(low2, upp2)

# B.2. Bonferroni Intervals
t <- qt(1-alpha/(2*p),df=n-1)

low1 <- xbar[1] - t*sqrt(S[1,1]/n)
upp1 <- xbar[1] + t*sqrt(S[1,1]/n)
low2 <- xbar[2] - t*sqrt(S[2,2]/n)
upp2 <- xbar[2] + t*sqrt(S[2,2]/n)

bonf1 <- c(low1, upp1)
bonf2 <-c(low2, upp2)

# SUBQUESTION C
par(mfrow=c(1,3), mar=c(4,4,1,0)+0.1)
{qqnorm(df$Tail.length, pch = 1, frame = FALSE, main="Q-Q Plot for Tail Length")
qqline(df$Tail.length, col = "steelblue", lwd = 2)}
{qqnorm(df$Wing.length, pch = 1, frame = FALSE, main="Q-Q Plot for Wing Length")
qqline(df$Wing.length, col = "steelblue", lwd = 2)}
plot(df$Tail.length, df$Wing.length,pch=19, cex=0.8, main = "Scatterplot of Tail vs Wing",
     xlab="Tail Length", ylab="Wing Length")
```