Multivariate Statistical Methods

Assignment 2

Allan Gholmi, Emma Wallentinsson, Rasmus Holm

2017-12-07

Question 1

a)

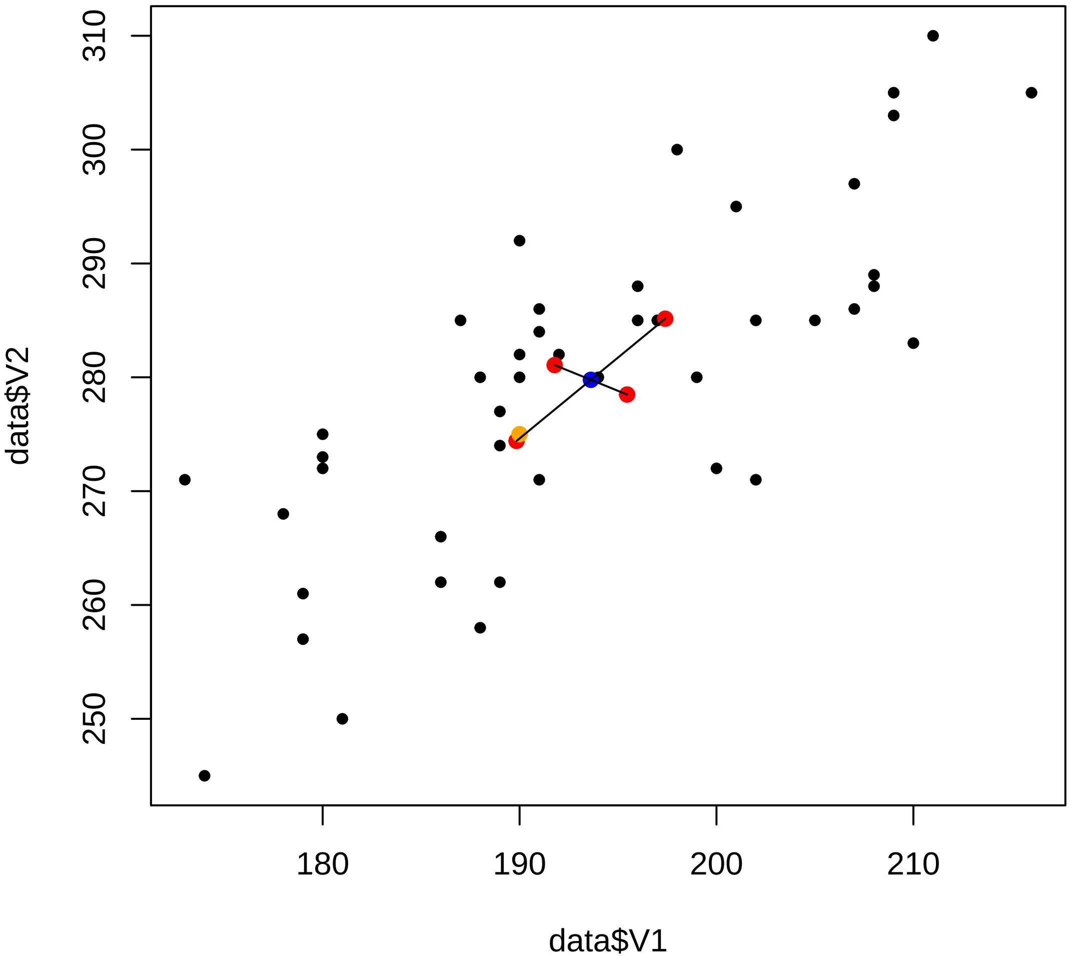
```
#> [1] "Outliers without correction"
#> [1] "COK"  "KORN" "MEX"  "PNG"  "SAM"
```

No clue what the multiple-testing correction procedure refers to.

b)

Question 2

a)

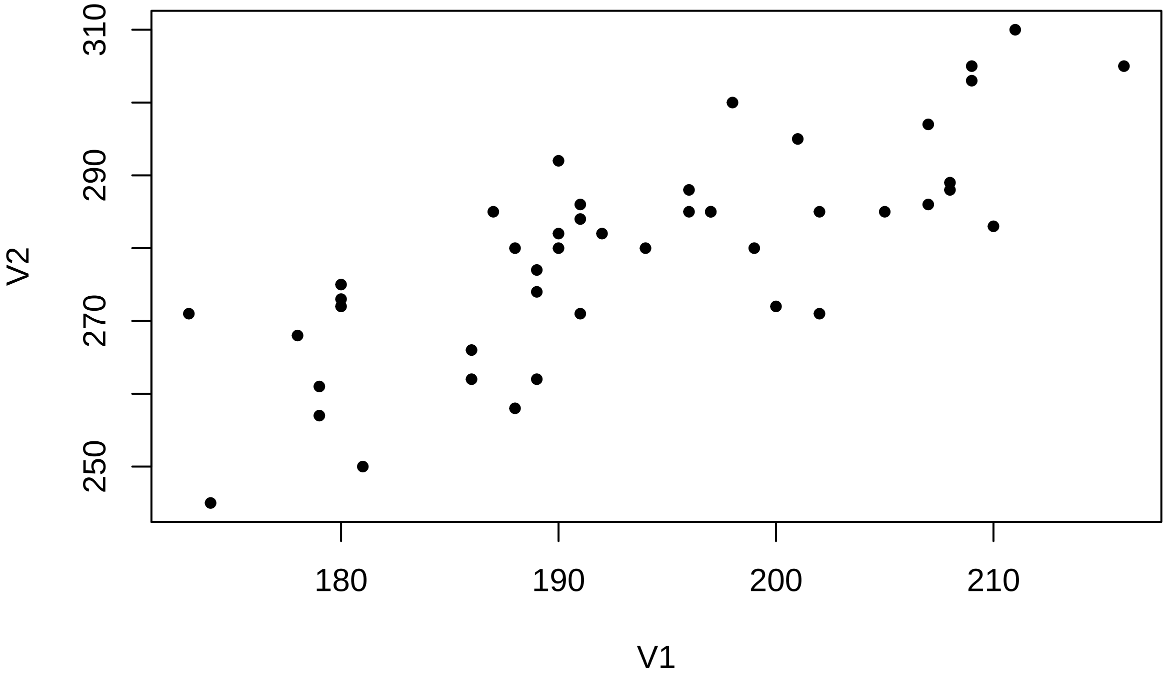


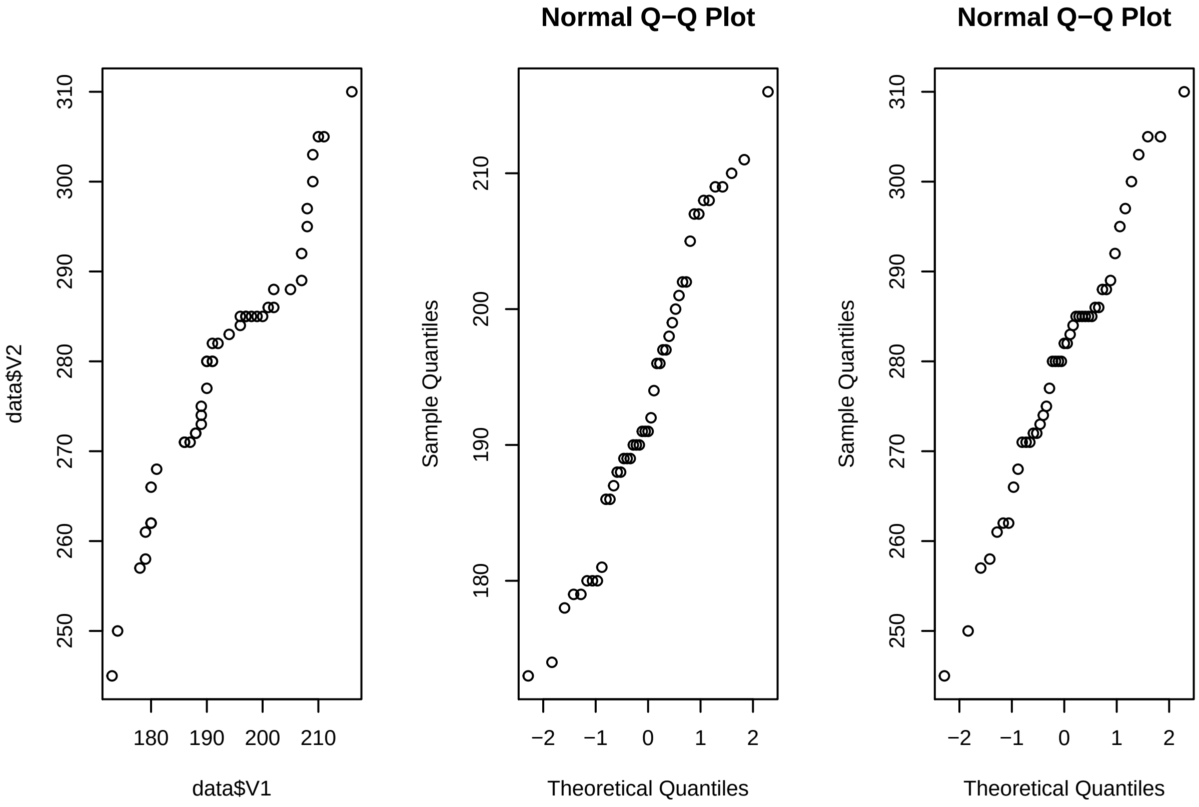
b)

```
#> [1] "T-square Intervals"
#>            V1       V2
#> [1,] 189.4217 274.2564
#> [2,] 197.8227 285.2992
#> [1] "Bonferroni Intervals"
#>            V1       V2
#> [1,] 189.8216 274.7819
#> [2,] 197.4229 284.7736
```

T-square test always gives wider confidence intervals since it takes the correlation between the measured variables into account. Bonferroni intervals are more precise if you are interested in the individual component means, but if you are interested in the overall data mean you should consider the T-square intervals.

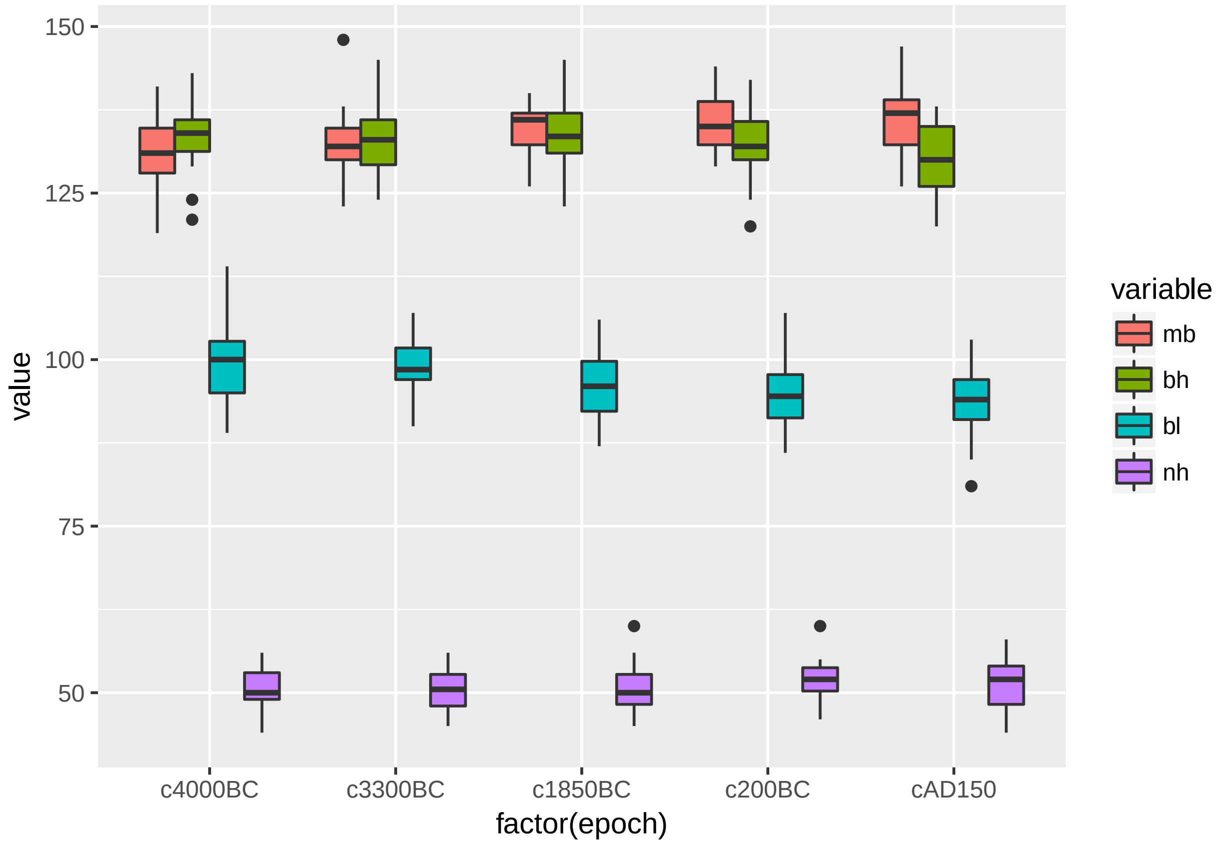
c)




data$V2
310
300
290
280
270
260
250
180
190
200
210
data$V1
Normal Q−Q Plot
Sample Quantiles
210
200
190
180
−2
−1
0
1
2
Theoretical Quantiles
Normal Q−Q Plot
Sample Quantiles
310
300
290
280
270
260
250
−2
−1
0
1
2
Theoretical Quantiles

Question 3

a)

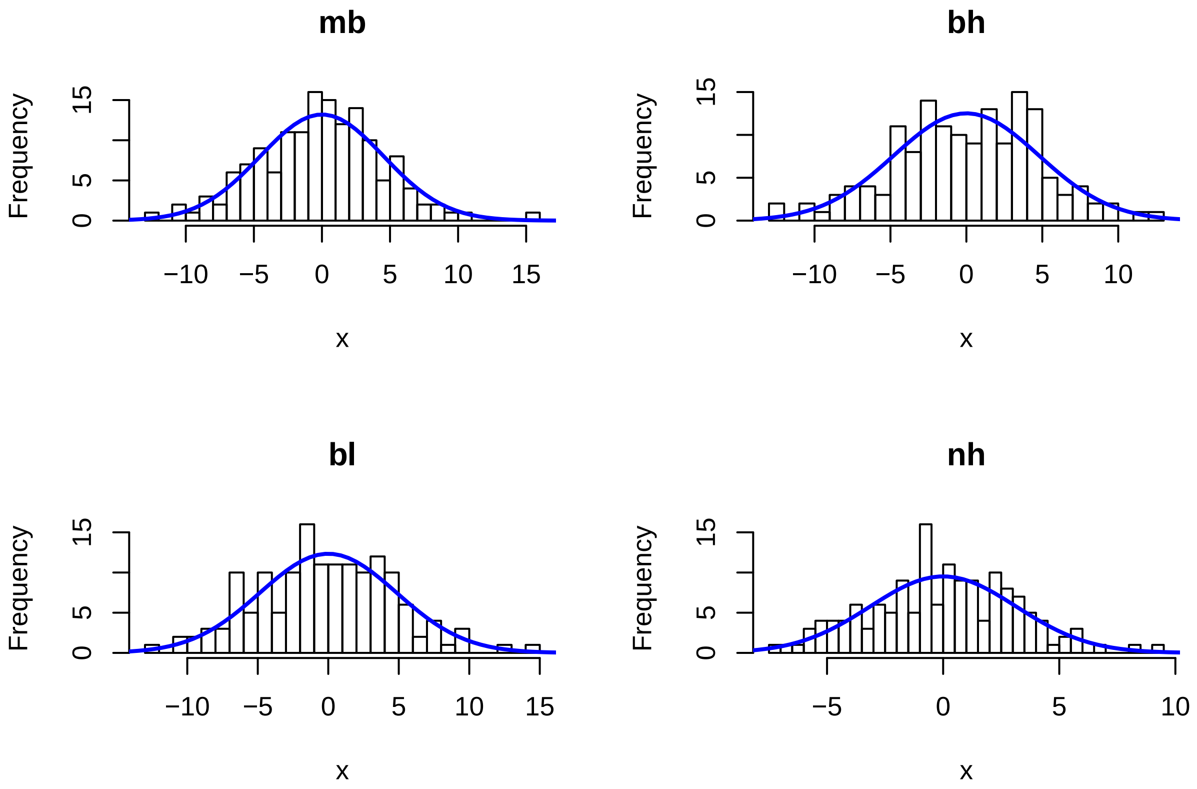


b)

```
group_means <- data %>%
    group_by(epoch) %>%
    summarise_all(funs(mean(., na.rm=TRUE)))


fit <- manova(cbind(mb, bh, bl, nh) ~ data$epoch, data)
```

c)


mb
Frequency
0
5
15
−10
−5
0
5
10
15
x
bh
Frequency
0
5
15
−10
−5
0
5
10
x
bl
Frequency
0
5
15
−10
−5
0
5
10
15
x
nh
Frequency
0
5
15
−5
0
5
10
x

Appendix

Code

```
# Question 1
data <- read.table("../data/T1-9.dat")
names(data) <- c("country", "100m", "200m", "400m", "800m", "1500m", "3000m", "marathon")
numeric_data <- data[, -1]

countries <- as.character(data$country)
X <- as.matrix(numeric_data)
means <- colMeans(X)
covariances <- cov(X)
X_central <- X - rep(1, nrow(X)) %*% t(means)

mdist_sq <- X_central %*% solve(covariances) %*% t(X_central)
country_mdist <- diag(mdist_sq)

significance_level <- 0.1
p <- ncol(X)
quantile <- qchisq(1 - significance_level, df=p)

outliers <- country_mdist > quantile
print("Outliers without correction")
countries[outliers]


# Question 2
data <- read.table("../data/T5-12.DAT")
x_bar <- colMeans(data)
S <- cov(data)
S_inv <- solve(S)

n <- nrow(data)
p <- ncol(data)

eigen_values <- eigen(S)$values
eigen_vectors <- eigen(S)$vectors

true_mean <- c(190, 275)
confidence_level <- 0.05

half_lengths <- sqrt(eigen_values) * sqrt((p * (n - 1)) / (n * (n - p)) *
                                          qf(1 - confidence_level, df1=p, df2=n - p))

p1 <- x_bar + eigen_vectors[, 1] * half_lengths[1]
p2 <- x_bar - eigen_vectors[, 1] * half_lengths[1]

p3 <- x_bar + eigen_vectors[, 2] * half_lengths[2]
p4 <- x_bar - eigen_vectors[, 2] * half_lengths[2]

x <- c(p1[1], p2[1], p3[1], p4[1])
y <- c(p1[2], p2[2], p3[2], p4[2])
```

```
plot(data$V1, data$V2, pch=20)
points(x, y, col="red", pch=20, cex=1.5)
points(true_mean[1], true_mean[2], col="orange", pch=20, cex=1.5)
points(x_bar[1], x_bar[2], col="blue", pch=20, cex=1.5)
segments(rep(x_bar[1], 4), rep(x_bar[2], 4), x, y)
Tsq_offset <- sqrt(p * (n - 1) * qf(1 - confidence_level, df1=p, df2=n - p) / (n - p) * diag(S) / n)
Tsq_confidence_interval <- rbind(x_bar - Tsq_offset, x_bar + Tsq_offset)

bonferroni_offset <- sqrt(diag(S) / n) * qt(1 - confidence_level / (2 * p), df=n - 1)
bonferroni_confidence_interval <- rbind(x_bar - bonferroni_offset, x_bar + bonferroni_offset)

print("T-square Intervals")
Tsq_confidence_interval

print("Bonferroni Intervals")
bonferroni_confidence_interval


# Question 3
library(heplots)
library(dplyr)
library(ggplot2)
library(reshape2)

data <- Skulls
numeric_data <- data[, -1]
colors <- as.numeric(data$epoch)
# pairs(numeric_data, col=colors)
mm <- melt(data, id="epoch")
ggplot(mm) +
    geom_boxplot(aes(x=factor(epoch), y=value, fill=variable))
group_means <- data %>%
    group_by(epoch) %>%
    summarise_all(funs(mean(., na.rm=TRUE)))


fit <- manova(cbind(mb, bh, bl, nh) ~ data$epoch, data)
```