732A97 - Multivariate statistical methods

Lab 2 - "Inference about mean vectors"

Alexander Karlsson (aleka769)

2019-11-28

Contents

Packages used can be seen below.

```
library(dplyr)
library(car)
```

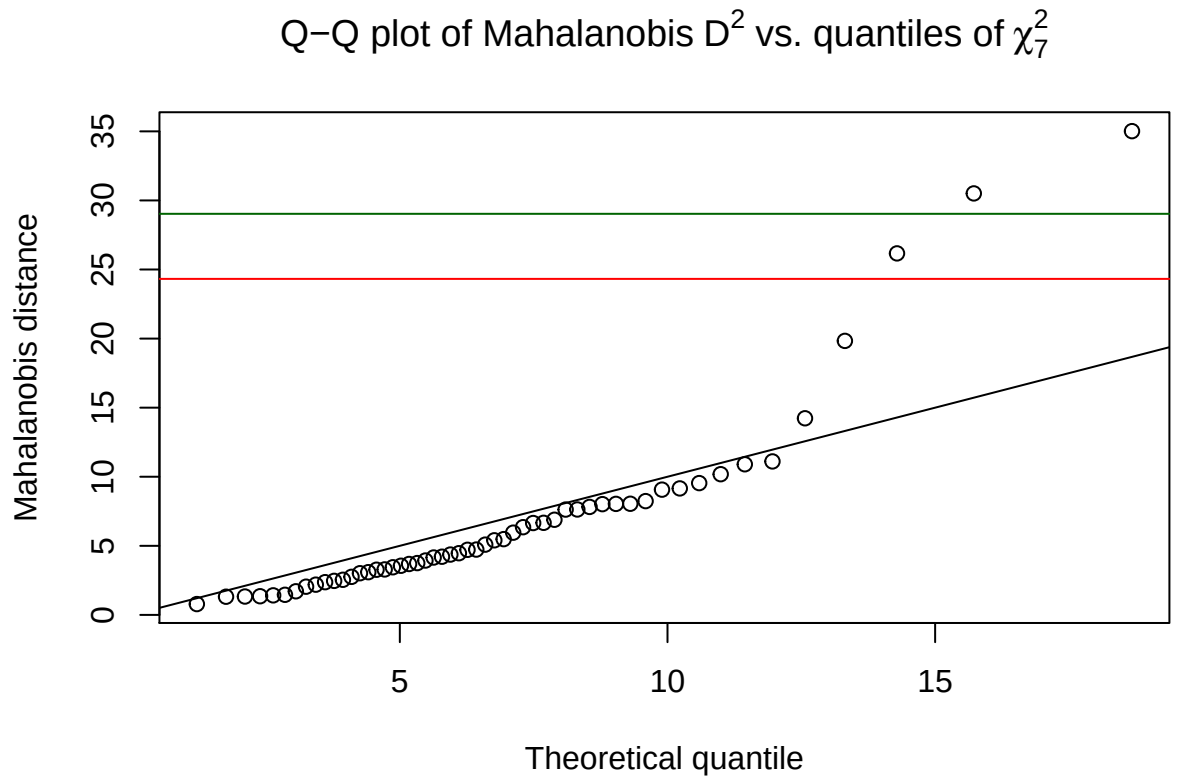
```
trackData = read.table(file        = "T1-9.dat",
                       col.names   = c("Country", "100m", "200m", "400m",
                                       "800m", "1500m", "3000m", "42000m"),
                       check.names = FALSE)

birdData = read.table(file        = "T5-12.dat",
                      col.names   = c("tailLength", "wingLength"))
```

Question 1 - Describing individual variables

Multiple testing correction in χ^2

Simultaneous testing is done with a change of α when computing the χ^2 critical value. For 0.01 % significance level without correction $1 - \alpha = 0.999$ and with correction $1 - (\alpha/p)$ is used as input to the χ^2-quantile function.

```
# Manipulate data:
X = as.matrix(trackData[2:8])
rownames(X) = trackData$Country
Xc = scale(X, scale = FALSE)

# Calculate covariance, distances:
V = cov(X)
Dsq = Xc %*% solve(V) %*% t(Xc)

# Plot quantile, quantile vs outlier position:
qqplot(qchisq(ppoints(54), df = 7), diag(Dsq),
       main = bquote("Q-Q plot of Mahalanobis" * ~D^2 *
                       " vs. quantiles of" * ~ chi[7]^2),
       xlab = "Theoretical quantile", ylab = "Mahalanobis distance")
abline(a = 0, b = 1)
abline(h = qchisq(1-0.001, 7), col = "red")
abline(h = qchisq(1-(0.001 / 7), 7), col = "darkgreen")
```



Countries KORN, PNG, SAM are outliers without correction and PNG, SAM are outliers with correction.

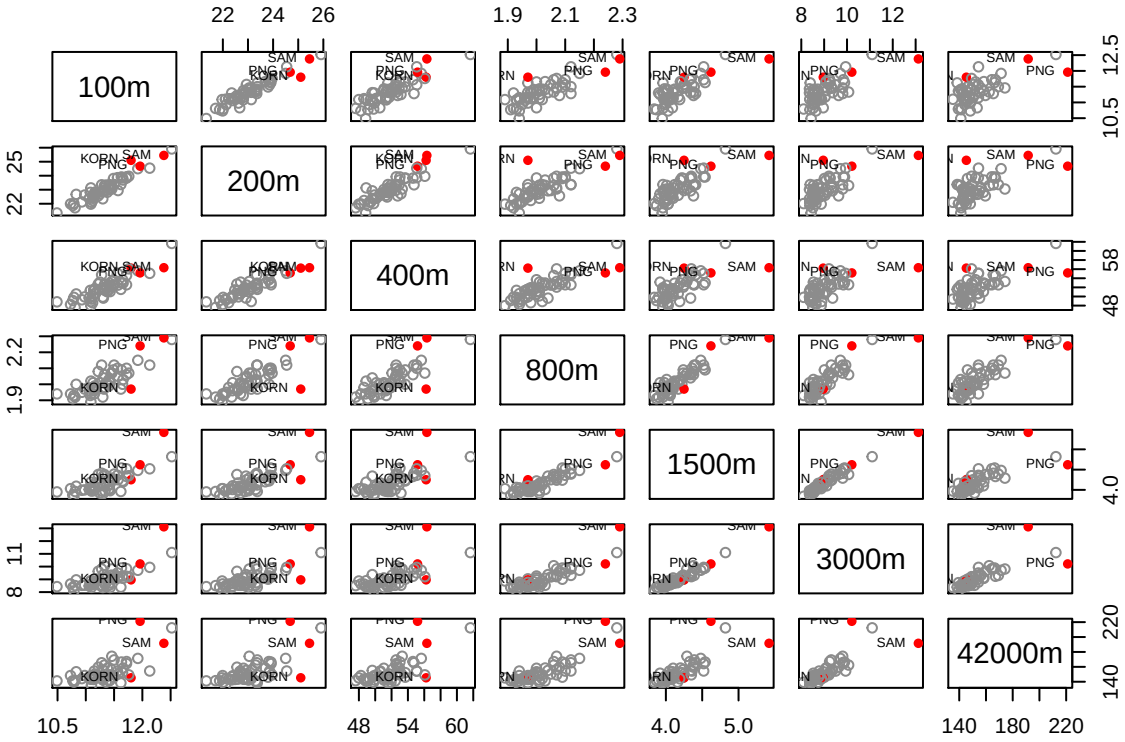
0.1 % is not a sensible level because many countries that deviate heavily from the black line (QQ-line). A more sensible approach is to use some standard significance-level, such as $\alpha = 0.05$ and use correction on that

significance level instead.

Mahalanobis and Euclidean contradiction

Mahalanobis calculations considers the covariance matrix for variables, whereas Euclidean distances considers the identity matrix. If a country is an outlier when computing Mahalanobis distance but not when computing the Euclidean distance, then it means this country is close to the mean of the population but behaves differently than others when looking at different variables (distances).

```
w_ = trackData$Country %in% c("KORN", "PNG", "SAM") # Index vector
n_ = trackData$Country                              # Name vector
c_ = ifelse(w_, "red", "gray55")                    # Color vector

pairs(x = trackData[2:8],
      panel = function (x, y, ...) {
        d = data.frame(n = n_, x, y)
        points(x, y, col = c_, pch = c(1,16)[w_+1])
        with(d[w_,], text(x, y, cex = 0.6, labels = n, pos = 2))
      })
```



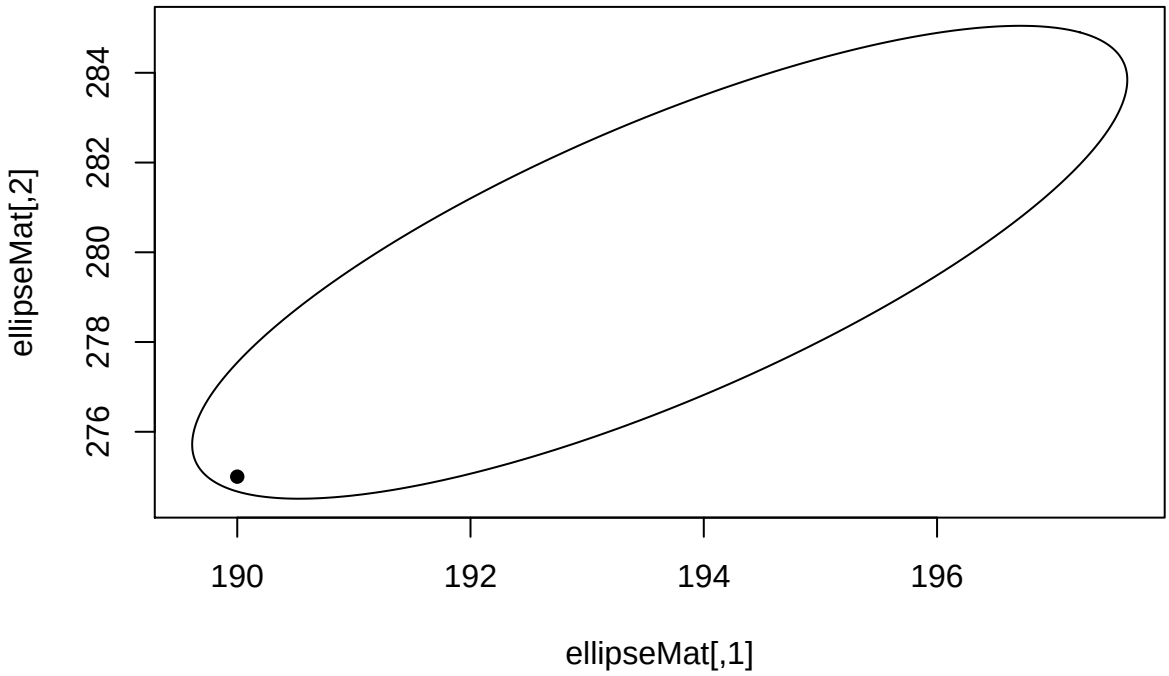
Question 2 - Relationships between the variables

Bivariate ellipse for sample mean

The mean-vector values given in the assignment are compared to the ellipse centered around data means $\bar{X}$ and directions given by the eigendecomposition of S.

The shape of the ellipse is determined according to equations given in lecture 5. The ellipse is a 95 % confidence region for the mean vector of female birds, which means that if the assumed mean vector given falls within this region then no differences can be assumed.

```
X = as.matrix(birdData)
n = nrow(X)
p = ncol(X)

Chi = qchisq(.95, p)

E = eigen(cov(X))$vectors
l = eigen(cov(X))$values

l1 = sqrt(l[1])*sqrt(Chi) / sqrt(n)
l2 = sqrt(l[2])*sqrt(Chi) / sqrt(n)

angles = seq(0, 2*pi, length.out = 500)

ellipseMat <- cbind(l1*cos(angles), l2*sin(angles)) %*% t(E)
ellipseMat <- ellipseMat + matrix(colMeans(X), 500,2,T)

plot(ellipseMat, type="l", lwd=1)
points(190, 275, pch=16)
```



Since the mean vector for male bidrs $\vec{\mu} = [190, 275]$ can be found within the 95% confidence region (ellipse) for the mean vector of female birds, there are no significant differences on 95% confidence level between the

female and male birds. However, one can discuss if the number of data points are sufficiently many. It is deemed that $n = 45$ is enough.

Bonferroni and T^2 intervals

T^2: H_0 is rejected if $n(\bar{x} - \vec{\mu}_0)^T S^{-1}(\bar{x} - \vec{\mu}_0) \geq \frac{(n-1)p}{np} F_{[p,n-p]}(\alpha)$.

```
n = nrow(X)
p = ncol(X)

xbar_k = colMeans(X)
S_kk   = diag(cov(X), names = F)

stdErrBonf = qt(p = 1-(.05/p), df = n-1) * (sqrt(S_kk) / sqrt(n))
stdErrT2   = sqrt(((n-1)*p)/(n*p)) * sqrt(qf(p = .95, df1 = p, df2 = n-p)) * (sqrt(S_kk) / sqrt(n))

data.frame(
  lower_Bonf = xbar_k - stdErrBonf,
  upper_Bonf = xbar_k + stdErrBonf,
  lower_T2   = xbar_k - stdErrT2,
  upper_T2   = xbar_k + stdErrT2
  ) %>% t()
```

```
##            tailLength wingLength
## lower_Bonf   190.3216   275.4392
## upper_Bonf   196.9228   284.1163
## lower_T2     190.7188   275.9613
## upper_T2     196.5257   283.5943
```

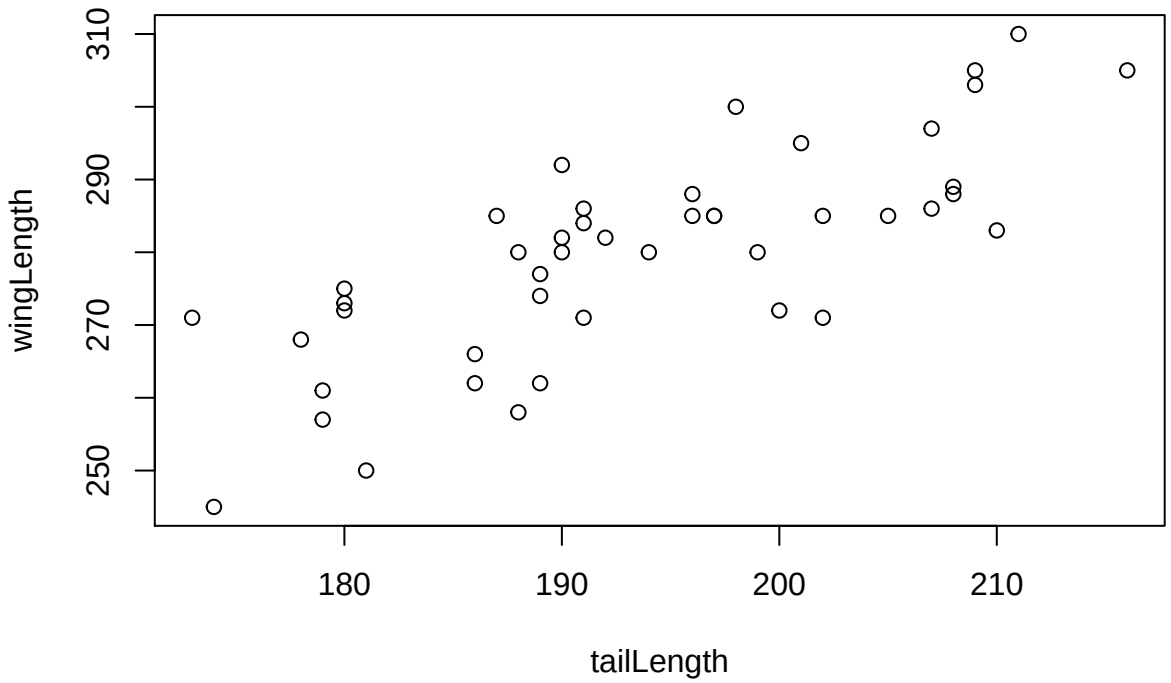
The T^2 intervals are a bit more narrow, for both variables. T^2 is also invariant under rotations and no independence assumptions must be made. Simultaneous Bonferroni testing assumes independence, but x_1 and x_2 are likely not independent. If two variables are independent, they are also uncorrelated, which is not the case for the data set that is analyzed. Wing length and tail length is correlated with the following sample correlation matrix:

```
cor(X) %>% round(2)
```

```
##            tailLength wingLength
## tailLength       1.00       0.77
## wingLength       0.77       1.00
```

Is data bivariate Gaussian?

```
plot(X)
```

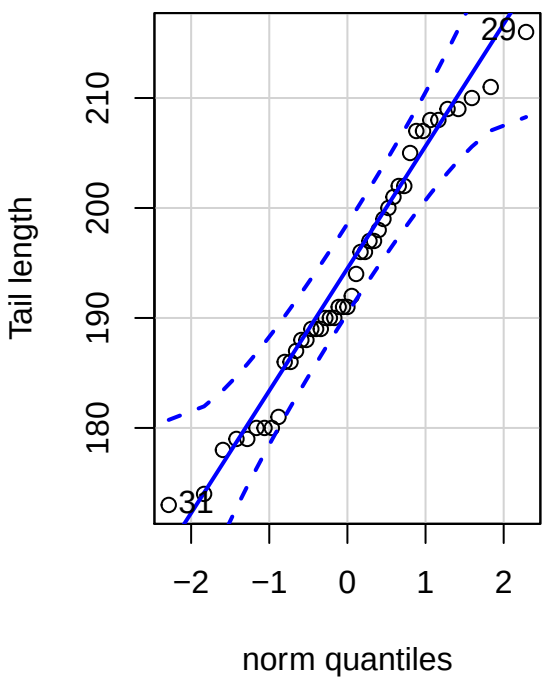


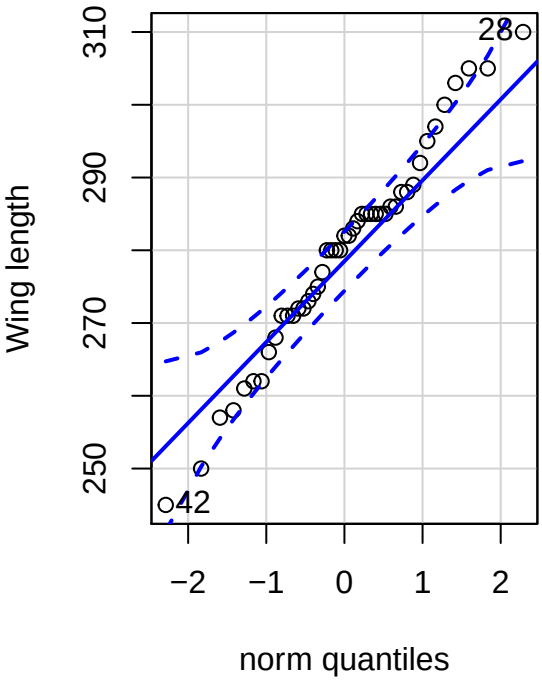
The data definitely looks approximately ellipse-shaped, and thus a bivariate normal with a positive correlation would fit well. One can study the marginal distributions separately in Qunatile-Quantile-plots, where the z-scores are calculated for each data point. The points should lie within a confidence interval for an independent(!) Gaussian to be assumed. For both variables (plotted below), these assumptions are a bit shaky.

```
par(mfrow = c(1,2))
qqPlot(x = X[,1], distribution = "norm", ylab = "Tail length")
```

```
## [1] 29 31
```

```
qqPlot(x = X[,2], distribution = "norm", ylab = "Wing length")
```


Tail length
180
190
200
210
29
31
−2
−1
0
1
2
norm quantiles


Wing length
250
270
290
310
28
42
−2
−1
0
1
2
norm quantiles

```
## [1] 42 28
```

Question 3 - Examining for extreme values