Lab 2 732A97 Multivariate Statistical Methods

Raymond Sseguya

2019-11-29

Inference about mean vectors

Question 1: Test of outliers

Consider again the data set from the T1-9.dat file, National track records for women. In the first assignment we studied different distance measures between an observation and the sample average vector. The most common multivariate residual is the Mahalanobis distance and we computed this distance for all observations.

a) The Mahalanobis distance is approximately chi-square distributed, if the data comes from a multivariate normal distribution and the number of observations is large. Use this chi-square approximation for testing each observation at the 0.1% significance level and conclude which countries can be regarded as outliers. Should you use a multiple-testing correction procedure? Compare the results with and without one. Why is (or maybe is not) 0.1% a sensible significance level for this task?

```
trackrcs <- read.table("T1-9.dat",
  col.names = c("countries", "x100m", "x200m",
        "x400m", "x800m", "x1500m", "x3000m", "marathon"))

trackrcs2 <- (trackrcs)[,-1]
rownames(trackrcs2) <- trackrcs[,1]

C <- cov((trackrcs)[,-1])
x_bar = apply(trackrcs2,1,mean)
d0 = as.matrix(trackrcs2-x_bar)
deviation = sqrt( d0%*%t(d0) )

d_sq_m <- d0%*%solve(C)%*%t(d0)
diagonal_vector3 <- diag(d_sq_m)
deviation_countries3 <-
  cbind.data.frame(countries = as.vector(trackrcs[,1]),diagonal_vector3)
deviation_countries_ordered3 <-
   deviation_countries3[order(-deviation_countries3$diagonal_vector3), ]


named_Mahalanobis <- as.vector(deviation_countries_ordered3[,2])
names(named_Mahalanobis) <- rownames(deviation_countries_ordered3)
ch_s <- combn(x=named_Mahalanobis, m=2,
      FUN = function(c){
```

```
      sg <<- 0.1/100
      pv <- chisq.test(x=c, p=rep(sg,2), rescale.p = TRUE)$p.value
      pvname <- paste0(names(c)[1]," ",names(c)[2])
      assign(pvname, pv)
      return(list(pv, pvname, xx=names(c)[1]))
      })

outliers_unsorted <- unlist(ch_s[3, which(ch_s[1,] == 0)])
summary(as.data.frame.character(outliers_unsorted), maxsum = 10)
```

```
##  outliers_unsorted
##  COK    : 52
##  PNG    : 52
##  SAM    : 51
##  GUA    : 49
##  BER    : 48
##  MRI    : 44
##  CRC    : 43
##  DOM    : 43
##  MAS    : 43
##  (Other):632
```

```
print(deviation_countries_ordered3[1:5,])
```

```
##     countries diagonal_vector3
## COK       COK          2094451
## PNG       PNG          2083382
## SAM       SAM          1793652
## GUA       GUA          1485602
## BER       BER          1472580
```

We can see from the summary that "COK", "PNG", "SAM", "GUA" and "BER" remain the top oultiers even with using the chi-sqaure test.

possibly using a multiple-testing correction procedure

```
chisq.test(deviation_countries_ordered3$diagonal_vector3,
           p=rep(0.1/100, nrow(deviation_countries_ordered3)), rescale.p = TRUE )
```

```
##
##  Chi-squared test for given probabilities
##
## data:  deviation_countries_ordered3$diagonal_vector3
## X-squared = 2034800, df = 53, p-value < 2.2e-16
```

We can see that a multiple-testing correction procedure is clearly a bad idea as it does not actually pinpoint the actual outliers but it tells us only that there are indeed some outliers.

correct way

```
### the critical value
which(deviation_countries3[,2] > qchisq(p=0.1/100, df=54))
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
## [24] 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
## [47] 47 48 49 50 51 52 53 54
```

b) One outlier is North Korea. This country is not an outlier with the Euclidean distance. Try to explain these seemingly contradictory results.

It seems in our results, North Korea is an outlier throughout.

Question 2: Test, confidence region and confidence intervals for a mean vector

Look at the bird data in file T5-12.dat and solve Exercise 5:20 of Johnson, Wichern. Do not use any extra R package or built-in test but code all required matrix calculations. You MAY NOT use loops!

```
birds <- read.table("T5-12.DAT", col.names = c("taillength", "winglength") )
```

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

```
S <- cov(birds); eigen(S)
```

```
## eigen() decomposition
## $values
## [1] 294.60898  34.62637
##
## $vectors
##           [,1]       [,2]
## [1,] 0.5753739 -0.8178905
## [2,] 0.8178905  0.5753739
```

```
p <- ncol(birds); n <- nrow(birds); alpha <- 0.05
c = sqrt(p*(n-1)/(n-p)*qf(p=(1-alpha), df1=p, df2=n-p ))
(eigen(S)$values)*c/sqrt(n)
```

```
## [1] 112.64249  13.23924
```

Yes. These are plausible values.

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?

95% T^2-intervals

```
mu0 <- c(190, 275)
# mu <- apply(birds, 2, mean)
# T_sq <- as.vector(n*t(mu-mu0)%*%solve(S)%*%(mu-mu0)); T_sq < c

c(mu0[1] - (c*sqrt(diag(S)[1]/n) ), mu0[1] + (c*sqrt(diag(S)[1]/n) ))
```

```
## taillength taillength
##   185.7995   194.2005
```

```
c(mu0[2] - (c*sqrt(diag(S)[2]/n) ), mu0[2] + (c*sqrt(diag(S)[2]/n) ))
```

```
## winglength winglength
##   269.4786   280.5214
```

95% Bonferroni intervals

```
t <- qt(p=( 1-(alpha/(2*p)) ), df=n-1)

c(mu0[1] - (t*sqrt(diag(S)[1]/n) ), mu0[1] + (t*sqrt(diag(S)[1]/n) ))
```

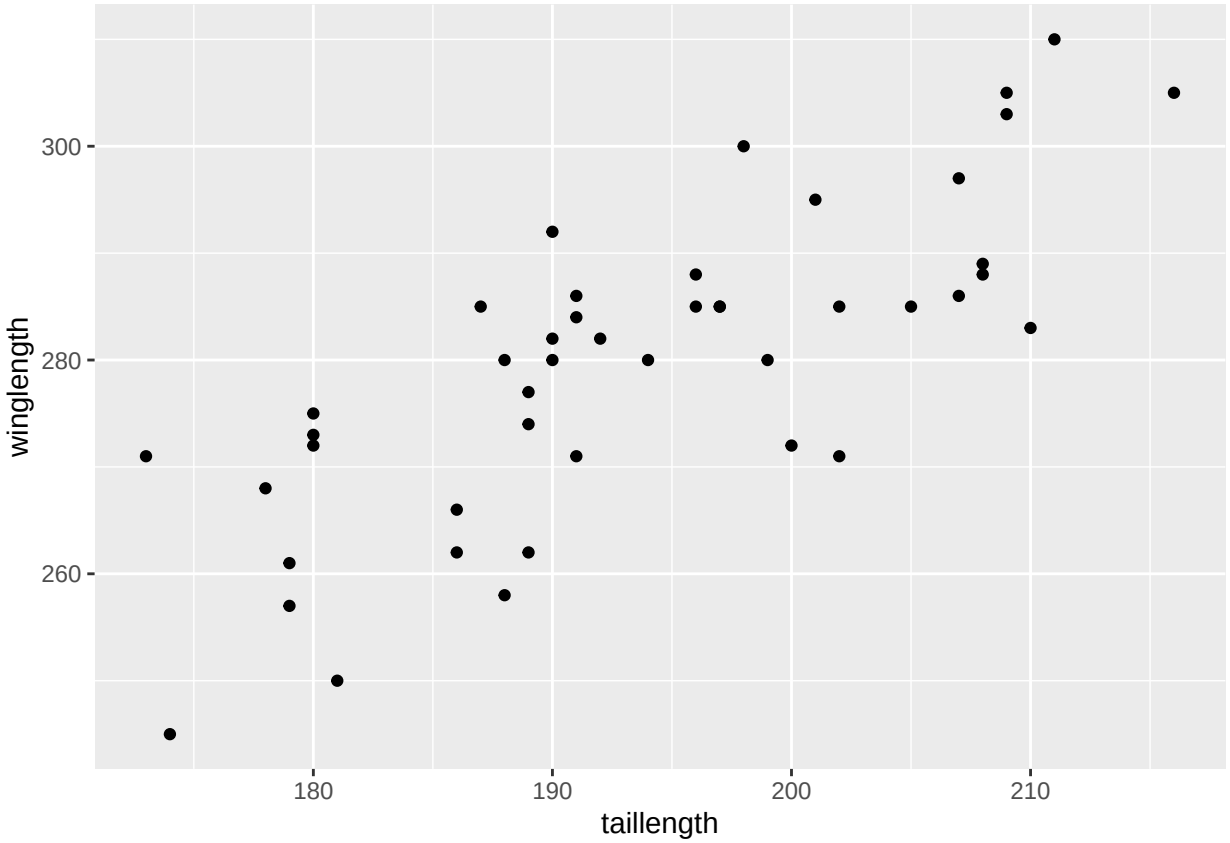
```
## taillength taillength
##   186.1993   193.8007
```

```
c(mu0[2] - (t*sqrt(diag(S)[2]/n) ), mu0[2] + (t*sqrt(diag(S)[2]/n) ))
```

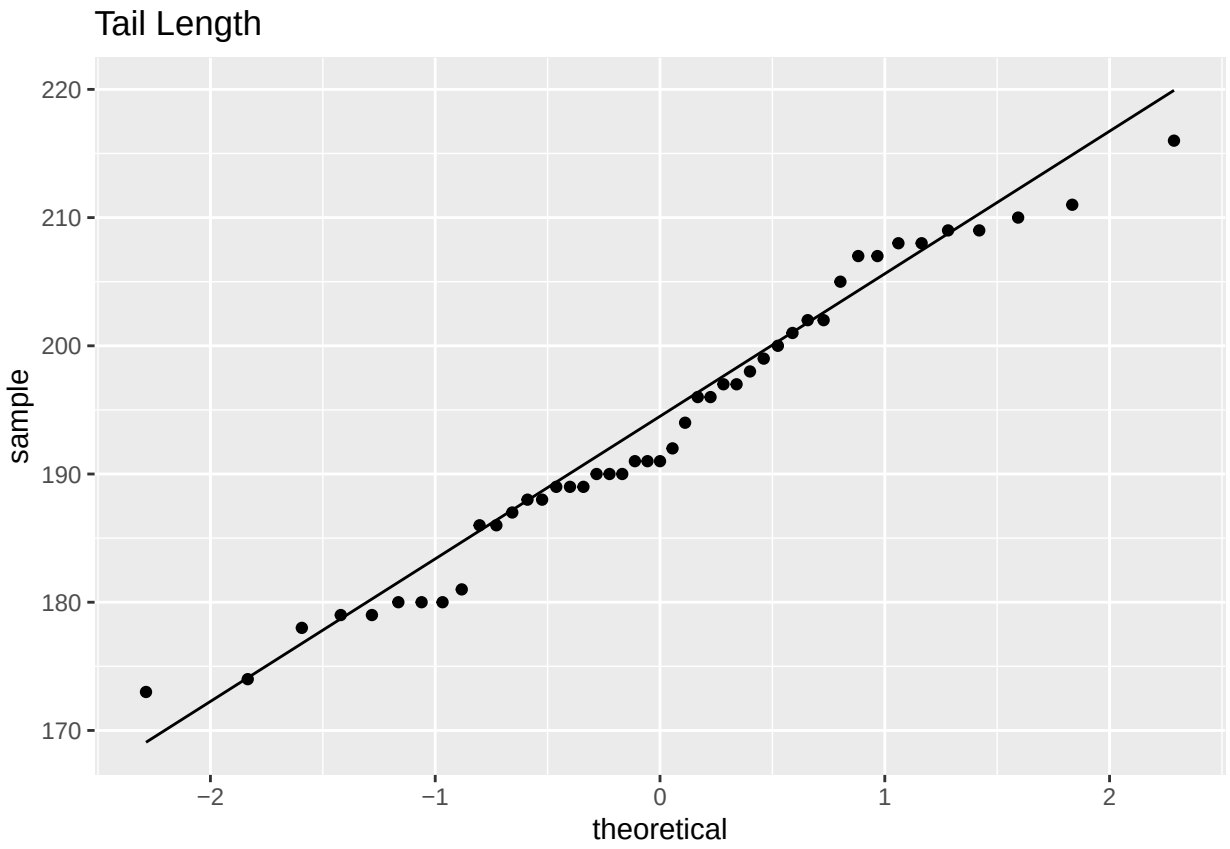
```
## winglength winglength
##   270.0041   279.9959
```

(c) Is the bivariate normal distribution a viable population model? Explain with reference to Q-Q plots and a scatter diagram.

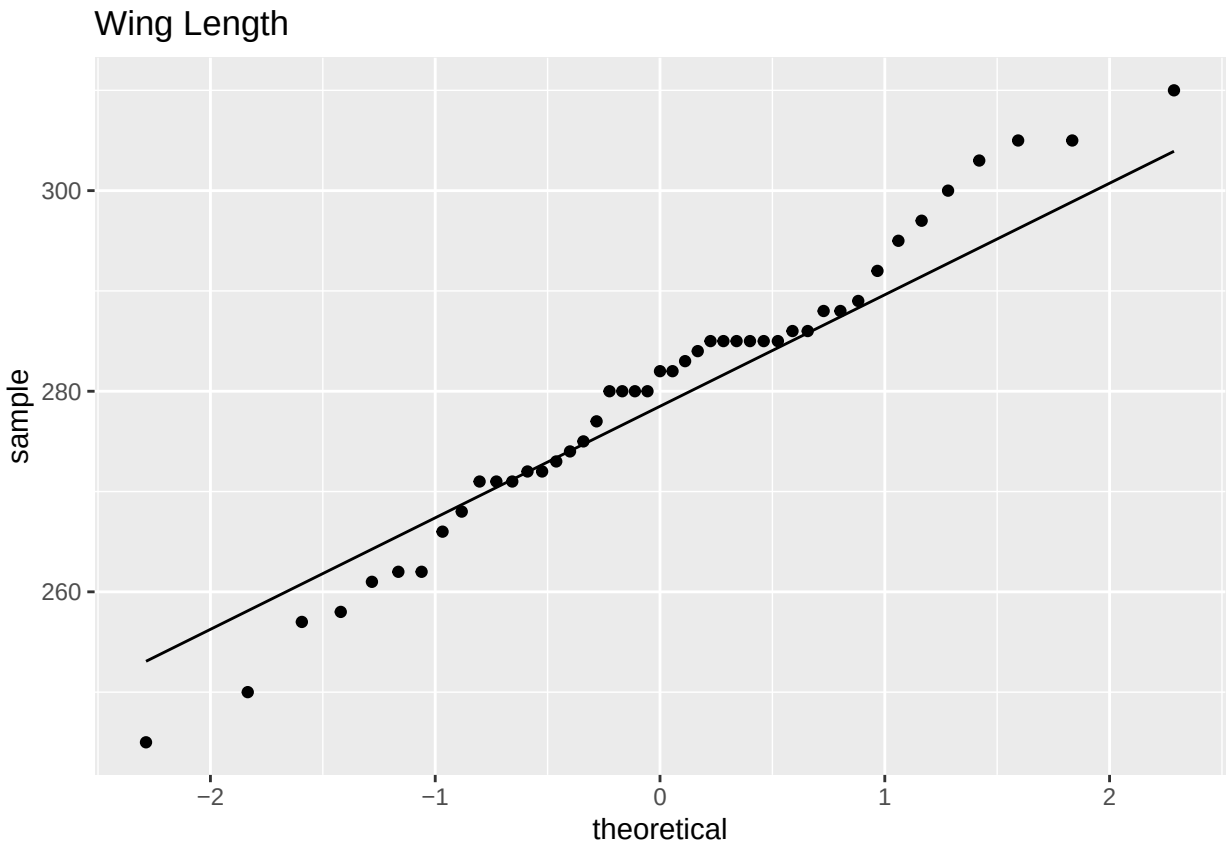
```
ggplot(data = birds)+aes(x=taillength, y=winglength)+
  geom_point(stat = "identity")
```




```
ggplot(data = birds)+aes(sample=taillength)+stat_qq()+stat_qq_line()+
  labs(title = "Tail Length")
```




```
ggplot(data = birds)+aes(sample=winglength)+stat_qq()+stat_qq_line()+
  labs(title = "Wing Length")
```



From the Q-Q plots, we can say that the data is normally distributed.

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.