Lab 2 732A97 Multivariate Statistical Methods

Raymond Sseguya

2019-11-28

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, 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.

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!