Multivariate Statistical Methods - Lab 02

Maximilian Pfundstein (maxpf364), Hector Plata (hecpl268), Aashana Nijhawan(aasni448), Lakshidaa Saigiridharan (laksa656)

2019-11-14

Contents

Focusing on the multivariate normal distribution, we will study methods for estimating, testing hypotheses about and comparing mean vectors. These methods are the multivariate generalizations of the univariate methods.

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.

1.1 Chi-Squared Approximation

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 percent 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 percent a sensible significance level for this task?

Answer: First we import, name and look at the track times.

```
track_times = read.table("data/T1-9.dat")
colnames(track_times) = c("country", "100m", "200m", "400m",
                  "800m", "1500m", "3000m", "marathon")
head(track_times)
```

```
##   country  100m  200m  400m 800m 1500m 3000m marathon
## 1     ARG 11.57 22.94 52.50 2.05  4.25  9.19   150.32
## 2     AUS 11.12 22.23 48.63 1.98  4.02  8.63   143.51
## 3     AUT 11.15 22.70 50.62 1.94  4.05  8.78   154.35
```

```
## 4     BEL 11.14 22.48 51.45 1.97  4.08  8.82   143.05
## 5     BER 11.46 23.05 53.30 2.07  4.29  9.81   174.18
## 6     BRA 11.17 22.60 50.62 1.97  4.17  9.04   147.41
```

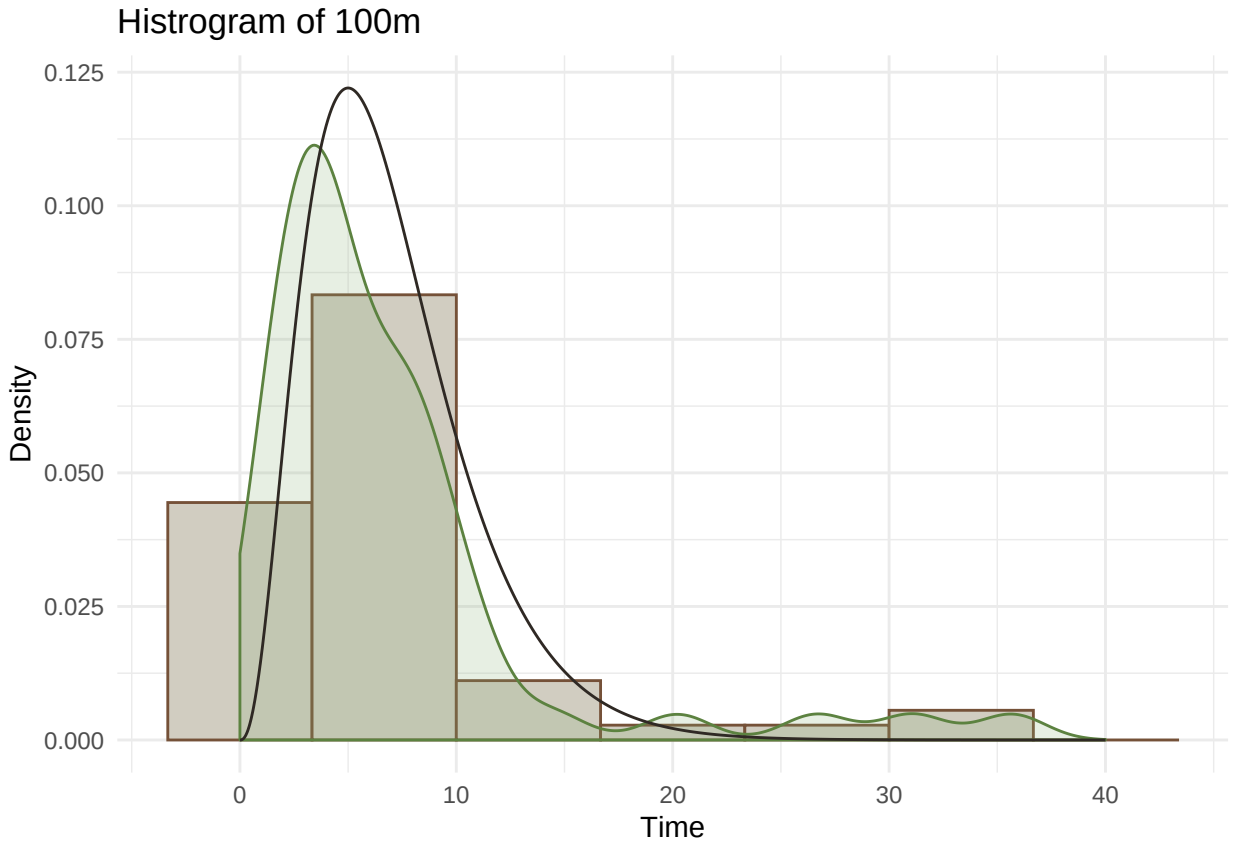
We reimport our function written in the previous lab for computing the Mahalanobis Distance as it is actually more convenient than the built-in function in R.

```
sample_variance = function(X) {

  X = as.matrix(X)

  identity = diag(nrow(X))
  one_n = matrix(1, nrow=nrow(X), ncol=1)

  inter =  identity - 1/nrow(X) * (one_n %*% t(one_n))

  return(1/nrow(X) * (t(X) %*% inter %*% X))
}

mahalanobis_distance = function(X) {
  X = as.matrix(X)

  V = sample_variance(X)
  ident = matrix(1, nrow=nrow(X), ncol=nrow(X))
  mu = 1/nrow(X) * (t(ident) %*% X)

  X_centered = X - mu

  return(diag(X_centered %*% solve(V) %*% t(X_centered)))
}
```

We set the degrees of freedom for the $\chi^2(\nu)$-distribution, which corresponds to the amount of features. We also calculate the Mahalanobis Distances. They should follow a $\chi^2(\nu)$-distribution with 7 degrees of freedom.

```
nu = ncol(track_times) - 1
D = mahalanobis_distance(track_times[,2:8])
```

The following plot shows the histogram of the Mahalanobis Distances with the respective density. The real $\chi^2(\nu)$-distribution with 7 degrees of freedom is outlined in black.



We define $\alpha = 0.001$ and we check for each observation if it lies within the $1 - \alpha$ percentile of the $\chi^2(\nu)$-distribution with 7 degrees of freedom. Finally we check which countries are the outliers. Our findings match the results from the previous lab.

TODO: multiple–testing correction, explain the significance level

```
alpha = 0.001

outlier_indeces = 1 - pchisq(D, nu) < alpha

track_times$country[outlier_indeces]
```

```
## [1] KORN PNG  SAM
## 54 Levels: ARG AUS AUT BEL BER BRA CAN CHI CHN COK COL CRC CZE DEN ... USA
```

1.2 Chi–Square Approximation for Outliers

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 percent 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 percent a sensible significance level for this task?

1.3 Different Outlier Reasoning

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

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!

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.

3.1 Exploring the Data

Explore the data first and present plots that you find informative.

3.2 Differing of Mean Vectors

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

3.3 Confidence Intervals

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.

4 Source Code

```
library(viridis)
library(ggplot2)
knitr::opts_chunk$set(echo = TRUE)

track_times = read.table("data/T1-9.dat")
colnames(track_times) = c("country", "100m", "200m", "400m",
                    "800m", "1500m", "3000m", "marathon")
head(track_times)


sample_variance = function(X) {

  X = as.matrix(X)

  identity = diag(nrow(X))
  one_n = matrix(1, nrow=nrow(X), ncol=1)
```

```
  inter =  identity - 1/nrow(X) * (one_n %*% t(one_n))

  return(1/nrow(X) * (t(X) %*% inter %*% X))
}

mahalanobis_distance = function(X) {
  X = as.matrix(X)

  V = sample_variance(X)
  ident = matrix(1, nrow=nrow(X), ncol=nrow(X))
  mu = 1/nrow(X) * (t(ident) %*% X)

  X_centered = X - mu

  return(diag(X_centered %*% solve(V) %*% t(X_centered)))
}


nu = ncol(track_times) - 1


nu = ncol(track_times) - 1
D = mahalanobis_distance(track_times[,2:8])


val = seq(0, 40, 0.01)
chi_sq_7 = dchisq(val, nu)

ggplot() +
  geom_histogram(aes(x = D, y=..density..),
                 color = "#755138", fill = "#D1CDC1",
                 bins = sqrt(nrow(track_times))) +
  geom_density(aes(x = D, y=..density..),
               color="#5C8240", fill="#8AB077", alpha = 0.2) +
  geom_line(aes(x = val, y = chi_sq_7), color = "#2F2924") +
  labs(title = "Histrogram of 100m",
       y = "Density",
       x = "Time", color = "Legend") +
  scale_color_viridis(discrete=FALSE) +
  theme_minimal()


alpha = 0.001

outlier_indeces = 1 - pchisq(D, nu) < alpha

track_times$country[outlier_indeces]
```