Question 1: Test of outliers

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?**

To answer this question properly we need to calculate again, as we did on the previous lab assignment, the Mahalanobis vector, which can be easily done by running the following commands:

```
> mean <- colMeans(dataset[, 2:8])
> Sx <- cov(dataset[, 2:8])
> D2M <- mahalanobis(dataset[, 2:8], mean, Sx)
>
> setNames(D2M, dataset[, 1])
       ARG        AUS        AUT        BEL        BER        BRA        CAN        CHI        CHN        COL
 6.6500027  3.9356705  4.3680848  2.4623238  7.6305669  2.1885513  3.0916940  4.1542602  3.0167240  5.3994790
       COK        CRC        CZE        DEN        DOM        FIN        FRA        GER        GBR        GRE
19.8340006  4.4647286 10.9014563  1.3174764  4.7284159  2.7556601  5.0862790  3.4467449  3.5572682  9.5403223
       GUA        HUN        INA        IND        IRL        ISR        ITA        JPN        KEN        KORS
 3.6846053  1.7086674  4.7128021  4.2129432  5.4848378  2.5421901  1.4553927  3.2695339  7.6254662  8.0349184
      KORN        LUX        MAS        MRI        MEX        MYA        NED        NZL        NOR        PNG
26.1671415 11.1088463  8.2371487  6.6649850 14.2309322  2.0429608  5.9522081  1.3505689  6.8880631 30.5072477
       PHI        POL        POR        ROM        RUS        SAM        SIN        ESP        SWE        SUI
 9.0658836  1.3311122  2.3695737  6.3491309  3.7506462 35.0140631  8.0163948  1.4160474  0.7846063  3.2910927
       TPE        THA        TUR        USA
10.1839962  7.8152191  8.0453684  9.1556967
```

Once this have been done, we can proceed to run the chi-square test on this Mahalanobis data, which returns this data:

```
> chisq.test(D2M)

        Chi-squared test for given probabilities

data:  D2M
X-squared = 364.07, df = 53, p-value < 2.2e-16
```

After that, and thanks to the *outliers* package (https://cran.r-project.org/web/packages/outliers/index.html) the outliers detection become considerably easier, proceeding as follows:

```
> library(outliers)
> setNames(scores(D2M, type="chisq", prob = 0.999), dataset[, 1])
  ARG   AUS   AUT   BEL   BER   BRA   CAN   CHI   CHN   COL   COK   CRC   CZE   DEN   DOM   FIN   FRA   GER   GBR
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  GRE   GUA   HUN   INA   IND   IRL   ISR   ITA   JPN   KEN  KORS  KORN   LUX   MAS   MRI   MEX   MYA   NED   NZL
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  NOR   PNG   PHI   POL   POR   ROM   RUS   SAM   SIN   ESP   SWE   SUI   TPE   THA   TUR   USA
FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

This output on the console shows, as a result of the *setNames* function, the country's acronym over a TRUE/FALSE indicator, which highlights the observations whose Mahalanobis distance lies beyond the indicated percentile. In this first case, the function works with the given significance of α=0.001.

Analyzing the table above, it can be noticed how only two countries are being identified as outliers. Those are: Papua New Guinea (*PNG*, 30.5072477) and Samoa (*SAM*, 35.0140631), which are the two most extreme cases amongst all the calculated distances.

The study could be finished here but seems effective to go a bit deeper and try to modify this quantile through some α modifications in order to find a better differentiation between outliers and average observations. More specifically, the Bonferroni correction enables the calibration of this parameter. Supposing that, for our dataset, each country would be compared once with each other country on the list, the number of total combinations comes to 1431. As a consequence, the chosen correction method results in a new α value of 1 - 0.761 = 0.239. Running again the *scores* function with this new parameter returns the following:

```
> setNames(scores(D2M, type="chisq", prob = 0.761), dataset[, 1])
  ARG   AUS   AUT   BEL   BER   BRA   CAN   CHI   CHN   COL   COK   CRC   CZE   DEN   DOM   FIN   FRA   GER   GBR
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  GRE   GUA   HUN   INA   IND   IRL   ISR   ITA   JPN   KEN  KORS  KORN   LUX   MAS   MRI   MEX   MYA   NED   NZL
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  NOR   PNG   PHI   POL   POR   ROM   RUS   SAM   SIN   ESP   SWE   SUI   TPE   THA   TUR   USA
FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

With this new limit, two new countries are identified as outliers: Cook Islands (*COK*, 19.8340006) and North Korea (*KORN*, 26.1671415). Those two values are far from being standard, therefore can be concluded that the correction is accurate. Beyond this conclusion, alpha could be even more reduced, depending on where we want to place the "outlier threshold". As a last example, setting α to 0.35 makes also Mexico (*MEX*, 14.2309322) an outlier:

```
> setNames(scores(D2M, type="chisq", prob = 0.65), dataset[, 1])
  ARG   AUS   AUT   BEL   BER   BRA   CAN   CHI   CHN   COL   COK   CRC   CZE   DEN   DOM   FIN   FRA   GER   GBR
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  GRE   GUA   HUN   INA   IND   IRL   ISR   ITA   JPN   KEN  KORS  KORN   LUX   MAS   MRI   MEX   MYA   NED   NZL
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
  NOR   PNG   PHI   POL   POR   ROM   RUS   SAM   SIN   ESP   SWE   SUI   TPE   THA   TUR   USA
FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

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

The reason for this is that Euclidean distance simply computes the ordinary straight line between two points and the Mahalanobis distance takes the covariate into account, which leads to elliptic decision boundaries in 2D, therefore narrower than the Eucliden circular, which helps to properly identify the outliers, as done in the previous section.