Lab 2 - Multivariate Statistical Methods

Gustav Sternelöv

Tuesday, November 24, 2015

Assignment 1

a)

In a test where each country is tested with a significance level of 0.1 %, the following rule is applied to determine whether a country is an outlier or not:

If the Mahalanbois distance is higher than 24.32, the country is considered to be an outlier.

The table below shows the six countries which has the highest calculated distances. Out of these six countres the first three of them are considered to be outliers. These are Samoa, Papua New Guinea and North Korea.

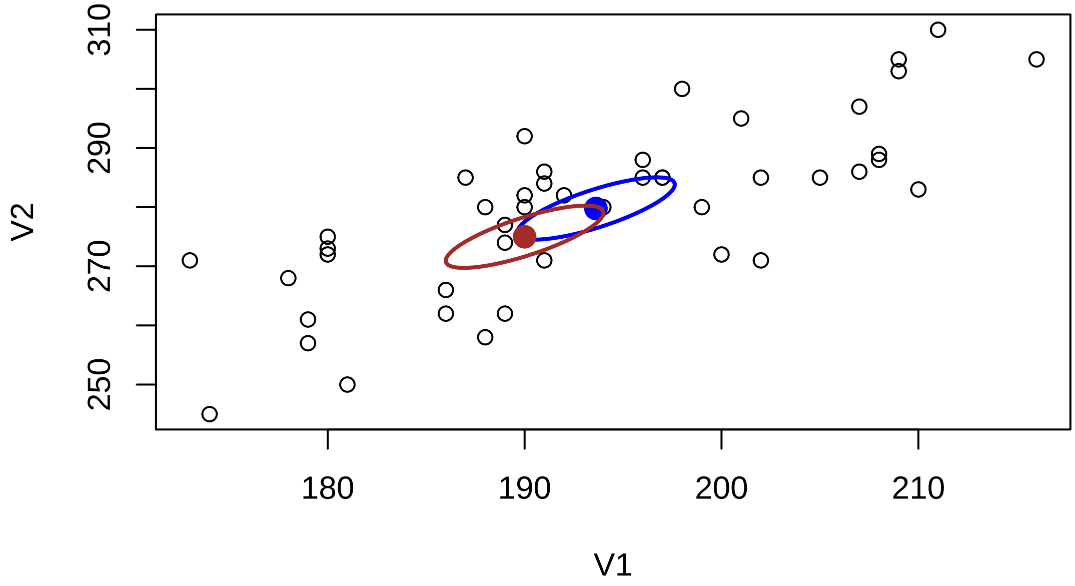
```
##    data...1. diag.MahanabisD.
## 46       SAM         35.01406
## 40       PNG         30.50725
## 31    KOR, N         26.16714
## 11       COK         19.83400
## 35       MEX         14.23093
## 32       LUX         11.10885
```

b)

The different results given by the Mahalanobis and the euclidean distance is due to the use of the covariance in the former distance measure. In this case the data set uses different units for different variables and this is not taken into account unless the covariance matrix is used. Since the Mahalanobis distances uses the covariances and the euclidean distances not uses them, we obtain different results depending on which distances that are calculated.

Assignment 2

a)



In the graph above the blue ellipse is the 95 % confidence ellipse for the population means for the female birds. The mean values for the male birds are illustrated by the brown dot and these values are thought to be plausible if they lie within the blue ellipse. Since this is the case, the dot lies just inside the ellipse, the values are considered to be plausible.

b)

95 % confidence interval for mean tail and wing length, Hotellings T^2:
Tail length: 189.4246525, 197.8197919
Wing length: 274.2601999, 285.2953557

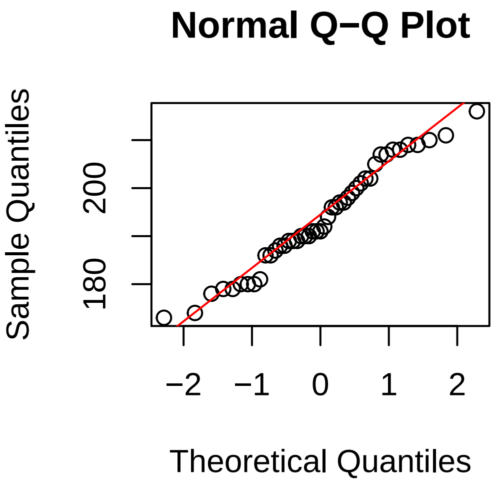
95 % confidence interval for mean tail and wing length, Bonferroni:
Tail length: 189.8227237, 197.4217207
Wing length: 274.7834524, 284.7721032

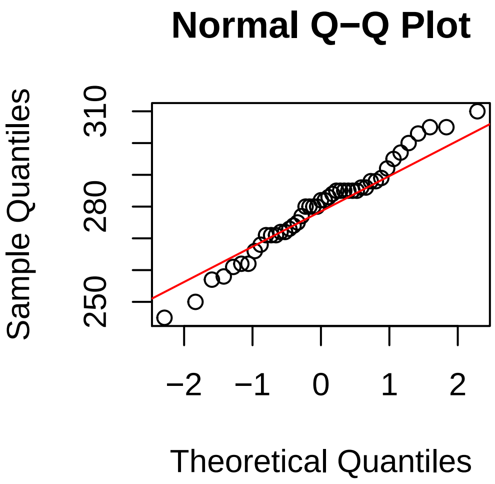
As expected the Bonferroni intervals are more narrow than the T^2-intervals, although the difference is very small.

An advantage with the T^2-intervals is that they works better when the number of specified mean components is large.

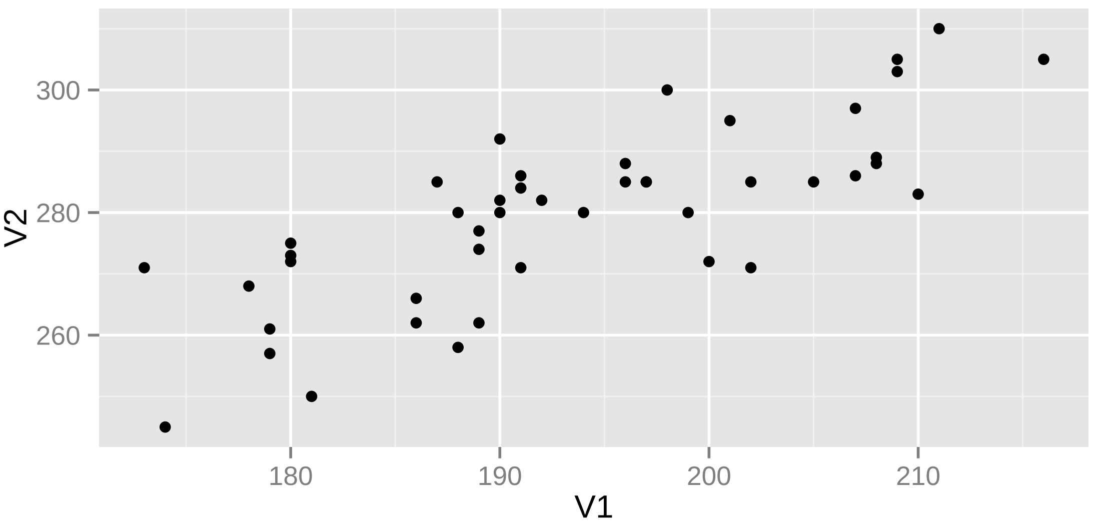
c)

To examine whether a bivariate normal distribution is a viable population model, Q-Q plots and a scatter diagram are produced.





In the Q-Q plot above to the left the values lies well alongside the red line. For the Q-Q plot to the right the values does not follow the red line equally well, but it still looks quite good. This would suggest that the bivariate normal distribution is a viable population model.



Regarding the scatter diagram above the observed points could be thought of as creating an ellipse. This also speaks for the viability of the bivariate normal distribution as a population model.

Assignment 3

The formula used for calculating simultaneous confidence intervals are presented below. The part to the left of the ± sign gives the difference and the part to the right gives the size of the interval.
$\bar{x}_{ki}$ - $\bar{x}_{li}$ $\pm t_{n-g}(\frac{\alpha}{pg(g-1)})\sqrt{\frac{\omega_{ii}}{n-g}\frac{1}{n_k}+\frac{1}{n_l}}$
Where n is equal to 90, g is equal to 3 and p is equal to 4. All time periods have the same amount of observations, so both n_k and n_l is always equal to 30. ω_{ii} is the diagonal values of the W matrix and this corresponds to the following vector of values (1785.4, 1924.3, 2153.0, 840.2), where the first value is equal to

ω_{11}, the second ω_{22} and so on.
$t_{n-g}(\frac{\alpha}{pg(g-1)})$ is a constant and can be computed to be equal to: 2.943

Four different values can be computed for the second part of the expression, $\sqrt{\frac{\omega_{ii}}{n-g}\frac{1}{n_k}+\frac{1}{n_l}}$, one for each response variable.
For the respective response variable then the following calculation is done to determine the interval size:

$$2.943 \times \sqrt{\frac{1785.4}{87}\frac{1}{30}+\frac{1}{30}} = 3.44$$
$$2.943 \times \sqrt{\frac{1924.3}{87}\frac{1}{30}+\frac{1}{30}} = 3.57$$
$$2.943 \times \sqrt{\frac{2153.0}{87}\frac{1}{30}+\frac{1}{30}} = 4.03$$
$$2.943 \times \sqrt{\frac{840.2}{87}\frac{1}{30}+\frac{1}{30}} = 2.36$$

The next step is to calcualte the differences between the mean values for the different periods. This is done for all of the for four response variables. This results in twelve differences, three for each response variable. Each difference is compared to corresponding interval size and if the interval not covers zero the mean values for the respective time periods are concluded to be different.

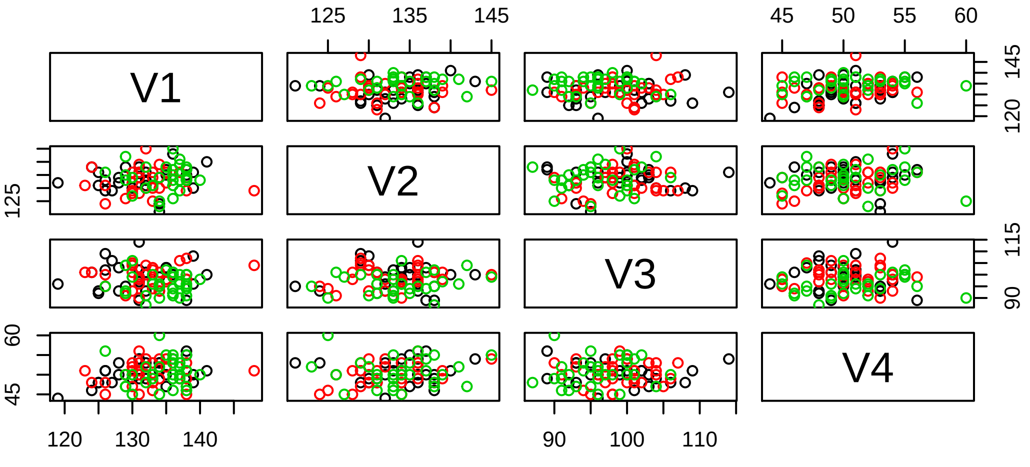
$(\tau_{11}-\tau_{21},\tau_{11}-\tau_{31}\ ,\tau_{21}-\tau_{31})$ =(-1, -3.1, -2.1) ± 3.44
$(\tau_{12}-\tau_{22},\tau_{12}-\tau_{32}\ ,\tau_{22}-\tau_{32})$ =(0.9, -0.2, -1.1) ± 3.57
$(\tau_{13}-\tau_{23},\tau_{13}-\tau_{33}\ ,\tau_{23}-\tau_{33})$ =(0.1, 3.1, 3) ± 4.03
$(\tau_{14}-\tau_{24},\tau_{14}-\tau_{34}\ ,\tau_{24}-\tau_{34})$ =(0.3, 0, -0.3) ± 2.36

Since it easily can be seen that all of the intervals covers zero the exact intervals are not calculated. The interpretation of this result is that none of the mean components seem to differ among the populations.

One of the assumptions made when conducting a MANOVA is that the covariance matrices are equal for the different populations. Another is that the random samples from different populations are independent and a third that each population is multivariate normal.

Regarding the independence of the samples it is hard to both reject and confirm this assumption, but it feels reasonable to consider them to be independent. The covariance matrices are compared and thought to be relatively similar, this could be investigated further with a Box's test but this is not done here.

One way to investigate if each population is multivariate normal is by looking at a scatter matrix. In the scatter matrix shown below the observations are coloured after the population they comes from.


125
135
145
45
50
55
60
V1
145
120
125
V2
115
V3
90
60
45
V4
120
130
140
90
100
110