Project 2: Multiple Analysis of Variance

DA 410

Marjorie Blanco

Part 1:

Download `testscoredata.txt` and read it in R or SAS.

Use Hotelling's T^2 test to test for a difference in the mean score vector of the boys and the mean vector of the girls. Make sure you include clear command lines and relevant output/results with hypotheses, test result(s) and conclusion(s)/interpretation(s).

	boy	girl
math	84.09833	82.9625
reading	81.78667	82.2800

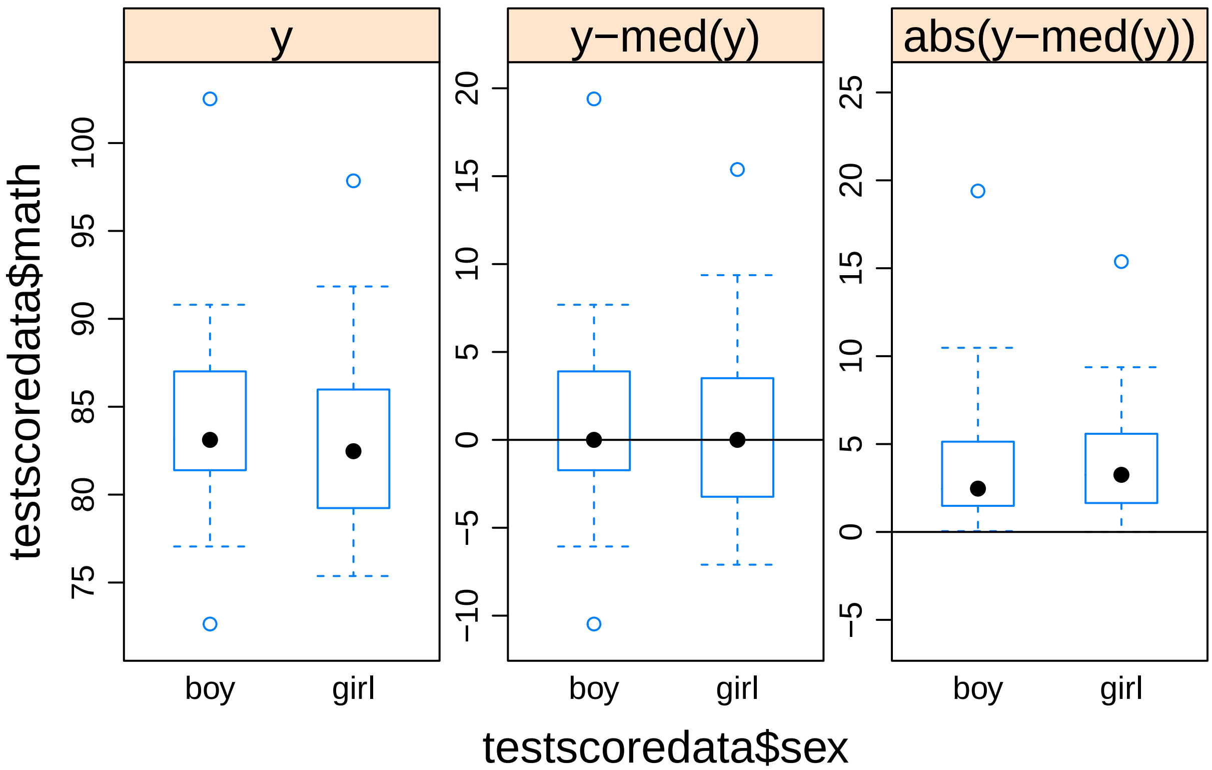
```
## $statistic
## [1] 18.35569
##
## $m
## [1] 0.4916667
##
## $df
## [1]  2 59
##
## $nx
## [1] 30
##
## $ny
## [1] 32
##
## $p
## [1] 2
```

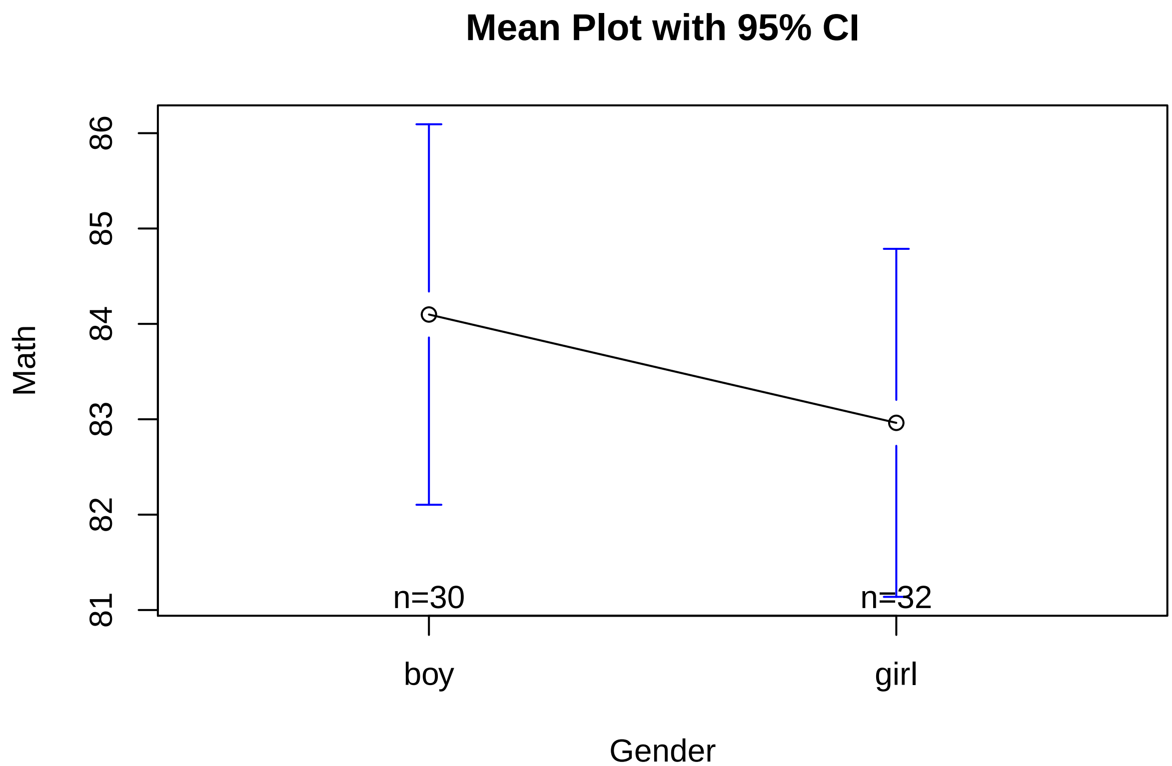
```
##
##  Hotelling's two sample T2-test
##
## data:  boys and girls
## T.2 = 9.0249, df1 = 2, df2 = 59, p-value = 0.0003805
## alternative hypothesis: true location difference is not equal to c(0,0)
```

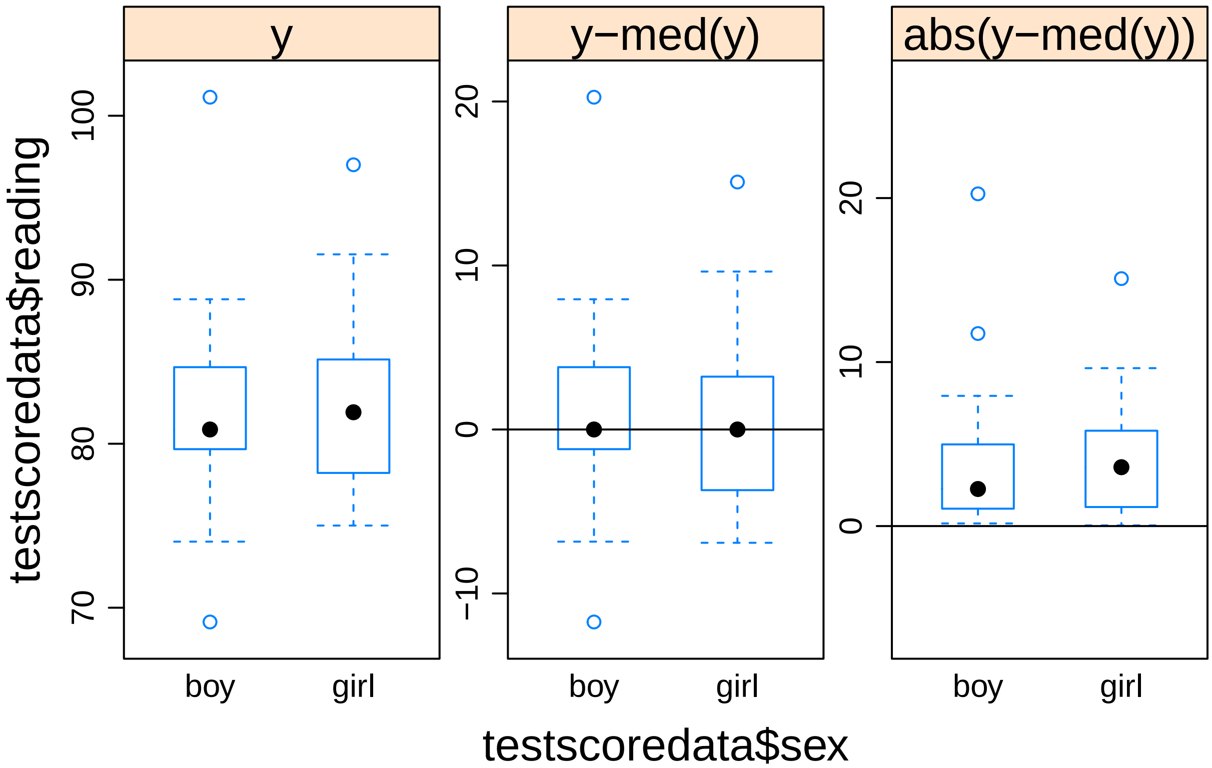
```
##
##  Hotelling's two sample T2-test
##
## data:  boys and girls
## T.2 = 18.356, df = 2, p-value = 0.0001033
## alternative hypothesis: true location difference is not equal to c(0,0)
```

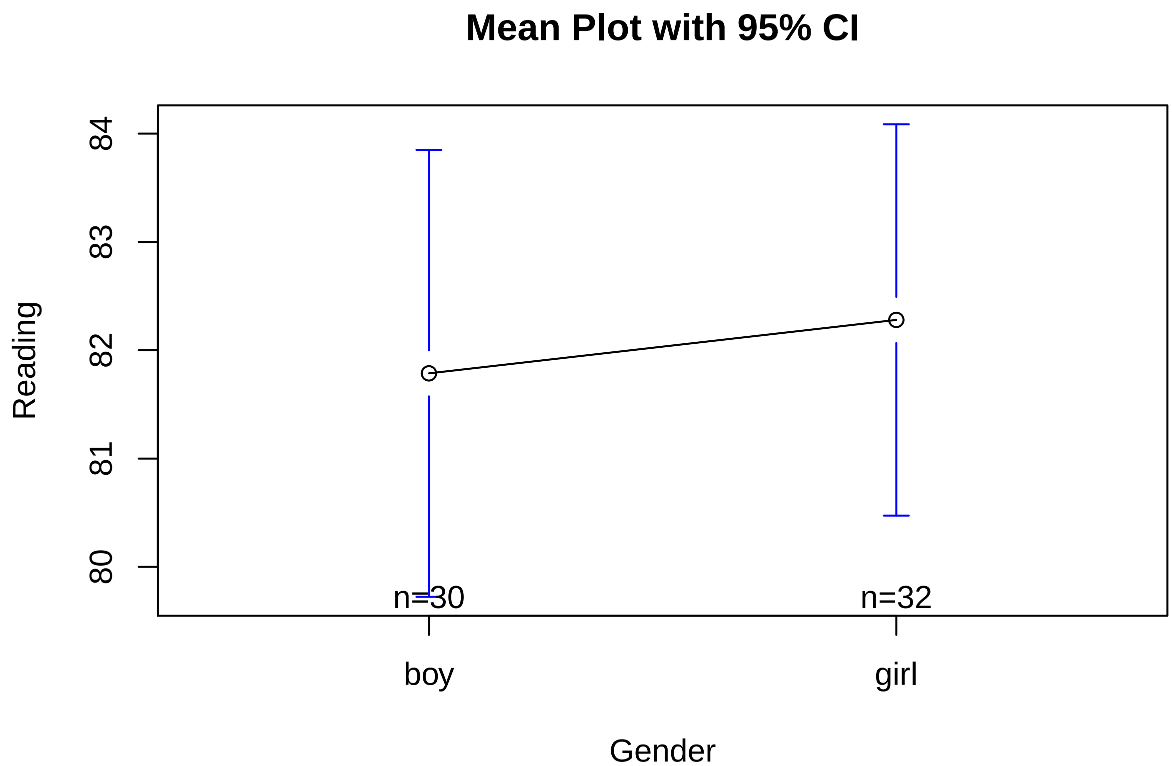
```
##                      Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
## testscoredata$sex  1          0.30593   9.0249      2     59 0.0003805 ***
## Residuals         60
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing the value of the test statistic T^2 with critical value, we are led to reject the hypothesis of equal mean vectors for the gender groups.




Mean Plot with 95% CI
Math
86
85
84
83
82
81
n=30
n=32
boy
girl
Gender


y
y−med(y)
abs(y−med(y))
testscoredata$reading
100
90
80
70
20
10
0
−10
boy
girl
testscoredata$sex



Part 2:

Suppose we have gathered the following data on female athletes in three sports. The measurements we have made are the athletes' heights and vertical jumps, both in inches. The data are listed as (height, jump) as follows:

Basketball Players:(66, 27), (65, 29), (68, 26), (64, 29), (67, 29)

Track Athletes:(63, 23), (61, 26), (62, 23), (60, 26)

Softball Players:(62, 23), (65, 21), (63, 21), (62, 23), (63.5, 22), (66, 21.5)

The following R code should read in the data as 3 vectors:

The data (athletes' heights and vertical jumps in inches) for female athletes from each of three sports are given in Table.

id	sport	height	jump
1	B	66.0	27.0
2	B	65.0	29.0
3	B	68.0	26.0
4	B	64.0	29.0
5	B	67.0	29.0
6	T	63.0	23.0
7	T	61.0	26.0
8	T	62.0	23.0
9	T	60.0	26.0
10	S	62.0	23.0
11	S	65.0	21.0
12	S	63.0	21.0
13	S	62.0	23.0
14	S	63.5	22.0
15	S	66.0	21.5

a) Use R to conduct the MANOVA F-test using Wilks' Lambda to test for a difference in (height, jump) mean vectors across the three sports. Make sure you include clear command lines and relevant output/results with hypotheses, test result(s) and conclusion(s)/interpretation(s)

We would then like to test if the properties are the same across the three sports.

$H_0 : \mu_1 = \mu_2$

H_1 : The two $\mu's$ are unequal

```
##           Df    Wilks approx F num Df den Df    Pr(>F)
## sport      2 0.035879   23.536      4     22 1.117e-07 ***
## Residuals 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Wilks's test rejects the hypothesis H_0 that the mean vector for the three sports are equal.

```
##           Df    Roy approx F num Df den Df    Pr(>F)
## sport      2 16.833      101      2     12 3.109e-08 ***
## Residuals 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Roy's test also rejects the hypothesis H_0 that the mean vector for the three sports are equal.

```
##           Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
## sport      2           17.396    43.49      4     20 1.355e-09 ***
## Residuals 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Hotelling-Lawley's test also rejects the hypothesis H_0 that the mean vector for the three sports are equal.

```
##           Df Pillai approx F num Df den Df   Pr(>F)
## sport      2 1.3041   11.244      4     24 2.78e-05 ***
## Residuals 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Pillai's test statistic test also rejects the hypothesis H_0 that the mean vector for the three sports are equal.

```
## Call:
## cbind(height, jump) ~ sport
##
## Descriptive:
##   sport n height   jump
## 1     B 5   66.0 28.000
## 2     S 6   63.5 21.833
## 3     T 4   61.5 24.500
##
## Wald-Type Statistic (WTS):
##       Test statistic df p-value
## sport        243.052  4       0
##
## modified ANOVA-Type Statistic (MATS):
##       Test statistic
## sport         90.795
##
## p-values resampling:
##       paramBS (WTS) paramBS (MATS)
## sport             0              0
```

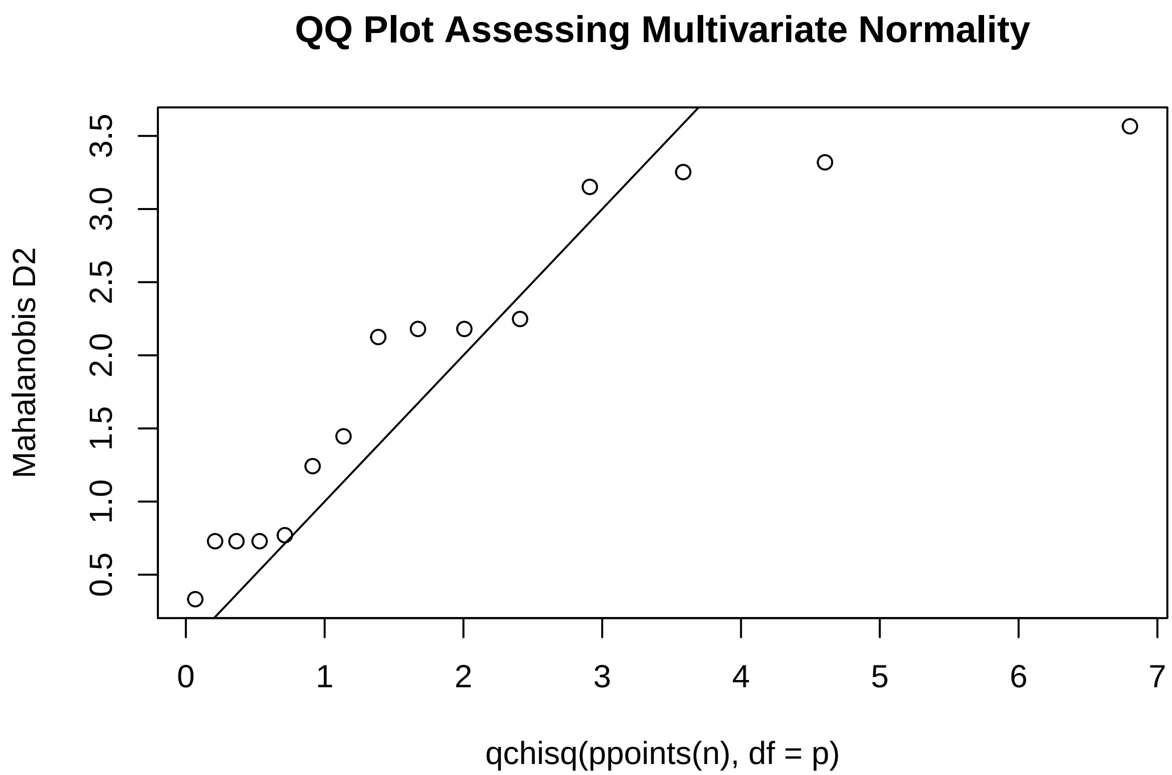
```
##  Response height :
##             Df Sum Sq Mean Sq F value  Pr(>F)
## sport        2 45.625 22.8125  9.7046 0.00311 **
## Residuals   12 28.208  2.3507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response jump :
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## sport        2 101.025  50.512  28.581 2.728e-05 ***
## Residuals   12  21.208   1.767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since each of the four statistical tests indicates that the mean vector of the two variates (height, jump) across the three sports are significantly different from each other, the measurements across the sports has to be different.

The hypothesis $H_0 : \mu_1 = \mu_2$ was rejected for the athletes data.

b) State the assumptions of your test and check to see whether assumptions are met. Do you believe your inference is valid? Why or why not?

Test Multivariate Normality



Significant departures from the line suggest violations of normality.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  athletes$height by athletes$sport
## Bartlett's K-squared = 0.1935, df = 2, p-value = 0.9078
```

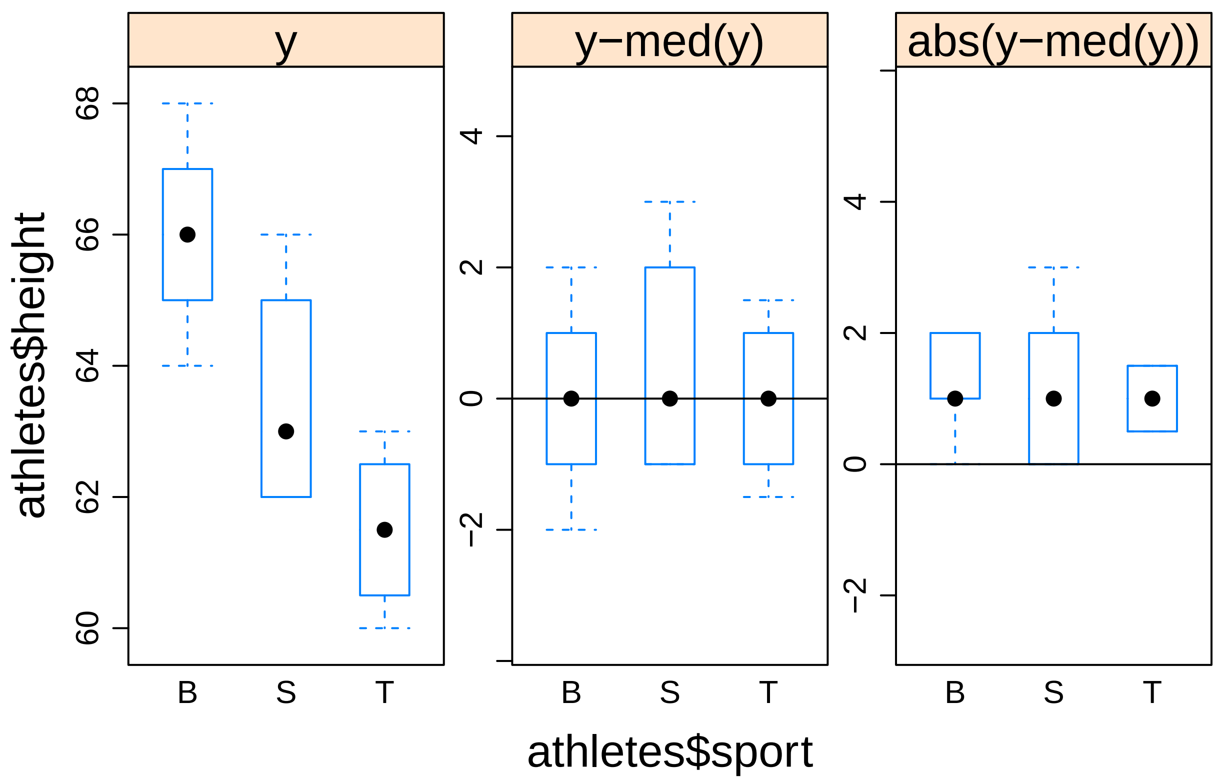
```
##
##  Bartlett test of homogeneity of variances
##
## data:  athletes$jump by athletes$sport
## Bartlett's K-squared = 1.1494, df = 2, p-value = 0.5629
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  athletes$height by athletes$sport
## Fligner-Killeen:med chi-squared = 0.23953, df = 2, p-value =
## 0.8871
```

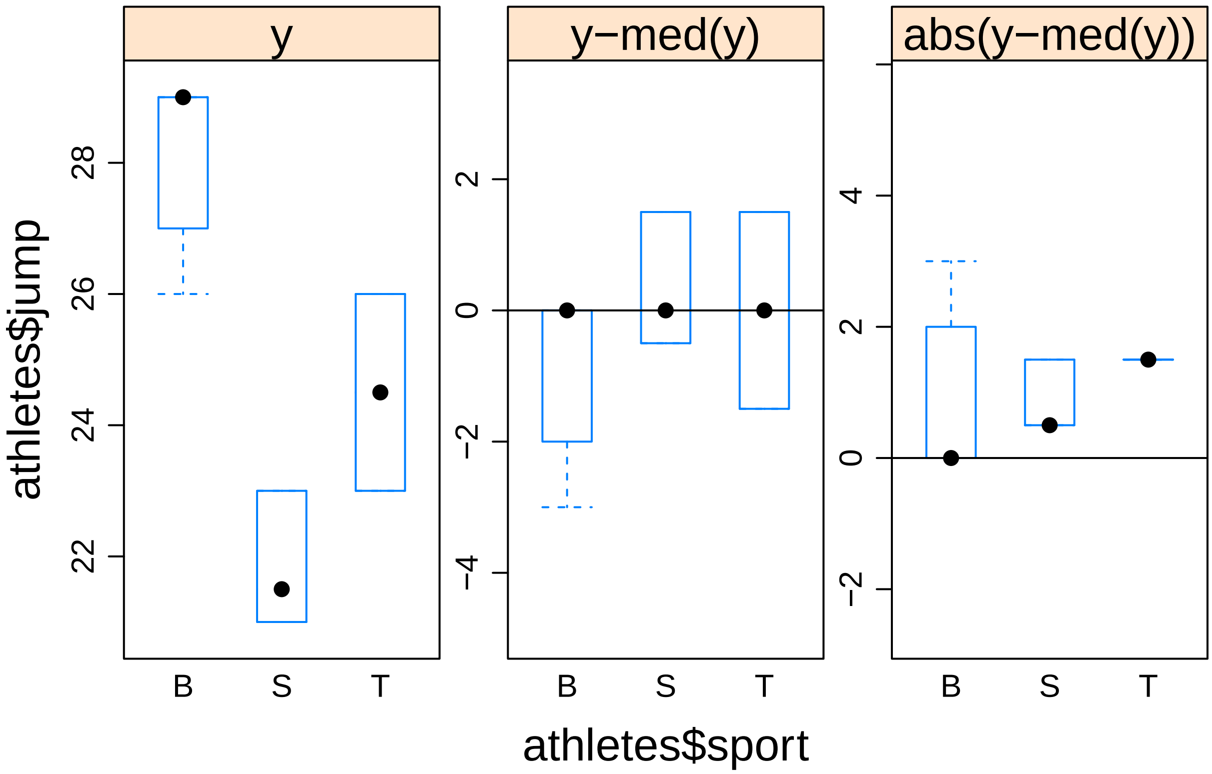
```
##
##  Fligner-Killeen test of homogeneity of variances
##
```

```
## data:  athletes$jump by athletes$sport
## Fligner-Killeen:med chi-squared = 1.0808, df = 2, p-value = 0.5825
```

```
##
##  hov: Brown-Forsyth
##
## data:  athletes$height
## F = 0.056426, df:athletes$sport = 2, df:Residuals = 12, p-value =
## 0.9454
## alternative hypothesis: variances are not identical
```




```
##
##  hov: Brown-Forsyth
##
## data:  athletes$jump
## F = 0.70714, df:athletes$sport = 2, df:Residuals = 12, p-value =
## 0.5125
## alternative hypothesis: variances are not identical
```

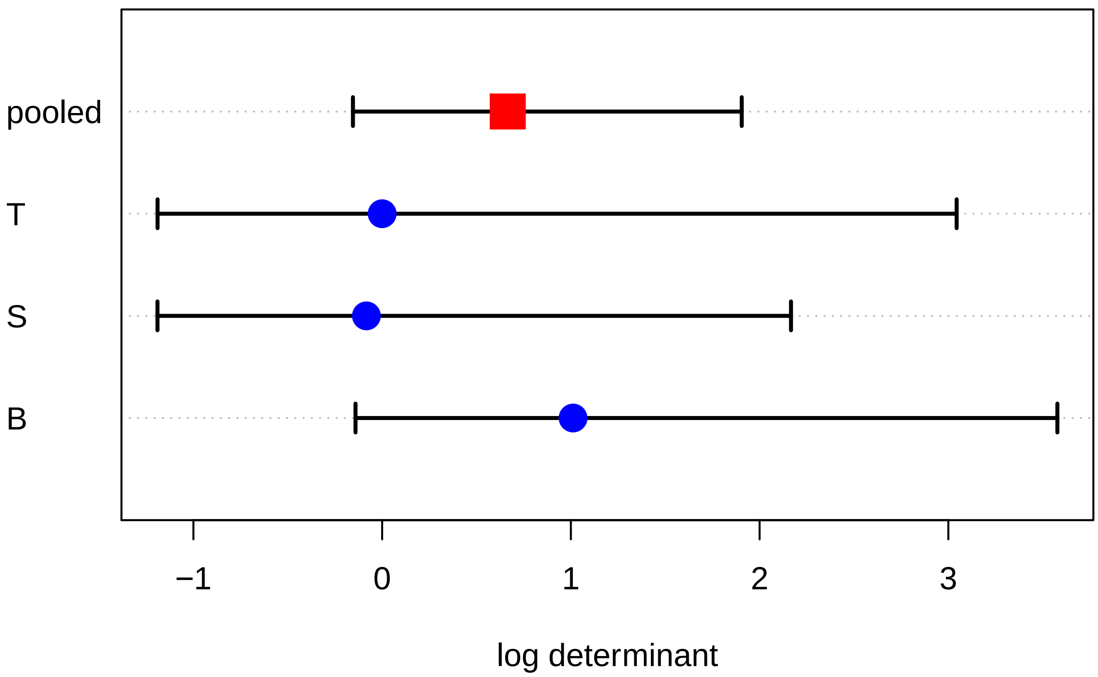


Equality of covariance matrices using Box's M-test

H_0: The observed covariance matrices for the dependent variables are equal across groups

H_1: The observed covariance matrices for the dependent variables are not equal across groups

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  cbind(athletes$height, athletes$jump)
## Chi-Sq (approx.) = 3.2506, df = 6, p-value = 0.7768
```



	B	S	T
height	2.5	2.7000000	1.666667
jump	2.0	0.9666667	3.000000

The non-significant test result, p-value $0.7768026 > 0.05$, the null hypothesis of equal variance-covariance matrices between groups fail to be rejected.

c) Use R to examine the sample mean vectors for each group. Make sure you include clear command lines and relevant output/results. Also comment on the differences among the groups in terms of the specific variables.

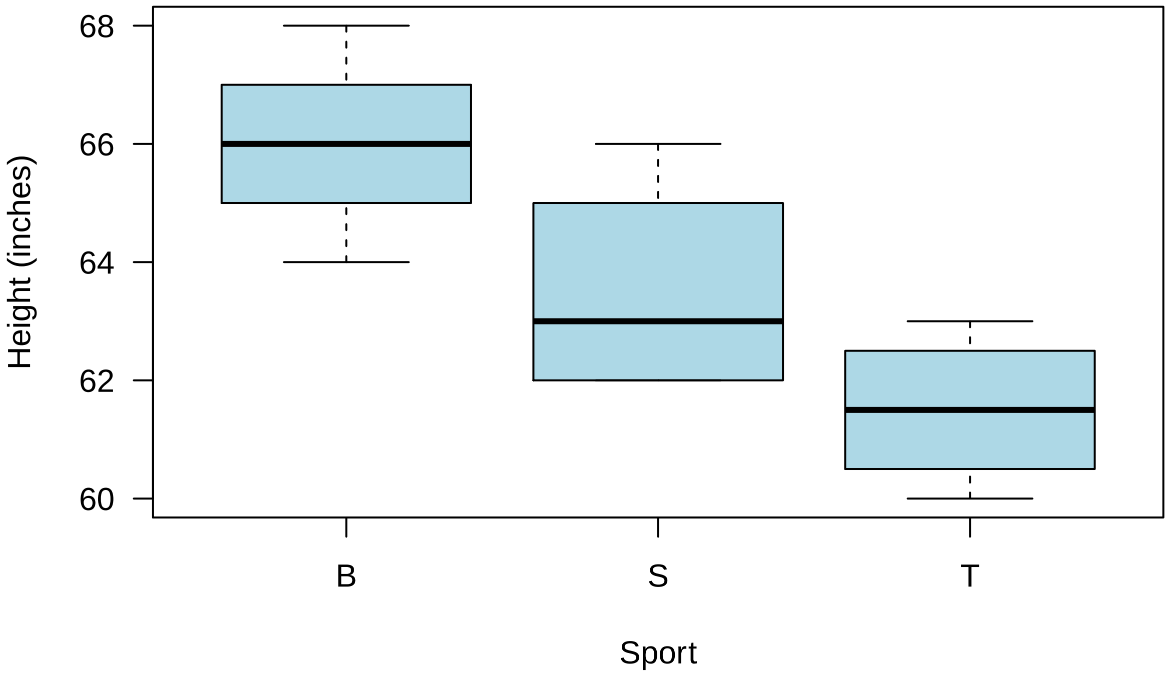
	B	S	T
height	66	63.50000	61.5
jump	28	21.83333	24.5

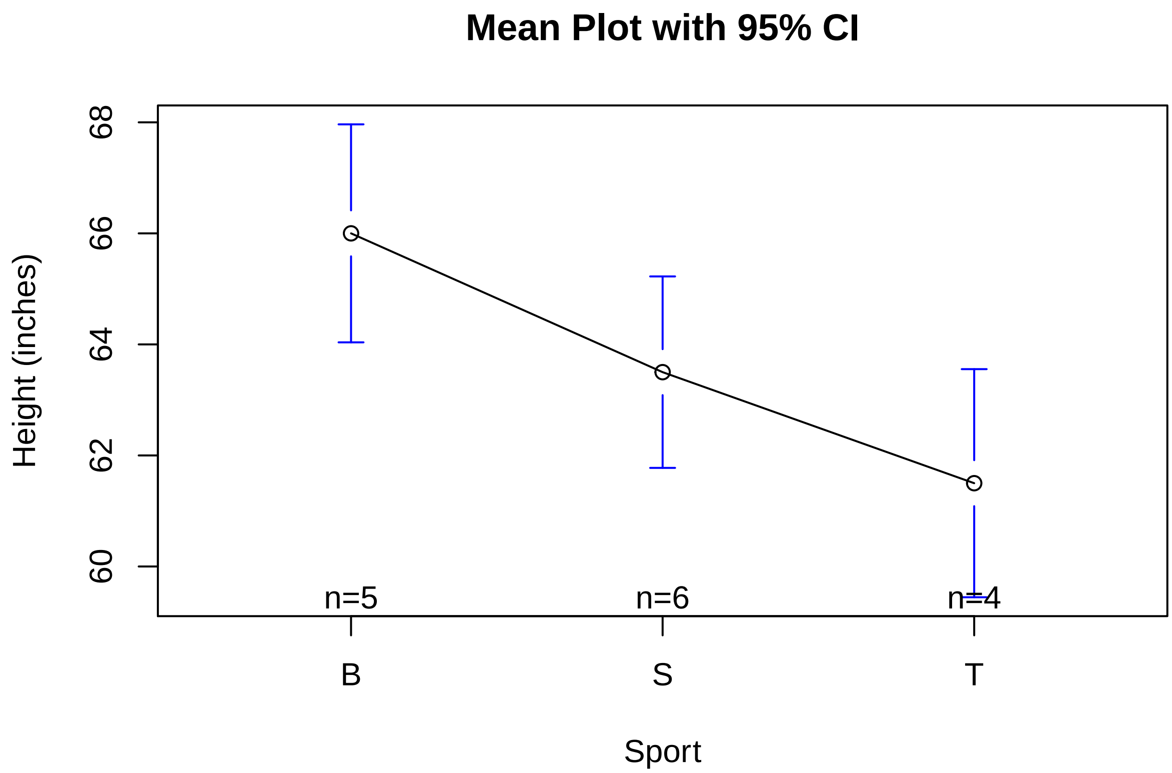
The two individual variables will be tested using the 0.05 level of significance.

First variable, y_1 = athletes' heights jumps (inches)

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## sport        2   45.9  22.950   9.663 0.00316 **
## Residuals   12   28.5   2.375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For F = 9.6631579 the p-value is 0.0031596, and we reject H_0

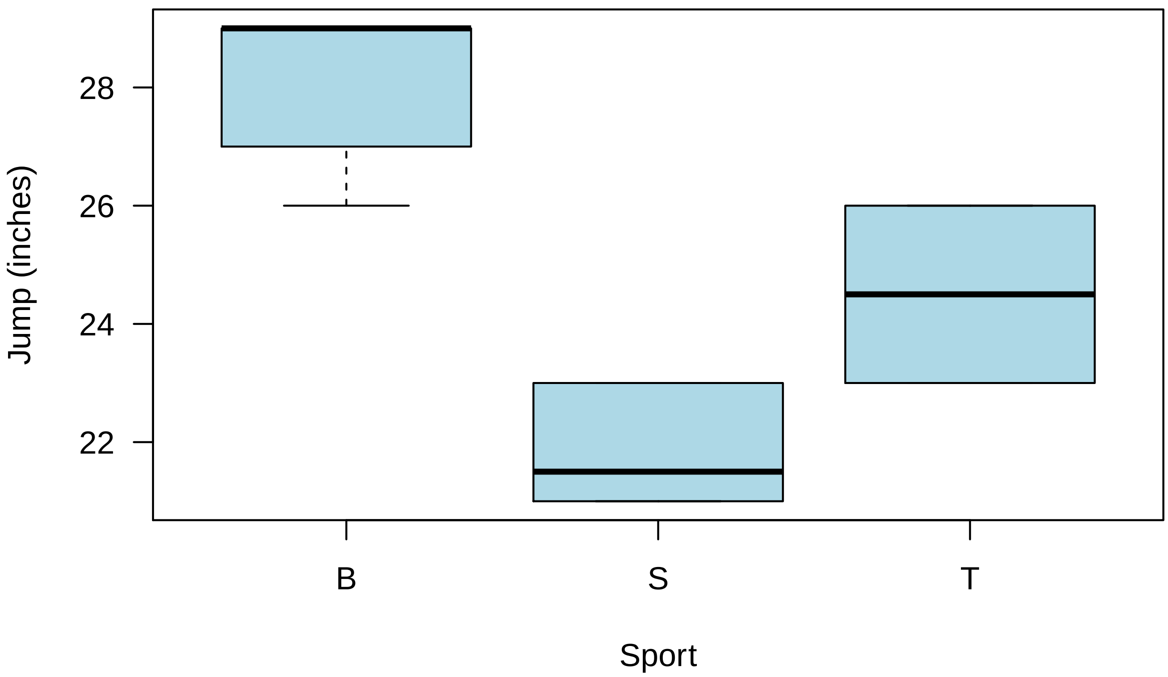


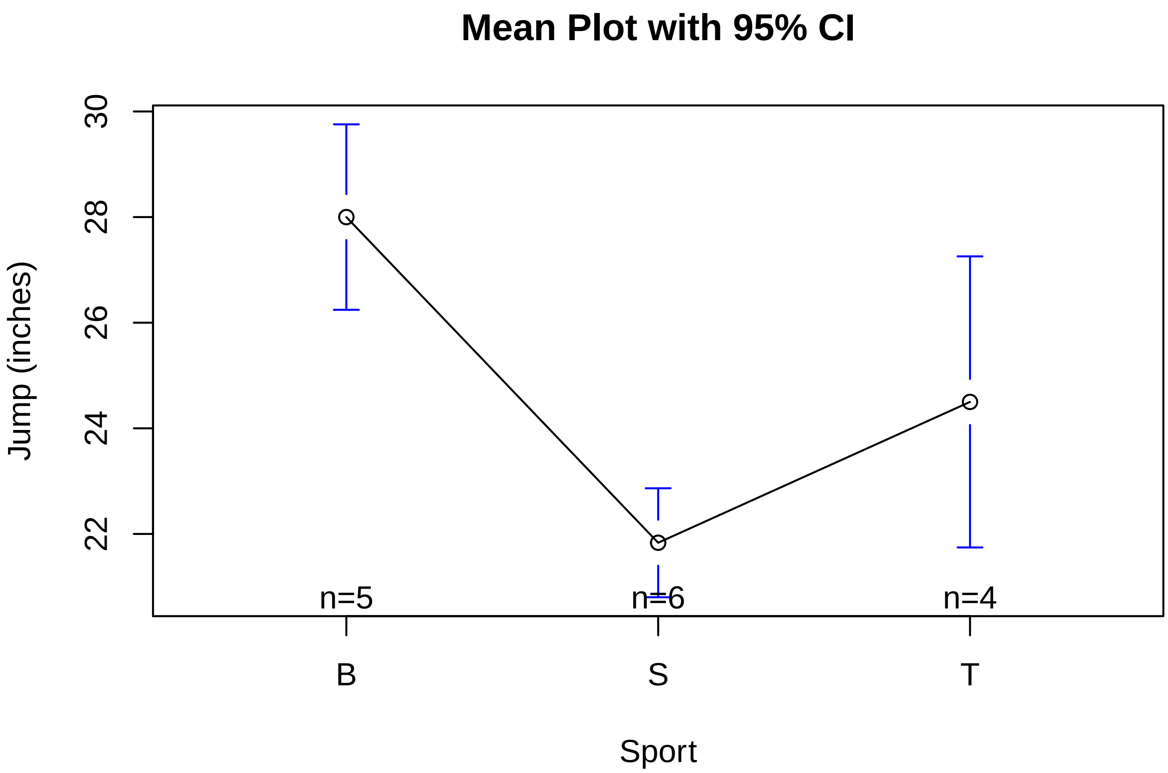


Second variable, y_2 = athletes' vertical jumps (inches)

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## sport        2 103.77   51.88   28.52 2.76e-05 ***
## Residuals   12  21.83    1.82                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For F = 28.5160305 the p-value is 0, and we reject H_0


Jump (inches)
28
26
24
22
B
S
T
Sport



From the output above, it can be seen that the two variables are highly significantly different among sport.