Assignment 1

Ruijie Sun

January 25, 2018

Solutions:

1. (30 marks) Consider the box plot below, drawn in R, based on the data in the file "juries.csv".

(a) (10 marks) Recall the 1.5IQR Rule which is used to identify potential outliers. Show, using this rule, how the two points identify as outliers.

Firstly, I calculate lowerbound of groupS and upperbound of groupNS through 1.5IQR Rule.

```
## [1] 6.7
```

```
## [1] 47.55
```

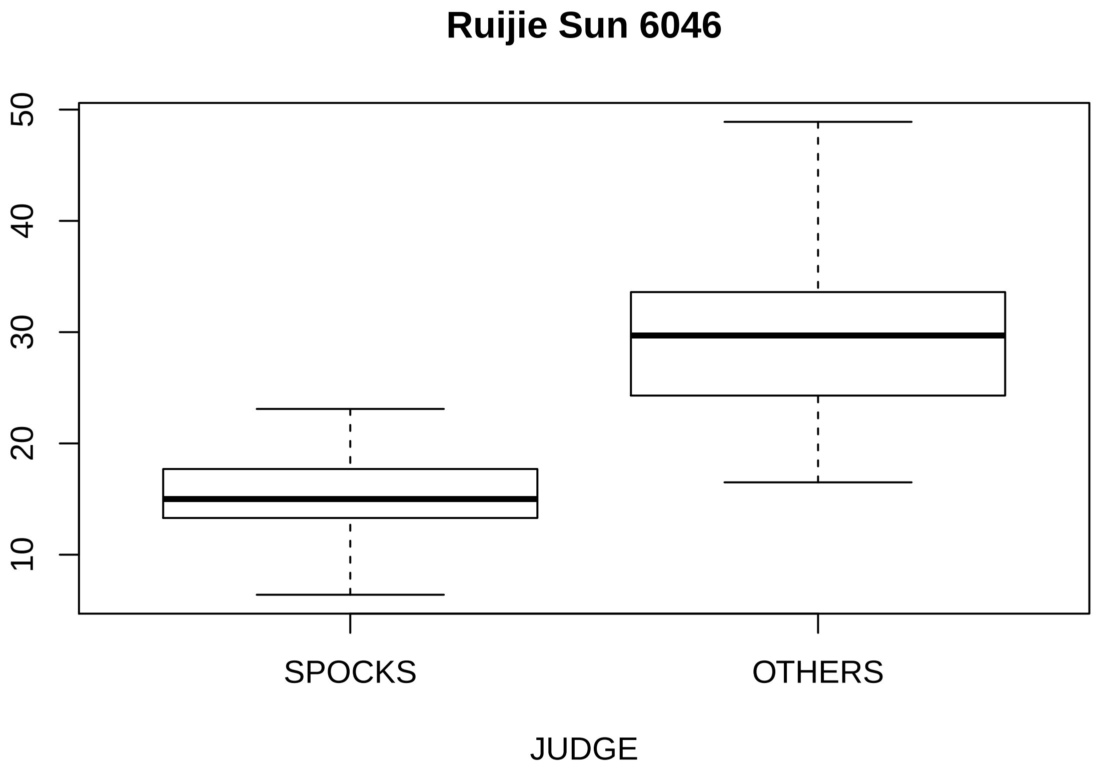
Then, I get the outlier number through the 1.5IQR Rule.

```
## [1] 6.4
```

```
## [1] 48.9
```

Since 6.4 is lower than groupS's lowerbound and 48.9 is bigger than groupNS's upperbound, so they are classified as outliers.

(b) (15 marks) Recreate the side-by-side box plots of percent of women on venires for Spock's judge and the other judges without identifying outliers.



(c) (5 marks) Comment on the difference between the skeletal box plot (which does not identify outliers) and the modified box plot (which identifies outliers).

1.The modified box plot draws outliers but the skeletal box plot does not draw outliers.

2.In modified box plot, the upper bound is maximum within Q3+1.5IQR in dataset and the lowerbound is minimum within Q1-1.5IQR in dataset. But in skeletal box plot, the upperbound is the maximum in the dataset and the lowerbound is the minimum in the data set.

2 Consider the data, "assign1data.csv" based on the heights of 166 students in our class and answer the questions that follow.

(a)(5 marks) Was the data based on an experiment or an observational study? Briefly discuss the limitations on the statistical inference we can draw from this data.

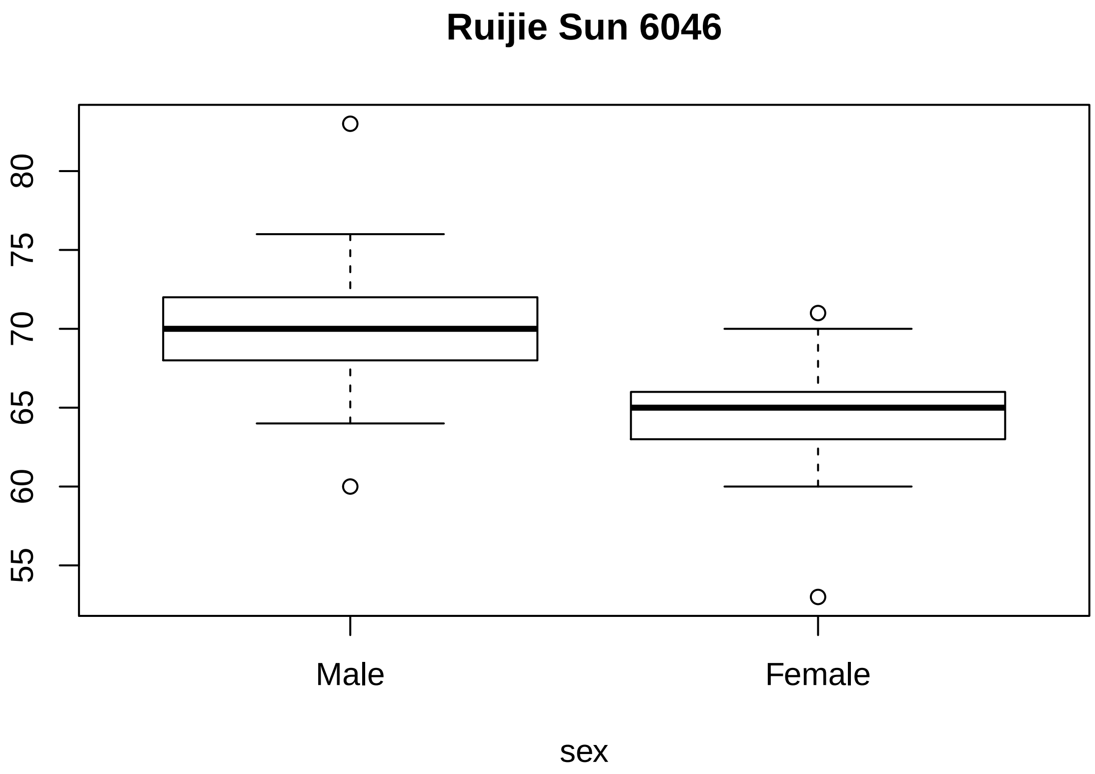
The data based on an observational study. Since we can only get association relation from observational study, so we can not get causality relation between gender and height.

(b) (5 marks) Which variables are categorical? Name the levels of each categorical variable.

Sex is a categorical variable. And the levels of sex are male(0) and female(1).

(c) (20 marks) Conduct an appropriate hypothesis test to determine whether there is a difference between the heights of Males and Females.

i. Side-by-side boxplots



ii. Null and Alternative Hypotheses

Null Hypothesis: the average height of male is equal to the average height of female.

Alternative Hypothesis: the average height of male is not equal to the average height of female.

iii. A test statistic and it's distribution

To test if the two sample have same variance.

```
##
##  F test to compare two variances
##
## data:  groupM and groupFM
## F = 1.2917, num df = 79, denom df = 85, p-value = 0.2471
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8365331 2.0015544
## sample estimates:
## ratio of variances
##           1.291697
```

p-value =0.2471 > 0.05. So the two groups have equal variances.

test statistics =

$$\frac{\overline{x}-\overline{y}}{s_p^2(\frac{1}{n_x}+\frac{1}{n_y})} = 12.076$$

follows t distribution with degree freedom 164, where

$$s_p^2 = \frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{n_x+n_y-2}$$

Result from two sample t-test:

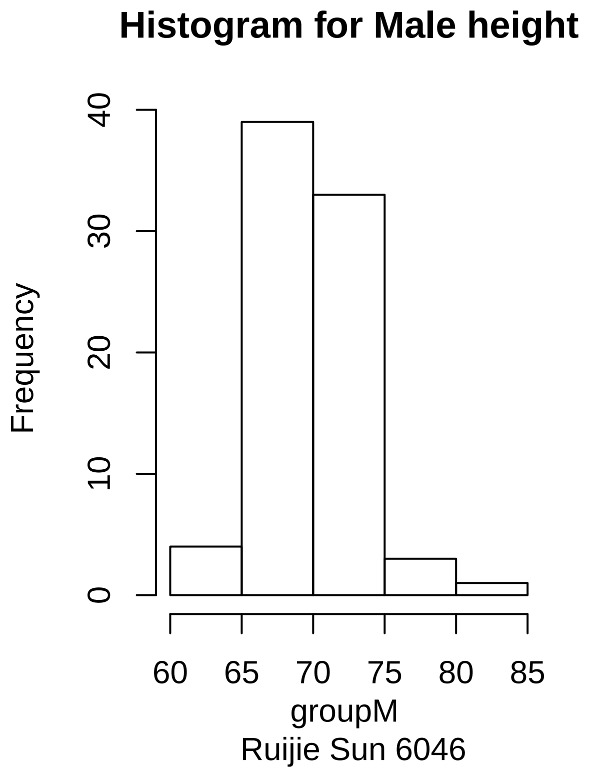
```
##
##  Two Sample t-test
##
## data:  groupM and groupFM
## t = 12.076, df = 164, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.691630 6.525811
## sample estimates:
## mean of x mean of y
##  70.22500  64.61628
```

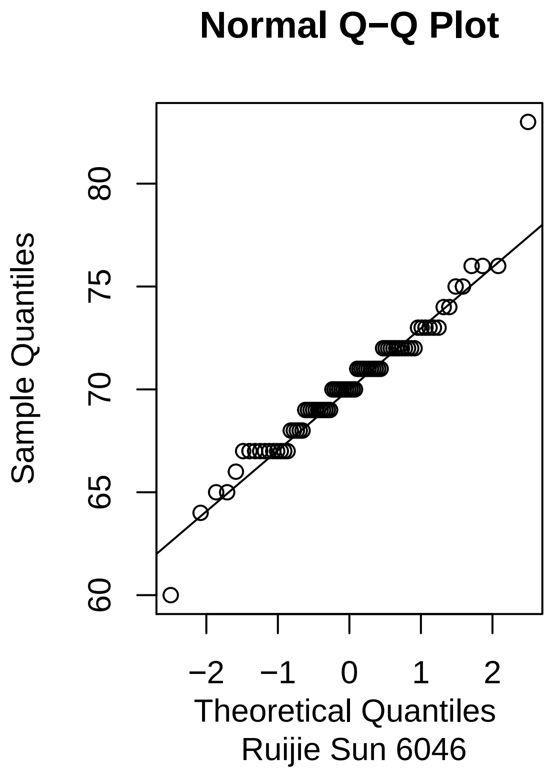
iv. Test assumptions

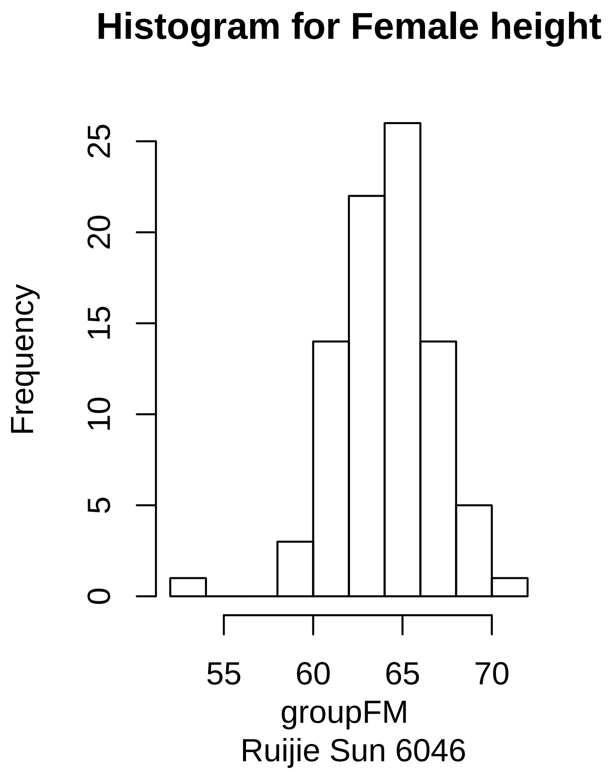
1.The two samples are iid from approximately Normal population.

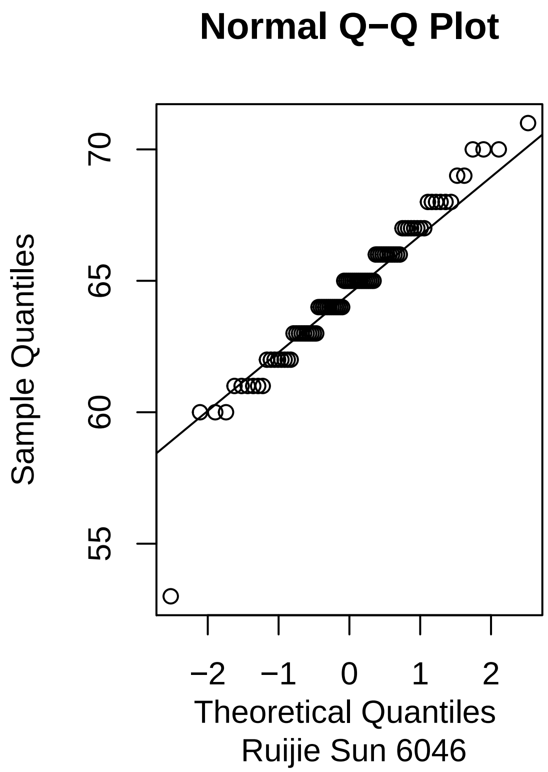
2.The two samples are independent of each other.

v. Test diagnostics (checking model assumptions)


Histogram for Male height
Frequency
0
10
20
30
40
60
65
70
75
80
85
groupM
Ruijie Sun 6046


Normal Q−Q Plot
Sample Quantiles
60
65
70
75
80
−2
−1
0
1
2
Theoretical Quantiles
Ruijie Sun 6046


Histogram for Female height
Frequency
0
5
10
15
20
25
55
60
65
70
groupFM
Ruijie Sun 6046


Normal Q−Q Plot
Sample Quantiles
55
60
65
70
−2
−1
0
1
2
Theoretical Quantiles
Ruijie Sun 6046

Compared to sample of male, sample of female is more approximately Normal. But sample of male is still acceptable for normality.

vi. P-value

```
##
##  Two Sample t-test
##
## data:  groupM and groupFM
## t = 12.076, df = 164, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.691630 6.525811
## sample estimates:
## mean of x mean of y
##  70.22500  64.61628
```

P-value < 2.2e-16

vii. Results (brief discussion and conclusion)

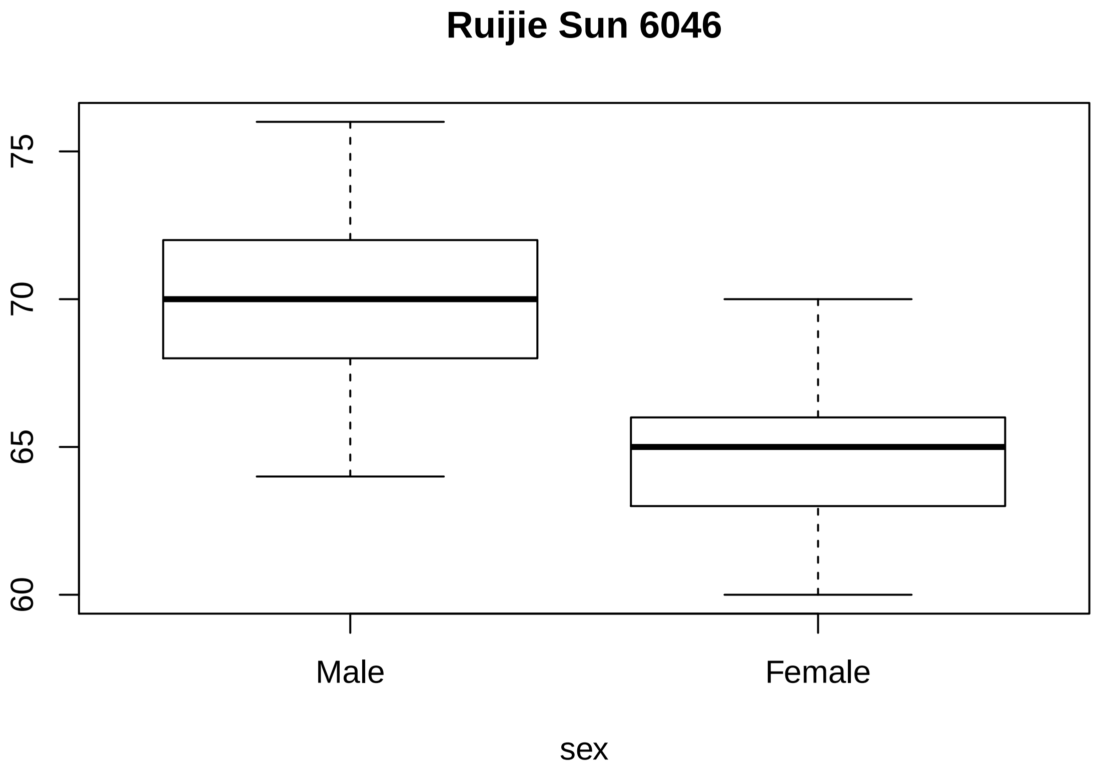
Since p-value < 0.05, there is sufficient evidence to reject Null Hypothesis. So the average height of male is different from average height of female.

(d) (5 marks) Name two(2) statistical methods which are equivalent to your method used in part (c) above.

1. Simple Linear regression(Dummy variable).
2. One-way ANOVA.

(e) (25 marks) Create a subset of the data by removing the row of observations whose 'id' matches the last 2 digits of your student number. For instance, this can be done in R by shivon.subset <- shivon.data[-100,] if my student number ends with '00'. Then redo the analyses of part (c) above with your data subset.

i. Side-by-side boxplots



ii. Null and Alternative Hypotheses

Null Hypothesis: the average height of male is equal to the average height of female. Alternative Hypothesis: the average height of male is not equal to the average height of female.

iii. A test statistic and it's distribution

```
##
##  F test to compare two variances
##
## data:  groupM and groupFM
## F = 1.2802, num df = 79, denom df = 84, p-value = 0.2655
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8278249 1.9857384
## sample estimates:
## ratio of variances
##           1.280222
```

p-value =0.2655 $>$ 0.05. So the two groups have equal variances.

test statistics =

$$\frac{\overline{x}-\overline{y}}{s_p^2(\frac{1}{n_x}+\frac{1}{n_y})} = 12.048$$

follows t distribution with degree freedom 163, where

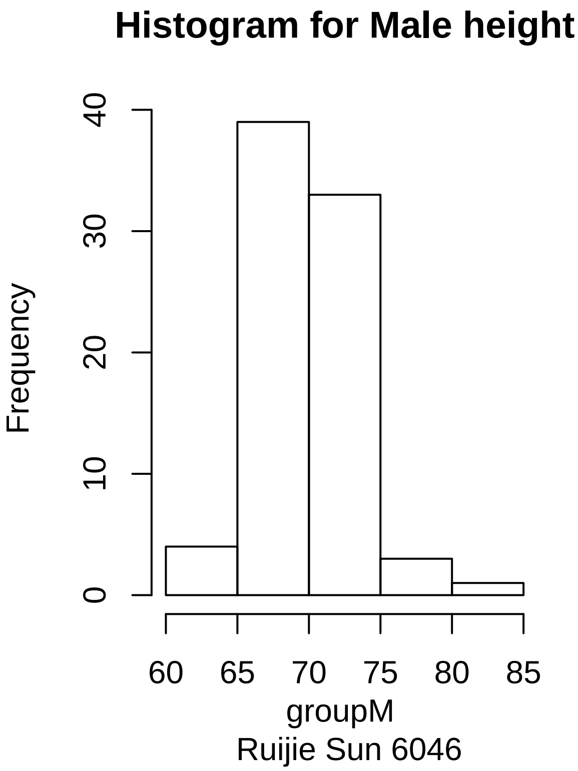
$$s_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$$

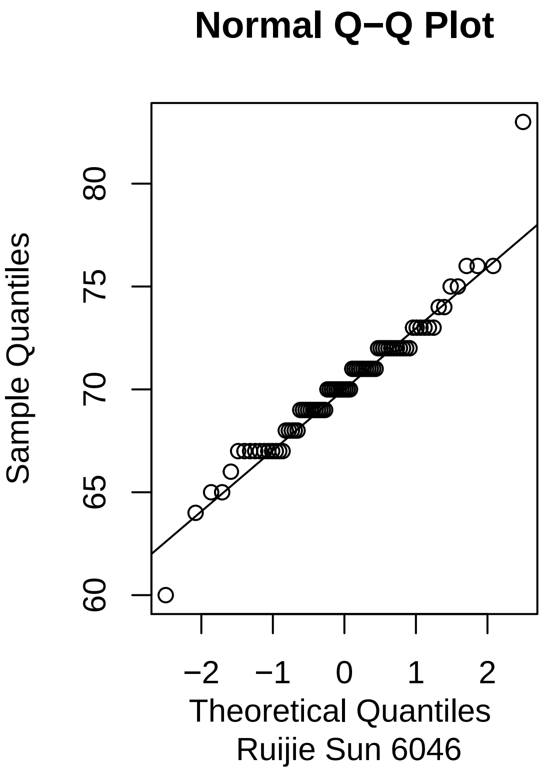
```
##
##  Two Sample t-test
##
## data:  groupM and groupFM
## t = 12.048, df = 163, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.703064 6.546936
## sample estimates:
## mean of x mean of y
##    70.225    64.600
```

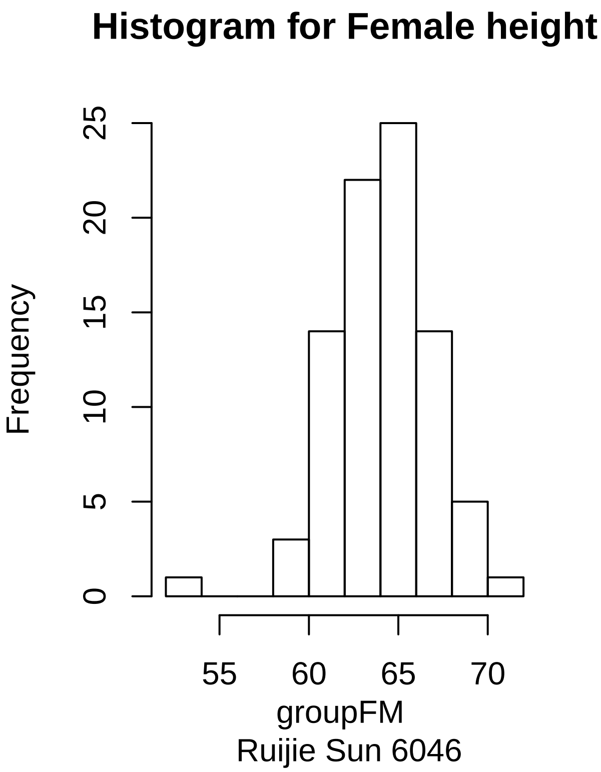
iv. Test assumptions

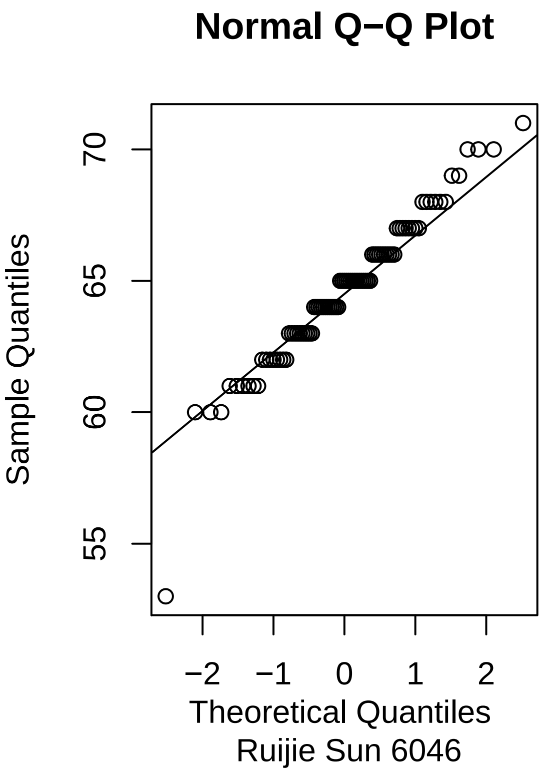
1.The two samples are iid from approximately Normal population. 2.The two samples are independent of each other.

v. Test diagnostics (checking model assumptions)


Histogram for Male height
Frequency
0
10
20
30
40
60
65
70
75
80
85
groupM
Ruijie Sun 6046


Normal Q−Q Plot
Sample Quantiles
60
65
70
75
80
−2
−1
0
1
2
Theoretical Quantiles
Ruijie Sun 6046


Histogram for Female height
Frequency
0
5
10
15
20
25
55
60
65
70
groupFM
Ruijie Sun 6046


Normal Q−Q Plot
Sample Quantiles
55
60
65
70
−2
−1
0
1
2
Theoretical Quantiles
Ruijie Sun 6046

Compared to sample of male, sample of female is more approximately Normal. But sample of male is still acceptable for normality.

vi. P-value

```
##
##  Two Sample t-test
##
## data:  groupM and groupFM
## t = 12.048, df = 163, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.703064 6.546936
## sample estimates:
## mean of x mean of y
##    70.225    64.600
```

P-value = 2.2e-16

vii. Results (brief discussion and conclusion)

Since p-value < 0.05, there is sufficient evidence to reject Null Hypothesis. So average height of male is different from average height of female.

(f) (10 marks) Compare your results of part (c) and part (e). Do you think that the observation removed was influential?

The result of part(c) is same as part(e). It is not influential.

Appendix

1. (30 marks) Consider the box plot below, drawn in R, based on the data in the file "juries.csv".

(a) (10 marks) Recall the 1.5IQR Rule which is used to identify potential outliers. Show, using this rule, how the two points identify as outliers.

```r
# Import the data
dat_juries <- read_csv("juries(1).csv")
```

```
## Parsed with column specification:
## cols(
##   PERCENT = col_double(),
##   JUDGE = col_character()
## )
```

```r
attach(dat_juries)
```

```
## The following objects are masked from dat_juries (pos = 5):
##
##     JUDGE, PERCENT
```

```r
groupS <- PERCENT[JUDGE == "SPOCKS"]
groupNS <- PERCENT[JUDGE != "SPOCKS"]
```

```r
summary(groupS)
summary(groupNS)
```

```r
groupS_IQR <- 17.70 - 13.30
groupNS_IQR <- 33.60 - 24.30

groupS_lowerbound <- 13.30 - 1.5*groupS_IQR
groupNS_upperbound <- 33.60 + 1.5*groupNS_IQR

groupS[groupS<groupS_lowerbound]
groupNS[groupNS>groupNS_upperbound]
```

(b) (15 marks) Recreate the side-by-side box plots of percent of women on venires for Spock's judge and the other judges without identifying outliers.

```r
boxplot(groupS, groupNS,xlab="JUDGE", names=c("SPOCKS","OTHERS"),range=0)
title("Ruijie Sun 6046")
```

(c) (5 marks) Comment on the difference between the skeletal box plot (which does not identify outliers) and the modified box plot (which identifies outliers).

2 Consider the data, "assign1data.csv" based on the heights of 166 students in our class and answer the questions that follow.

(a)(5 marks) Was the data based on an experiment or an observational study? Briefly discuss the limitations on the statistical inference we can draw from this data.

(b) (5 marks) Which variables are categorical? Name the levels of each categorical variable.

(c) (20 marks) Conduct an appropriate hypothesis test to determine whether there is a difference between the heights of Males and Females.

```r
data <- read_csv("assign1data.csv")
attach(data)
head(data)
```

```r
groupM <- height[sex == "Male"]
groupFM <- height[ sex == "Female"]
```

i. Side-by-side boxplots

```r
boxplot(groupM, groupFM,xlab="sex", names=c("Male","Female"))
title("Ruijie Sun 6046")
```

ii. Null and Alternative Hypotheses

iii. A test statistic and it's distribution

```
var.test(groupM,groupFM)
```

```
t.test(groupM,groupFM,var.equal = T)
```

iv. Test assumptions

v. Test diagnostics (checking model assumptions)

```
par(mfrow=c(1,2))
hist(groupM,main = "Histogram for Male height",xlab = "groupM \n Ruijie Sun 6046")
qqnorm(groupM, xlab = "Theoretical Quantiles \n Ruijie Sun 6046")
qqline(groupM)
```

```
par(mfrow=c(1,2))
hist(groupFM,main = "Histogram for Female height",xlab = "groupFM \n Ruijie Sun 6046")
qqnorm(groupFM,xlab = "Theoretical Quantiles \n Ruijie Sun 6046")
qqline(groupFM)
```

vi. P-value

```
t.test(groupM,groupFM,var.equal = T)
```

vii. Results (brief discussion and conclusion)

(d) (5 marks) Name two(2) statistical methods which are equivalent to your method used in part (c) above.

(e) (25 marks) Create a subset of the data by removing the row of observations whose 'id' matches the last 2 digits of your student number. For instance, this can be done in R by shivon.subset <- shivon.data[-100,] if my student number ends with '00'. Then redo the analyses of part (c) above with your data subset.

```
data_subset <- subset(data, id != 46)
attach(data_subset)
```

```
groupM <- height[sex == "Male"]
groupFM <- height[ sex == "Female"]
```

i. Side-by-side boxplots

```
boxplot(groupM, groupFM,xlab="sex", names=c("Male","Female"),outline=FALSE)
title("Ruijie Sun 6046")
```

ii. Null and Alternative Hypotheses

iii. A test statistic and it's distribution

```
var.test(groupM,groupFM)
```

```
t.test(groupM,groupFM,var.equal = T)
```

iv. Test assumptions

v. Test diagnostics (checking model assumptions)

```
par(mfrow=c(1,2))
hist(groupM,main = "Histogram for Male height",xlab = "groupM \n Ruijie Sun 6046")
qqnorm(groupM, xlab = "Theoretical Quantiles \n Ruijie Sun 6046")
qqline(groupM)
```

```
par(mfrow=c(1,2))
hist(groupFM,main = "Histogram for Female height",xlab = "groupFM \n Ruijie Sun 6046")
qqnorm(groupFM,xlab = "Theoretical Quantiles \n Ruijie Sun 6046")
qqline(groupFM)
```

vi. P-value

```
t.test(groupM,groupFM,var.equal = T)
```

vii. Results (brief discussion and conclusion)

(f) (10 marks) Compare your results of part (c) and part (e). Do you think that the observation removed was influential?