MATH40005 Coursework Spring 2021

Tristan Peroy, CID: 01854740

Introduction

This project aims to decide if the difference between the average heights of people in countries X and Y is significant.

Question 1

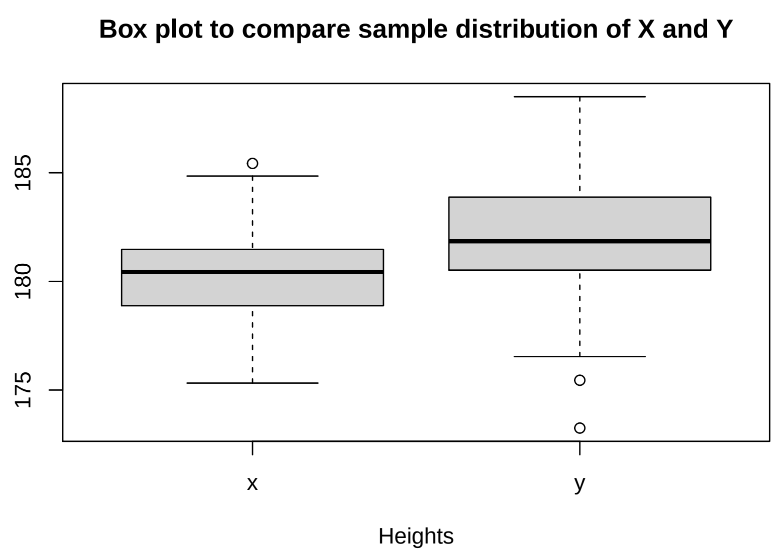
Read in the data.

```
df1 <- read.table("x_data.txt", sep=",", header=T)
x=df1$x
df2 <- read.table("y_data.txt", sep=",", header=T)
y=df2$y
```

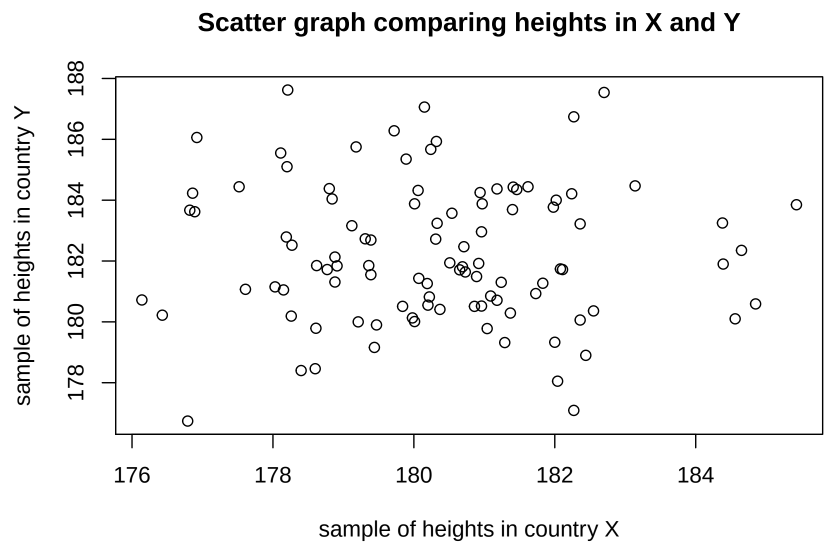
Question 2

The datasets have different sizes, so sample 100 elements of each and make a scatter plot. We also make a boxplot.

```
boxplot(x,y,names=c('x','y'),xlab='Heights',main='Box plot to compare sample distribution of X and Y')
```



```
x_sample=sample(x,100)
y_sample=sample(y,100)
plot(x_sample,y_sample,xlab="sample of heights in country X",
ylab="sample of heights in country Y",main='Scatter graph comparing heights in X and Y')
```



Question 3

The observations are $x_1, x_2, \ldots, x_n$ and the random variables are $X_1, X_2, \ldots, X_n$. Likewise for $y_1, y_2, \ldots, y_n$. The null hypothesis is:

- H_0 The average height of people in country X is equal to the average height of people in country Y.

The alternative hypothesis is:

- H_1 The average height of people in country X is different to the average height of people in country Y.

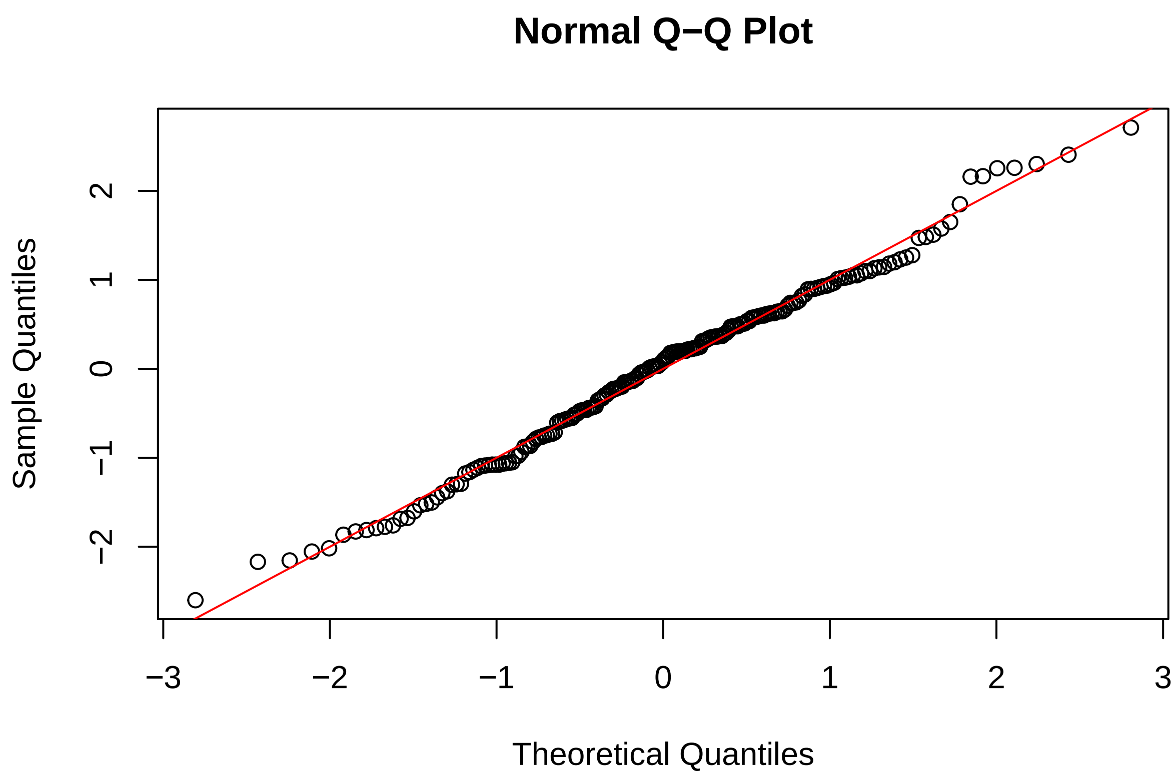
I plan to use Student's two-sample test, as we are comparing two population means. We need to assume the $X_1, X_2, \ldots, X_n$ follow a normal distribution with θ_1 mean and σ_1^2 distribution, while the $Y_1, Y_2, \ldots, Y_n$ follow a normal distribution with θ_2 mean and σ_2^2 distribution. I will assume also that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ in order to obtain a pooled sample variance. In summary, the means are unknown, the variances are unknown but assumed equal and we assume the random variables X and Y are independent.

The significance threshold is $\alpha = 0.01$.

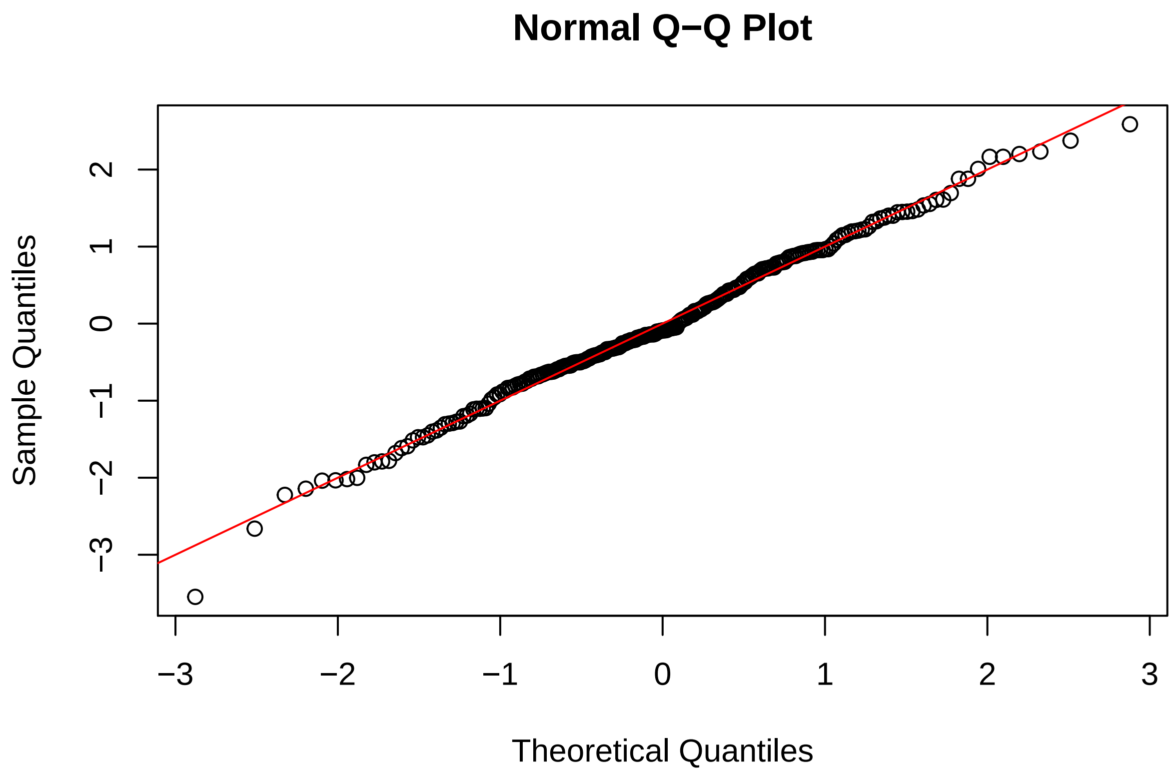
Question 4

I assumed normal distributions for the data. Test this using a qqplot on the samples. First standardize the data.We can see the data is normally distributed as assumed. We cannot test the mean or the variance as we only have a sample. In both cases the qqplot matches the y=x line quite closely so they both follow a normal distribution.

```r
new_x=(x-mean(x))/sqrt(var(x))
qqnorm(new_x)
abline(0, 1, col="red")
```




```r
new_y=(y-mean(y))/sqrt(var(y))
qqnorm(new_y)
abline(0, 1, col="red")
```




```r
mean(x);sd(x)
```

```
## [1] 180.2695
```

```
## [1] 1.903123
```

```r
mean(y);sd(y)
```

```
## [1] 182.0665
```

```
## [1] 2.485809
```

Question 5

We compute test statistic t and critical threshold to compare to the modulus of t. The threshold is from a t-distribution with n+m-2 levels of freedom where n and m are the lengths of x and y, and 1-alpha/2.

```r
sample_var_x=((1/(length(x)-1))*sum((x-mean(x))**2))
sample_var_y=((1/(length(y)-1))*sum((y-mean(y))**2))
sample_var_x
```

```
## [1] 3.621879
```

```r
sample_var_y
```

```
## [1] 6.179245
```

```r
pooled_sample_var=((length(x)-1)*sample_var_x+(length(y)-1)*sample_var_y)/(length(x)+length(y)-2)
pooled_sample_var
```

```
## [1] 5.043272
```

```r
t=(mean(x)-mean(y))/(sqrt(pooled_sample_var)*(sqrt(1/length(x)+1/length(y))))
t
```

```
## [1] -8.43486
```

```r
critical_threshold=qt(1-0.01/2,length(x)+length(y)-2)
critical_threshold
```

```
## [1] 2.586848
```

Question 6

In conclusion, the absolute value of the statistic t is greater than the critical value, therefore the null hypothesis is rejected, and the data supports the case that the average height of people in countries X and Y is different.