MATH40005 Coursework Spring 2021

Xiao Xuan Tan, [redacted]

Introduction

Testing if there is a statistically significant difference between the average heights of the people in countries X and Y.

Question 1

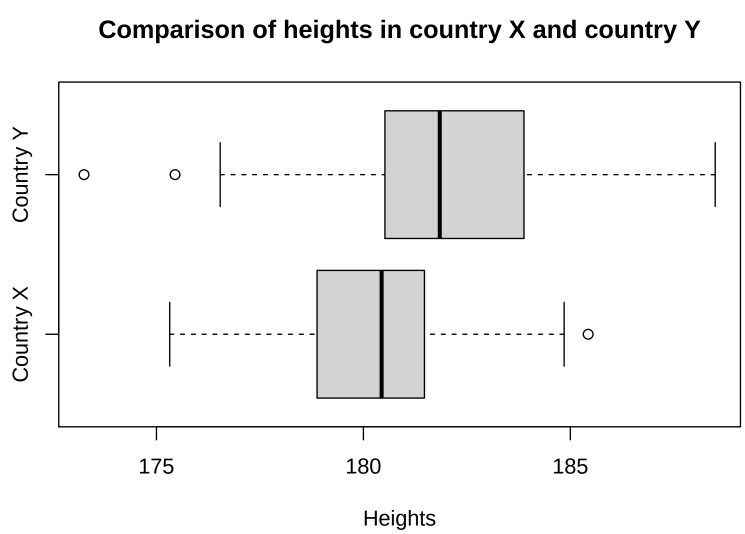
Read in the data.

```
df1<-read.table("~/Desktop/MATH40005_coursework_2021_questions/x_data.txt",
                sep=",",header=T)
df2<-read.table("~/Desktop/MATH40005_coursework_2021_questions/y_data.txt",
                sep=",",header=T)
x<-df1$x
y<-df2$y
```

Question 2

This box plot compares the median, quartiles, maximum and minimum of the heights in country X and Y.

```
boxplot(x,y,horizontal=TRUE,names=c("Country X","Country Y"),xlab="Heights",
        main="Comparison of heights in country X and country Y")
```



Question 3

Suppose n people are randomly selected in country X and their heights are the random variables $X_1, X_2, \ldots, X_n$, which are assumed to be independent and identically distributed following a normal distribution with unknown mean θ_1 and unknown variance σ_1^2. The observations are $x_1, x_2, \ldots, x_n$.

Suppose m people are randomly selected in country Y and their heights are the random variables $Y_1, Y_2, \ldots, Y_m$, which are assumed to be independent and identically distributed following a normal distribution with unknown mean θ_2 and unknown variance σ_2^2. The observations are $y_1, y_2, \ldots, y_m$.

Assume that σ_1^2=σ_2^2.

The null hypothesis is:

- H_0: $\theta_1 = \theta_2$

The alternative hypothesis is:

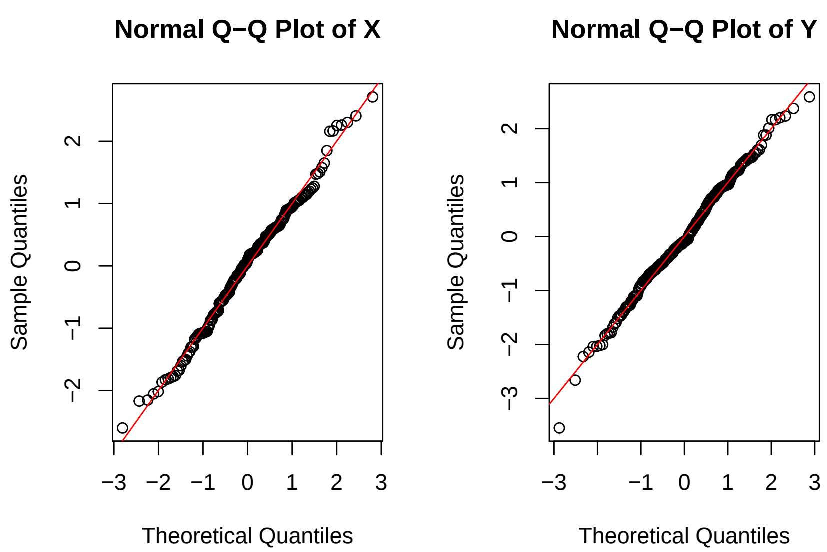
- H_1: $\theta_1 \neq \theta_2$

I plan to use the student's two sample test to test the hypothesis. The significance threshold is $\alpha = 0.05$.

Question 4

Assumption 1: Most of the points in both of the Q-Q plots lie along the line y = x, and so X and Y can each be assumed to be normally distributed.

```
layout(matrix(c(1,2),nrow=1,ncol=2,byrow=FALSE))
z<-(x-mean(x))/sd(x)
qqnorm(z,main="Normal Q-Q Plot of X")
abline(0,1,col="red")
w<-(y-mean(y))/sd(y)
qqnorm(w,main="Normal Q-Q Plot of Y")
abline(0,1,col="red")
```



Assumption 2: The ratio of the standard deviation of x and y is approximately 1, thus the variances of X and Y can be assumed to be equal.

```
cat("The standard deviation of x is: ", sd(x), "\n", sep="")
```

```
## The standard deviation of x is: 1.903123
```

```
cat("The standard deviation of y is: ", sd(y), "\n", sep="")
```

```
## The standard deviation of y is: 2.485809
```

```
cat("The ratio of the standard deviations of x and y is: ",sd(y)/sd(x), "\n", sep="")
```

```
## The ratio of the standard deviations of x and y is: 1.306173
```

Question 5

The t-statistic is

$$T = \frac{(\bar{X} - \bar{Y}) - (\theta_1 - \theta_2)}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

is the pooled sample variance.

```
n<-length(x)
m<-length(y)
x_bar<-mean(x)
y_bar<-mean(y)
x_var<-var(x)
y_var<-var(y)
pooled_variance<-((n-1)*x_var+(m-1)*y_var)/(n+m-2)
t<-(x_bar-y_bar)/(sqrt(pooled_variance)*sqrt(1/n+1/m))
cat("The test statistic t=", t, "\n", sep="")
```

```
## The test statistic t=-8.43486
```

```
alpha<-0.05
nu<-n+m-2
c<-qt(1-alpha/2,nu)
cat("The critical threshold is |t|>",c,"\n",sep="")
```

```
## The critical threshold is |t|>1.965273
```

Question 6

```
cat("Since the realised statistic |t|=",abs(t),">",c,", the null hypothesis is rejected.",
    "\n",sep="")
```

```
## Since the realised statistic |t|=8.43486>1.965273, the null hypothesis is rejected.
```

The data supports the case that there is a statistically significant difference between the average heights of people in countries X and Y.