Tutorial 2 Solution

Question 1

(a) Q: What is the response variable in this study?

Ans: FEV.

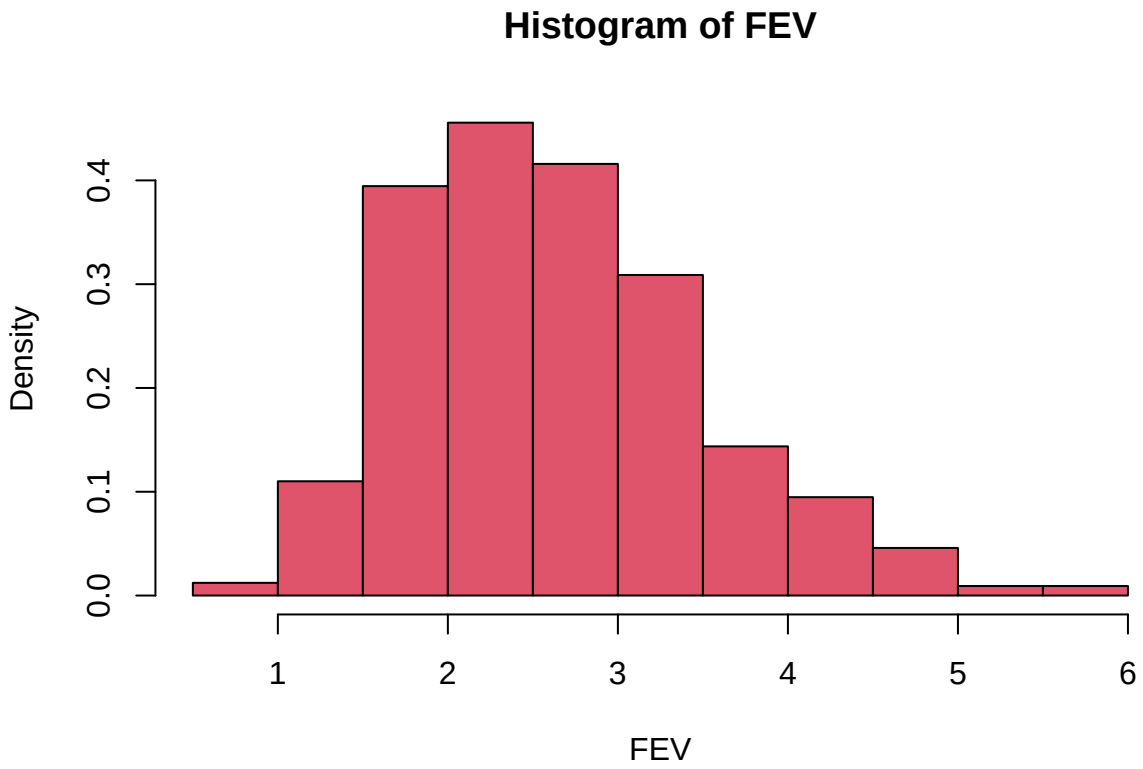
(b) Q: Create a histogram of FEV and comment on it.

```
fev = read.csv("C:/Data/FEV.csv")
names(fev)
```

```
## [1] "ID"     "Age"    "FEV"    "Hgt"    "Sex"    "Smoke"  "height"
```

```
attach(fev)
```

```
#Qb
hist(FEV, col = 10, freq= FALSE)
```

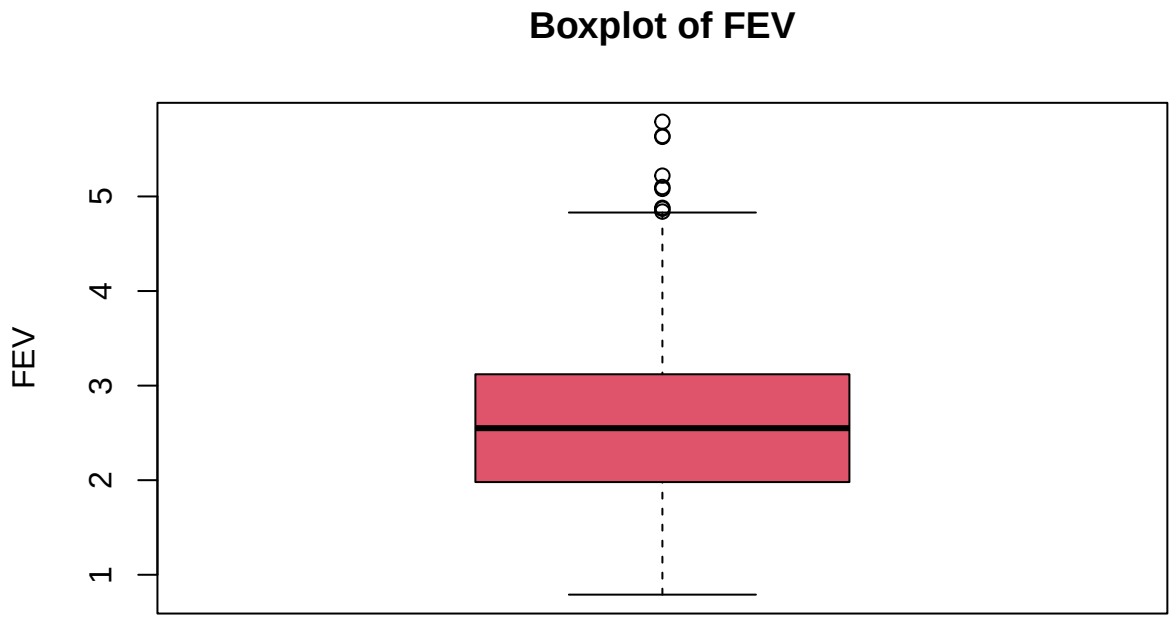


The histogram shows that the sample is unimodal. Compared to the overlaid normal density curve, the distribution looks slightly right-skewed. Most of the observations are within a range of 0.5 to 6 – there are no observations that are separated from the rest. However, this does not mean there are no outliers.

(c) Q: Create a boxplot of FEV and identify how many outliers there are. Investigate your data and

comment on these outliers.

```
#Qc
boxplot(FEV, col = 10, ylab = "FEV", main = "Boxplot of FEV")

#outlier values
out = boxplot(FEV, col = 10, ylab = "FEV", main = "Boxplot of FEV")$out
```




```
# get the index of the outliers:
index = which(FEV %in% c(out))
index
```

```
## [1] 321 452 464 517 609 624 632 648 649
```

```
#information of all the outliers:
fev[c(index),]
```

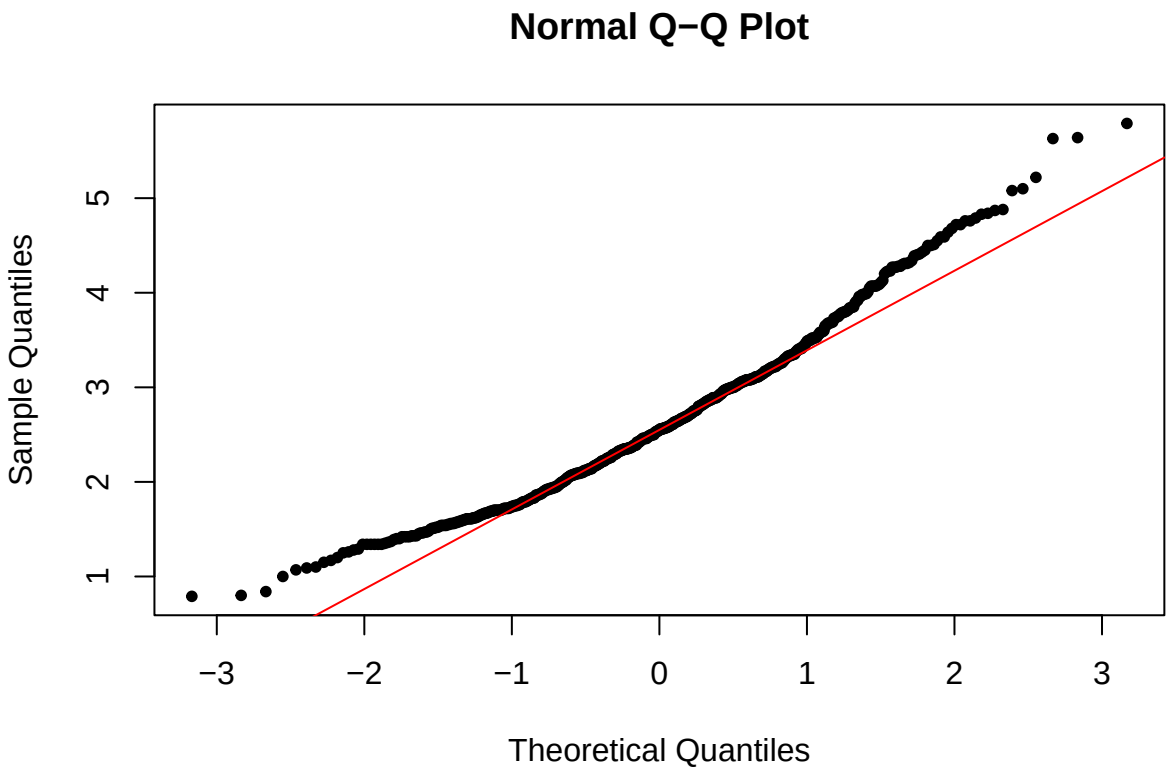
```
##        ID Age  FEV Hgt Sex Smoke height
## 321  2142  14 4.84  72   1     0   1.83
## 452 33041  12 5.22  70   1     0   1.78
## 464 37241  13 4.88  73   1     0   1.85
## 517 49541  13 5.08  74   1     0   1.88
## 609  6144  19 5.10  72   1     0   1.83
## 624 25941  15 5.79  69   1     0   1.75
## 632 37441  17 5.63  73   1     0   1.85
## 648 71141  17 5.64  70   1     0   1.78
## 649 71142  16 4.87  72   1     1   1.83
```

There are 9 outliers. Check the information of these 9 outliers, we would see that all these outliers are males,

most (8/9) are non-smokers, and they are rather tall.

(d) Q: Generally, is the sample of FEV normally distributed?

```
qqnorm(FEV, pch = 20)
qqline(FEV, col = "red")
```



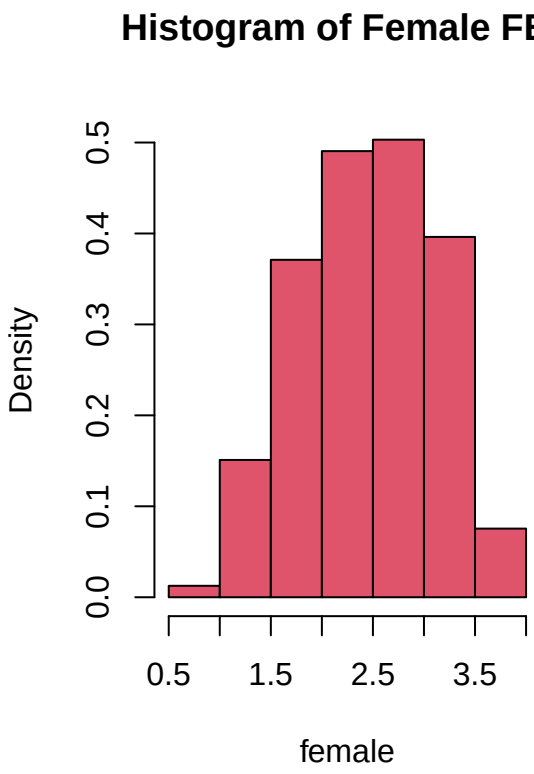
From the qq plot, on the left tail, the sample quantiles are smaller than expected (theoretical quantile) hence the left tail is shorter than normal.

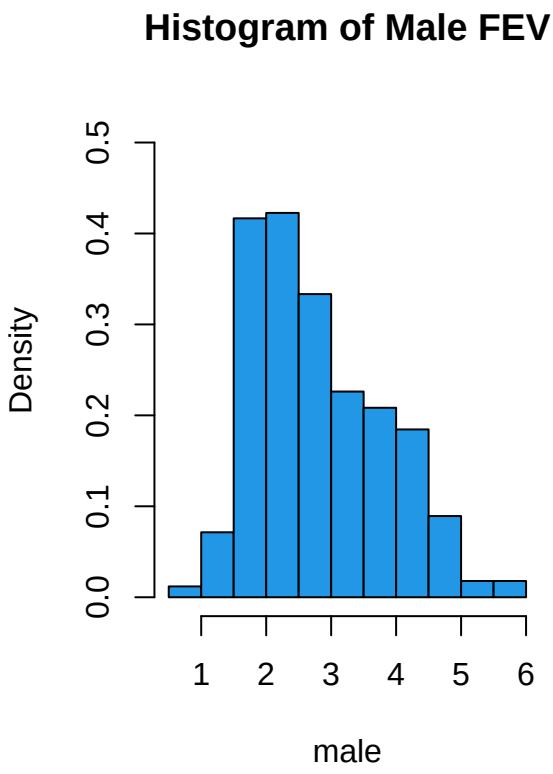
On the right side, the sample quantiles are larger than expected, hence the right tail is longer than normal.

Conclude: Combining with the histogram of FEV, it's clear that the sample of FEV is not normally distributed and quite right skewed.

(e) Q: Create separate histograms for male and female FEV, then obtain separate numerical summaries for males and female FEV. Comment on what you observe.

```
female = FEV[which(Sex==0)] # or FEV[Sex==0]

male = FEV[which(Sex==1)] # or FEV[Sex==1]

opar <- par(mfrow=c(1,2)) #arrange a figure of 1 row and 2 columns (to contain the 2 histograms)

hist(female, col = 2, freq= FALSE, main = "Histogram of Female FEV", ylim = c(0,0.52))

hist(male, col = 4, freq= FALSE, main = "Histogram of Male FEV", ylim = c(0,0.52))
```






```
par(opar)

median(female)
```

```
## [1] 2.49
```

```
IQR(female)
```

```
## [1] 1.04
```

```
summary(female)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.790   1.950   2.490   2.452   2.990   3.840
```

```
var(female)
```

```
## [1] 0.4169424
```

```
median(male)
```

```
## [1] 2.605
```

```
summary(male)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   2.007   2.605   2.813   3.535   5.790
```

```
IQR(male)
```

```
## [1] 1.5275
```

```
var(male)
```

```
## [1] 1.006866
```

The summary could be:

The shapes of the two histograms are quite different. Both are unimodal, but for females, it is almost symmetrical but for males it is quite right-skewed.

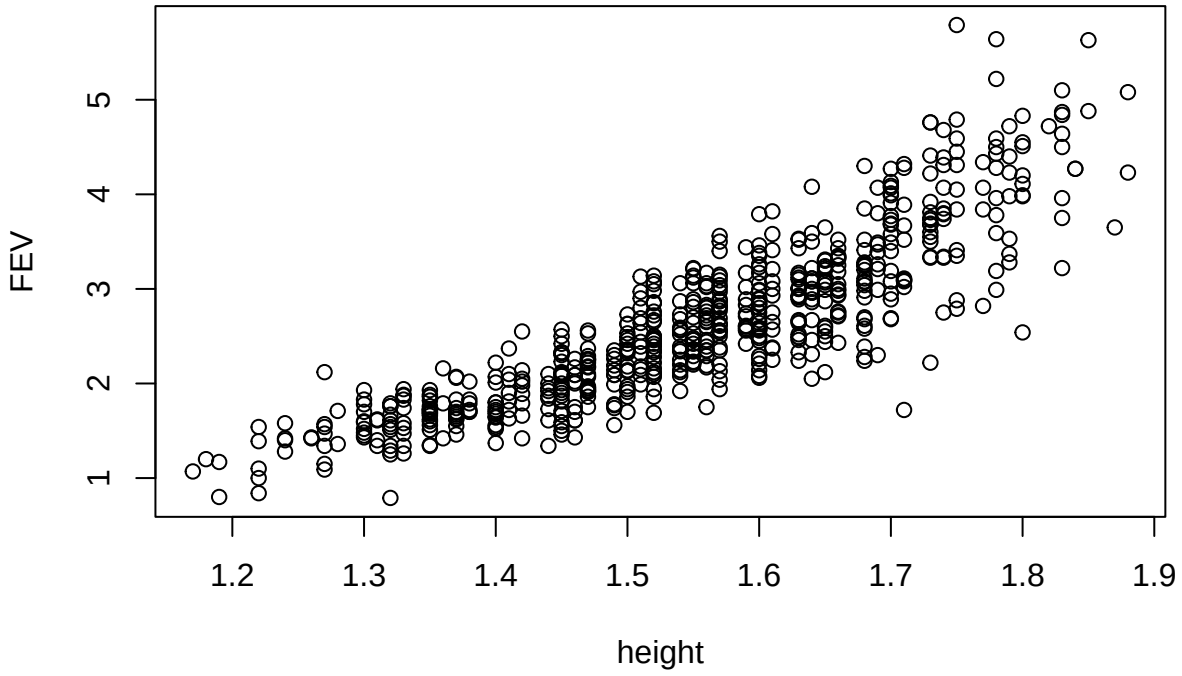
The median FEV for females is much lower than that for males (2.49 compared to 2.605).

In addition, the variability in the male group is higher than the variability in the female group.

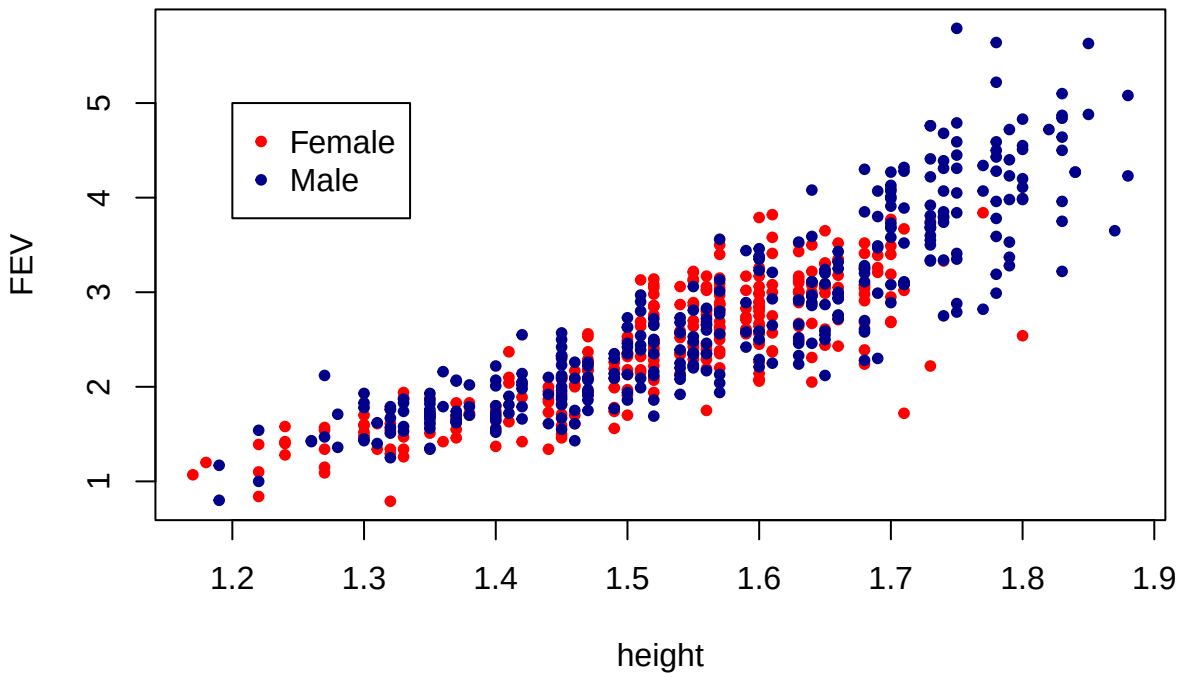
The respective IQR are 1.54 and 1.05.

(f & g) Q: Create a scatterplot with height (in metres) on the x-axis and FEV on the y-axis.

```
plot(height, FEV)
```




```
plot(height, FEV, type = "n")
points(female ~ height[which(Sex==0)], col = "red", pch = 20)
points(male ~ height[which(Sex==1)], col = "darkblue", pch = 20)
legend(1.2, 5, legend = c("Female", "Male"), col = c("red","darkblue"), pch=c(20,20))
```




```
cor(FEV, height)
```

```
## [1] 0.8675619
```

The computed correlation is quite high, and it is clear from the plot that there is a strong positive linear association between the two variables overall. The range of FEV for males appears larger than the range for females, as does the range of heights. The variability of FEV at lower heights does seem to be slightly less than the variability of FEV at greater heights.

Question 2

The Fibonacci numbers is a sequence of numbers $\{F_n\}$ defined by the following recursive relationship

$$F_n = F_{n-1} + F_{n-2}, \quad n > 3$$

with $F_1 = F_2 = 1$.

(a) Write the code to create a vector *Fibo* that contains the first 45 terms of the sequence.

(b) Report the 40th term of the Fibonacci sequence. Write the code to determine the smallest n such that F_n is larger than 5,000,000 (five million). Report the value of that F_n.

```
#(a)
Fibo = numeric(45)
Fibo[1:2] = 1
for(i in 3:45) {
Fibo[i] = Fibo[i-1] + Fibo[i-2]
}
```

```r
#(b)
Fibo[40]
```

```
## [1] 102334155
```

```r
# the 40th term is: 102334155

n = sum(Fibo<=5000000) + 1 ;n # 34
```

```
## [1] 34
```

```r
 # OR can use this code
n = max(which(Fibo<=5000000)) + 1; n
```

```
## [1] 34
```

```r
Fibo[n] # 5702887
```

```
## [1] 5702887
```