2. Explorando bases

Francisco Castorena, A00827756

2023-08-16

```
library(e1071)
library(moments)
```

```
##
## Attaching package: 'moments'
```

```
## The following objects are masked from 'package:e1071':
##
##     kurtosis, moment, skewness
```

```
df = read.csv('mc-donalds-menu-1.csv')
```

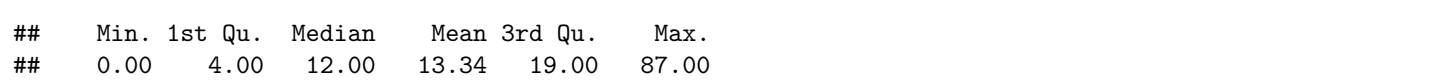
```
head(df)
```

```
##    Category                           Item   Serving.Size Calories
## 1 Breakfast                 Egg McMuffin 4.8 oz (136 g)      300
## 2 Breakfast            Egg White Delight 4.8 oz (135 g)      250
## 3 Breakfast             Sausage McMuffin 3.9 oz (111 g)      370
## 4 Breakfast    Sausage McMuffin with Egg 5.7 oz (161 g)      450
## 5 Breakfast Sausage McMuffin with Egg Whites 5.7 oz (161 g)      400
## 6 Breakfast         Steak & Egg McMuffin 6.5 oz (185 g)      430
##   Calories.from.Fat Total.Fat Total.Fat....Daily.Value. Saturated.Fat
## 1               120        13                        20             5
## 2                70         8                        12             3
## 3               200        23                        35             8
## 4               250        28                        43            10
## 5               210        23                        35             8
## 6               210        23                        36             9
##   Saturated.Fat....Daily.Value. Trans.Fat Cholesterol
## 1                            25         0         260
## 2                            15         0          25
## 3                            42         0          45
## 4                            52         0         285
## 5                            42         0          50
## 6                            46         1         300
##   Cholesterol....Daily.Value. Sodium Sodium....Daily.Value. Carbohydrates
## 1                          87    750                    31            31
## 2                           8    770                    32            30
## 3                          15    780                    33            29
## 4                          95    860                    36            30
```

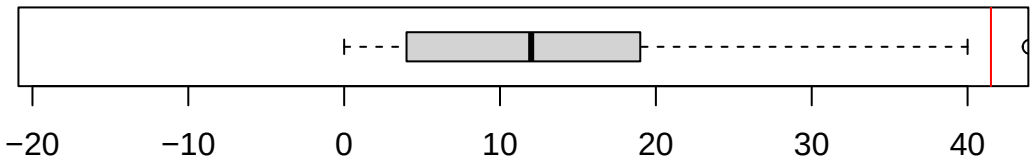
```
## 5                          16    880                    37            30
## 6                         100    960                    40            31
##   Carbohydrates....Daily.Value. Dietary.Fiber Dietary.Fiber....Daily.Value.
## 1                            10             4                            17
## 2                            10             4                            17
## 3                            10             4                            17
## 4                            10             4                            17
## 5                            10             4                            17
## 6                            10             4                            18
##   Sugars Protein Vitamin.A....Daily.Value. Vitamin.C....Daily.Value.
## 1      3      17                        10                         0
## 2      3      18                         6                         0
## 3      2      14                         8                         0
## 4      2      21                        15                         0
## 5      2      21                         6                         0
## 6      3      26                        15                         2
##   Calcium....Daily.Value. Iron....Daily.Value.
## 1                      25                   15
## 2                      25                    8
## 3                      25                   10
## 4                      30                   15
## 5                      25                   10
## 6                      30                   20
```

Análisis de la variable Protein

```
X = df$Protein
q1=quantile(X,0.25)  #Cuantil 1 de la variable X
q3=quantile(X,0.75)
ri= q3-q1   #Rango intercuartílico de X
par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
y1 = q1-1.5*ri
y2 = q3+1.5*ri
boxplot(X,horizontal=TRUE,ylim=c(y1,y2))
abline(v=q3+1.5*ri,col="red")  #linea vertical en el límite de los datos atípicos o extremos
X1= X[X<q3+1.5*ri]  #En la matriz M, quitar datos más allá de 3 rangos intercuartílicos arriba de q3 de
summary(X)
```



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    4.00   12.00   13.34   19.00   87.00
```



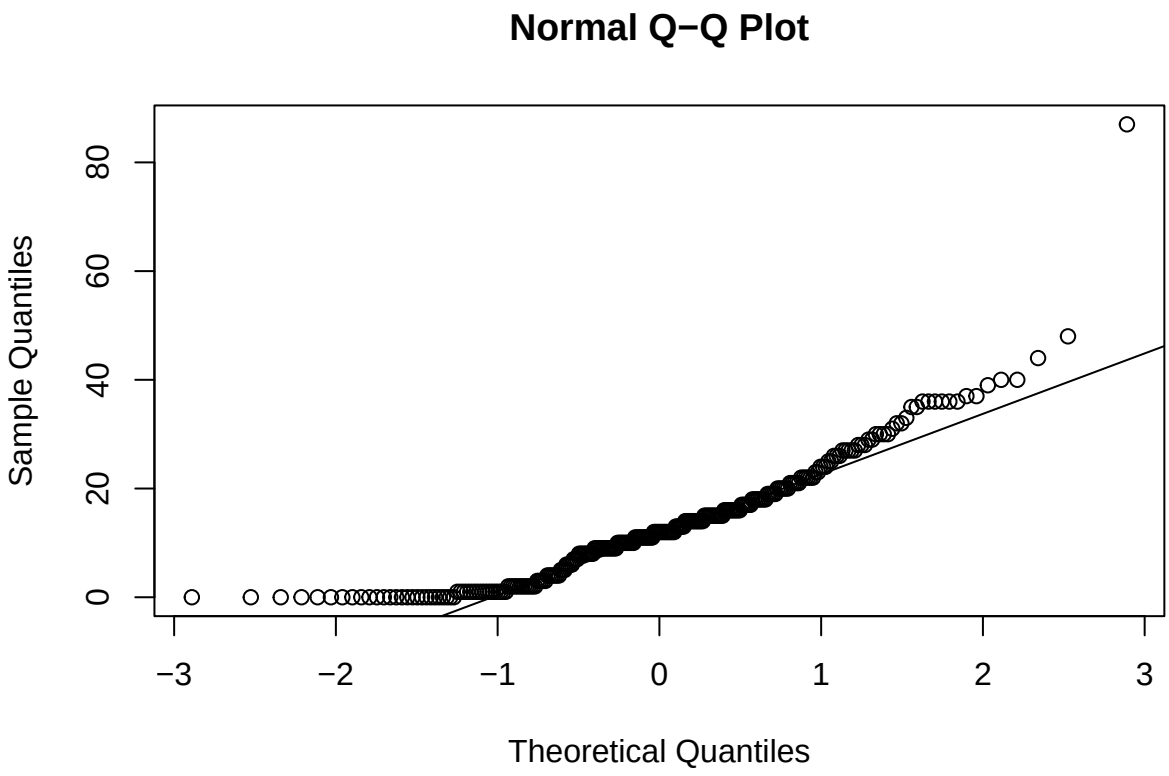
En el boxplot anterior se puede observar la distribución de los datos de la variable Protein del dataset, aquellos puntos que excedan la linea roja serán eliminados debido a que estan más de 3 cuartiles arriba del tercer cuartil, por lo que son considerados datos extremos.

```
summary(X1)
```

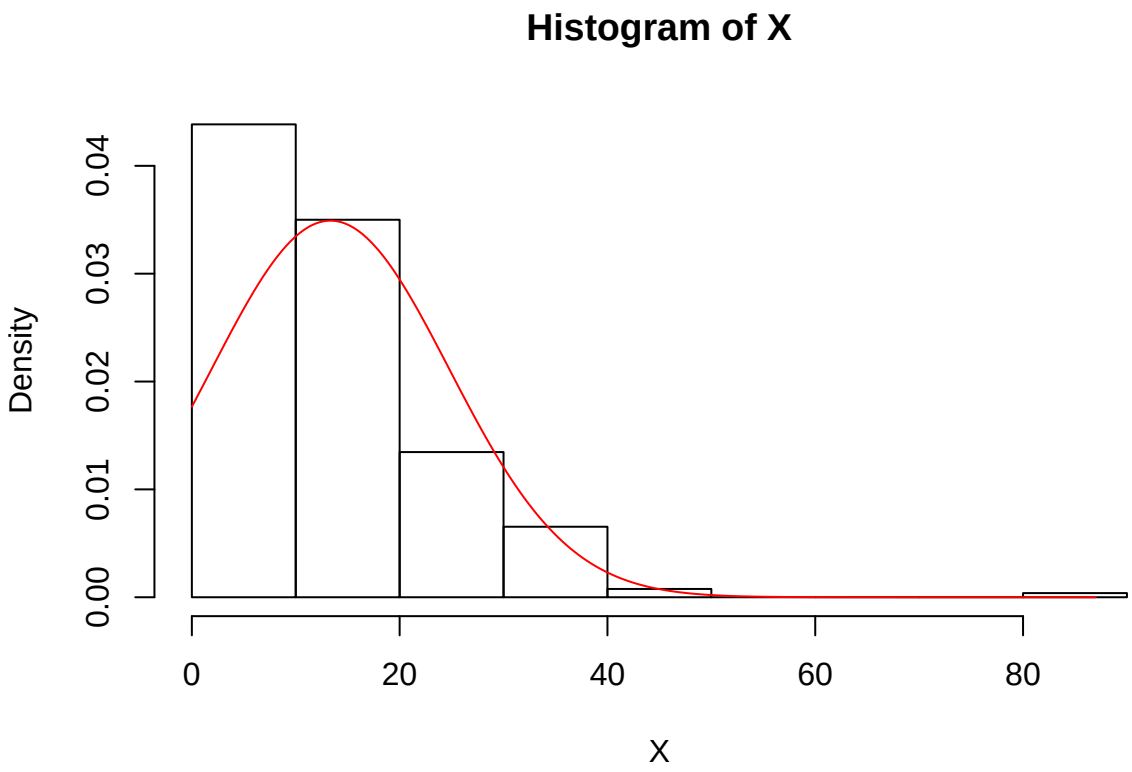
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     4.0    12.0    12.8    18.0    40.0
```

Podemos observar que al eliminar los datos atípicos el valor de la media y el tercer cuartil cambian un poco, sin embargo ya no tenemos valores que esten más de 3 cuartiles por encima del tercer cuartil.

```
qqnorm(X)
qqline(X)
```

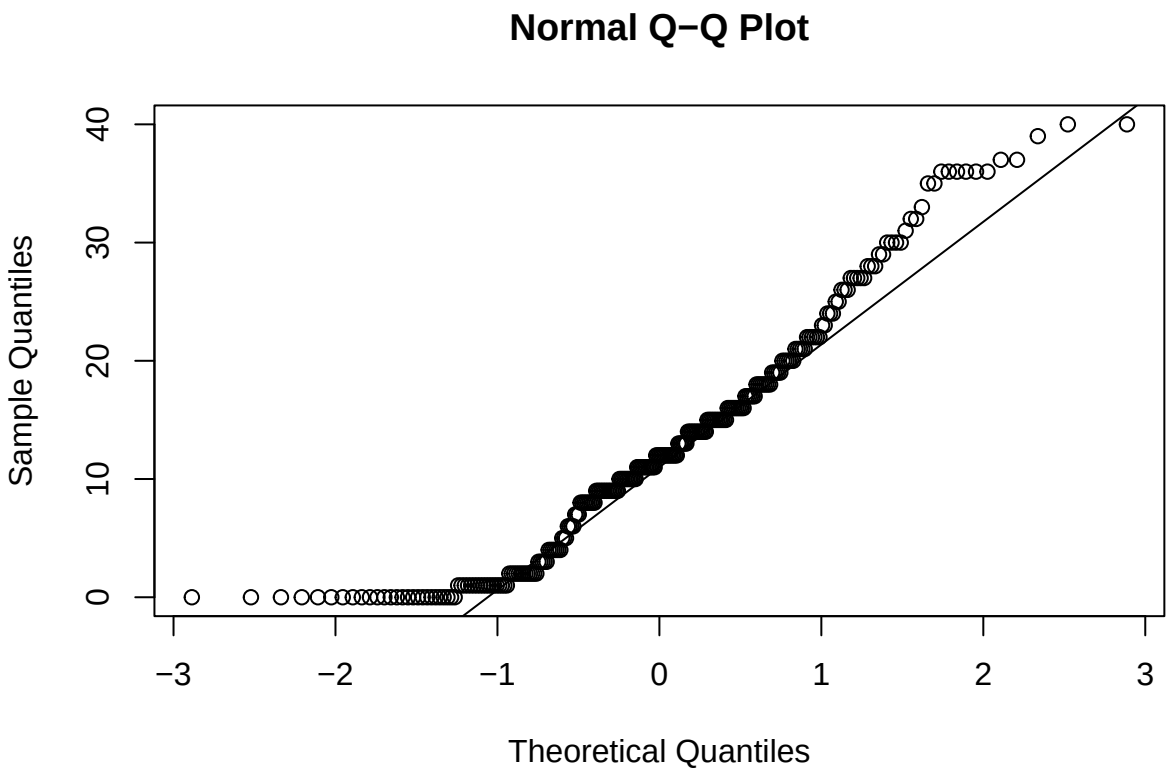



```r
hist(X,prob=TRUE,col=0)
x=seq(min(X),max(X),0.1)
y=dnorm(x,mean(X),sd(X))
lines(x,y,col="red")
```

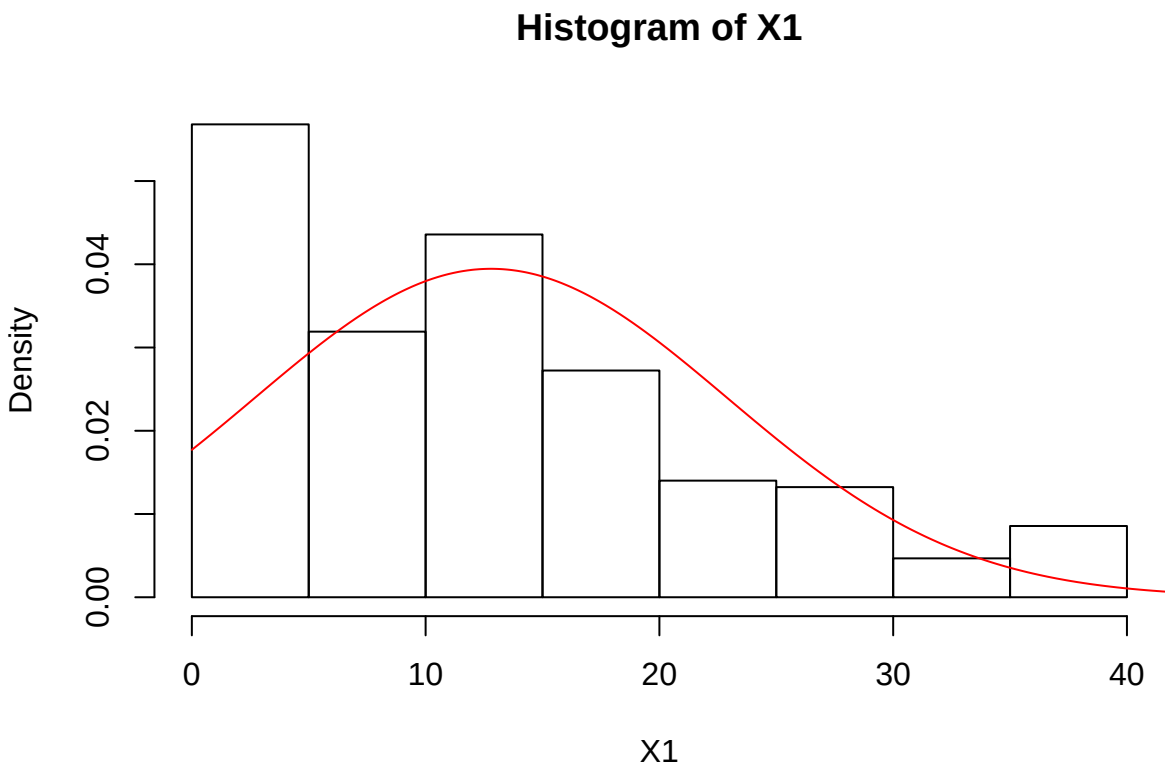


Podemos ver en la gráfica de residuos el valor atípico en la esquina superior derecha, lo cual hace que se tengan colas más pesadas, por otra parte, en el histograma se observa un sesgo a la derecha, tal vez asemejándose a una distribución Weibull más que a una normal.

```
qqnorm(X1)
qqline(X1)
```




```r
hist(X1,prob=TRUE,col=0)
x=seq(min(X1),max(X),0.1)
y=dnorm(x,mean(X1),sd(X1))
lines(x,y,col="red")
```



Podemos observar como después de hacer la limpieza de datos atípicos, el histograma tiene un menor sesgo a la derecha, aunque aun se tiene un sesgo considerabe para rechazar que esta variable se comporte como una distribución normal, debido a que únicamente se tenía un dato atípico es lógico que el peso de eliminar un solo dato sea poco para cambiar de manera significativa la distribución de los valores de la variable.

```
library(moments)
skewness(X)
```

```
## [1] 1.570794
```

```
kurtosis(X)
```

```
## [1] 8.86355
```

Podemos ver para la medida de sesgo (skewness) que esta es mayor a 0, lo cual nos indica que se tiene un sesgo hacia la derecha como se pudo observar claramente en el histograma anterior, por otra parte se observa que la curtosis es de 8.86, esta al ser mayor a tres es considerada como leptocúrtica lo cual nos indica que los datos están mayormente agrupados alrededor de la media.

```
skewness(X1)
```

```
## [1] 0.6742646
```

```r
kurtosis(X1)
```

```
## [1] 2.83266
```

Podemos ver como para los datos tratados el sesgo disminuyo lo cual acerca esto más a una distribución normal, igualmente la curtosis es muy cercana a tres, asemejandose a una curtosis mesocúrtica, esto indica una distribucion de datos más amplia alrededor de la media.

```r
library(nortest)
lillie.test(X)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X
## D = 0.12153, p-value = 5.806e-10
```

el p-value es muy bajo, se rechaza la hipótesis nula de distribución normal.

```r
lillie.test(X1)
```

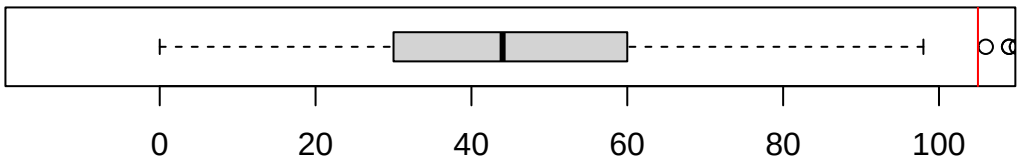
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X1
## D = 0.10281, p-value = 6.406e-07
```

De la misma forma se sigue rechazando la hipótesis nula, aún después de la eliminación de datos atípicos la distribución no es considerada normal.

Análisis de la variable Carbohydrates

```r
X = df$Carbohydrates
q1=quantile(X,0.25)  #Cuantil 1 de la variable X
q3=quantile(X,0.75)
ri= q3-q1   #Rango intercuartílico de X
par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
y1 = q1-1.5*ri
y2 = q3+1.5*ri
boxplot(X,horizontal=TRUE,ylim=c(y1,y2))
abline(v=q3+1.5*ri,col="red")  #linea vertical en el límite de los datos atípicos o extremos
X1= X[X<q3+1.5*ri]  #En la matriz M, quitar datos más allá de 3 rangos intercuartílicos arriba de q3 de
summary(X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   44.00   47.35   60.00  141.00
```



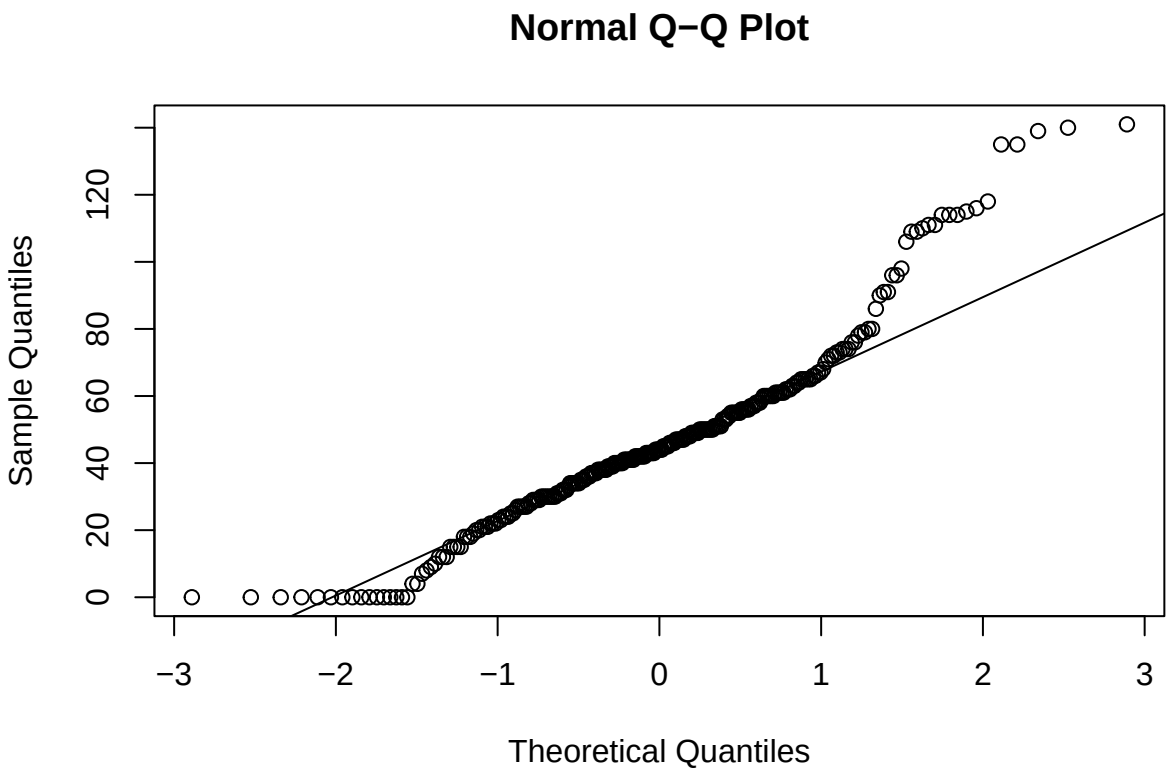
En esta variable se puede observar una mayor cantidad de datos atípicos que serán eliminados debido a su lejanía en comparación con los otros datos.

```
summary(X1)
```

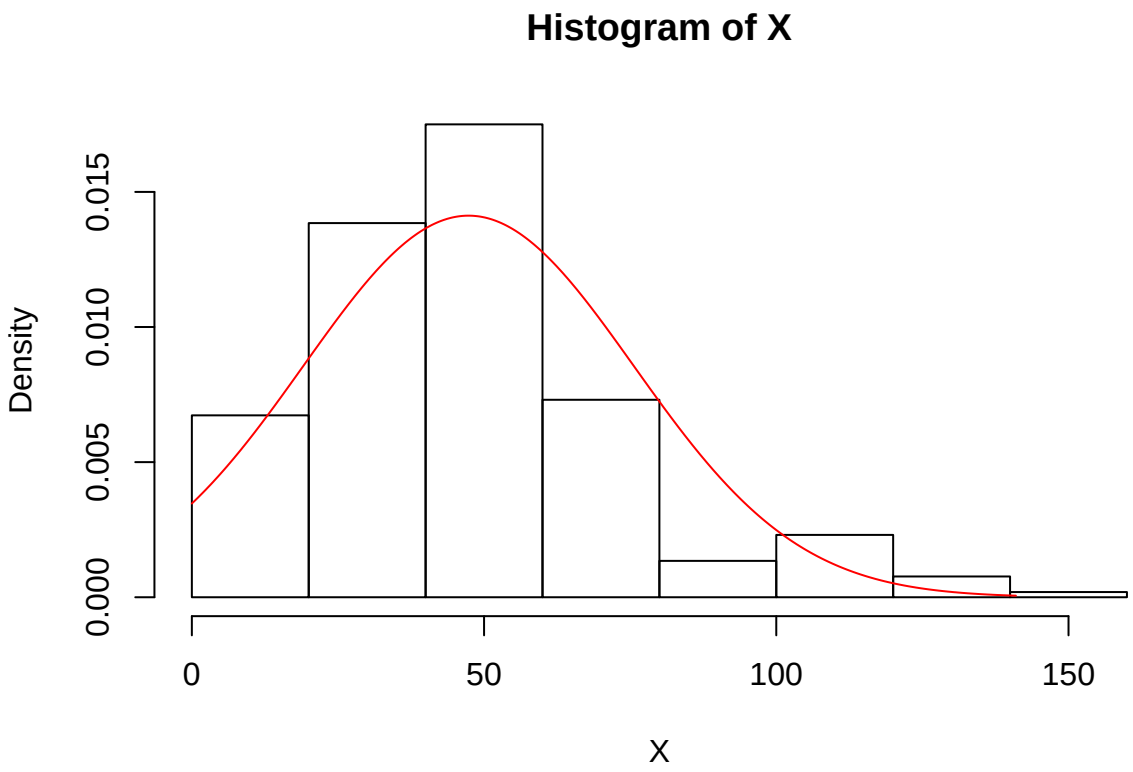
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   43.00   42.28   56.00   98.00
```

El valor de la media y el tercer cuartil cambian un poco después de la eliminación de datos atípicos, sin embargo ya no tenemos valores que estén más de 3 cuartiles por encima del tercer cuartil.

```
qqnorm(X)
qqline(X)
```

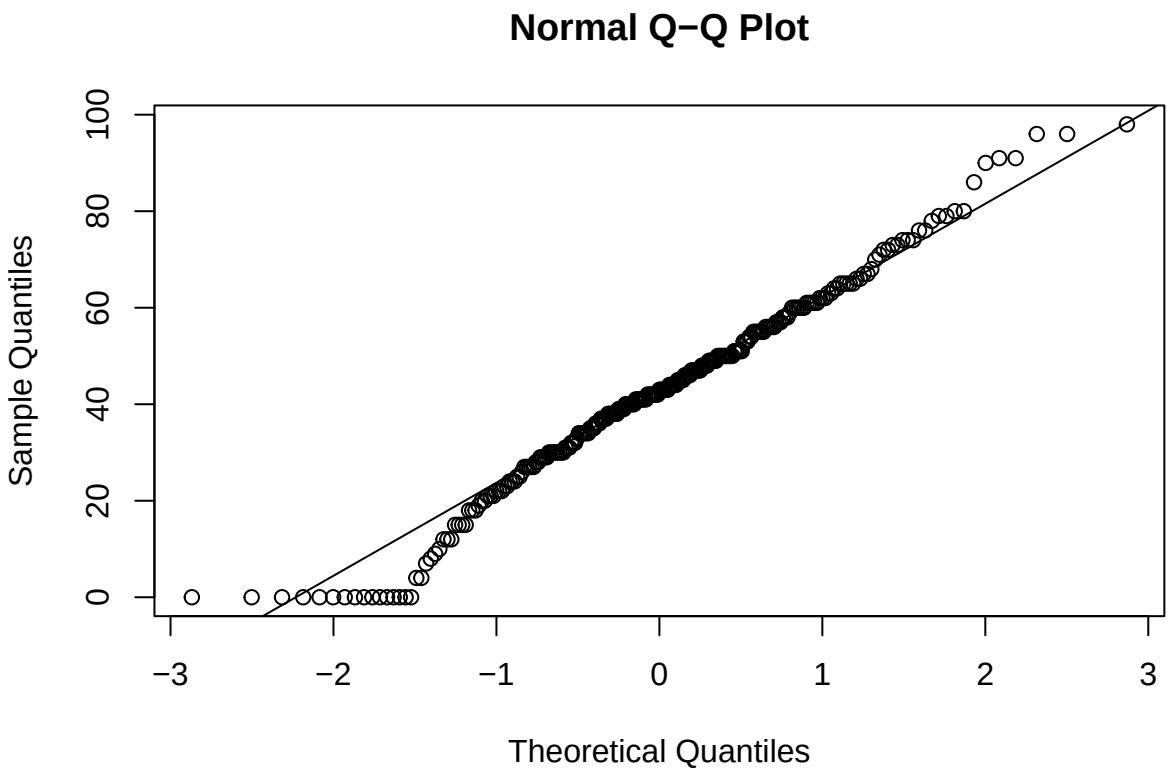



```r
hist(X,prob=TRUE,col=0)
x=seq(min(X),max(X),0.1)
y=dnorm(x,mean(X),sd(X))
lines(x,y,col="red")
```

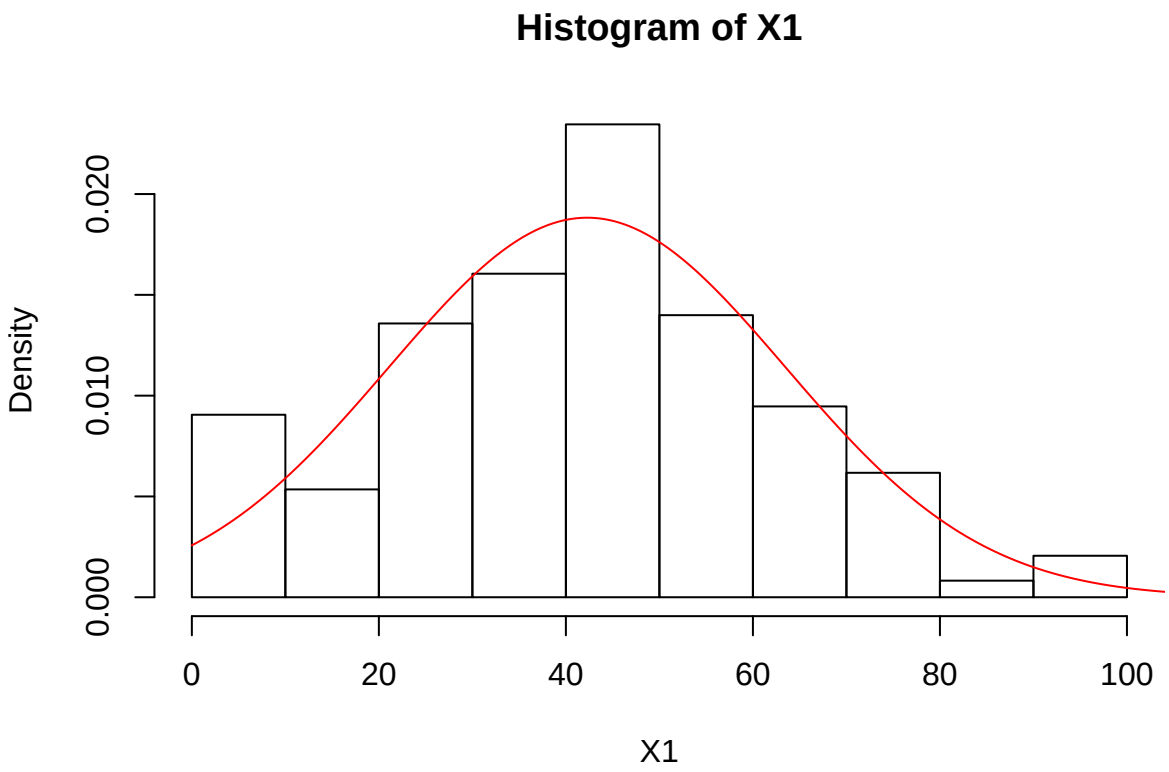


En la gráfica de residuos se observan colas pesadas, sobretodo para los valores superiores de la variable, por otra parte se tiene un sesgo a la derecha en el histograma mostrado, aunque este es menor al de la variable Protein antes analizada.

```
qqnorm(X1)
qqline(X1)
```




```r
hist(X1,prob=TRUE,col=0)
x=seq(min(X1),max(X),0.1)
y=dnorm(x,mean(X1),sd(X1))
lines(x,y,col="red")
```



Se observan colas menos pesadas y una gran disminución en el sesgo derecho que se tenía en el histograma, a primera vista podemos decir que el histograma se asemeja mucho a lo que es una distribución normal.

```
library(moments)
skewness(X)
```

```
## [1] 0.9074253
```

```
kurtosis(X)
```

```
## [1] 4.357538
```

Podemos ver algo de sesgo hacía la derecha y un valor de curtosis mayor a 3, lo cual indica una distribución leptocúrtica que distribuye los datos mayormente alrededor de la media.

```
skewness(X1)
```

```
## [1] -0.02861759
```

```
kurtosis(X1)
```

```
## [1] 2.931357
```

Podemos ver como el sesgo es mucho menor de lo que se tenía anteriormente, incluso el valor ahora es negativo lo cual indica un poco de sesgo hacía la izquierda, por otra parte ahora tenemos una distribución mesocúrtica.

```
lillie.test(X)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X
## D = 0.098548, p-value = 2.081e-06
```

el p-value es muy bajo, se rechaza la hipótesis nula de distribución normal.

```
lillie.test(X1)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X1
## D = 0.045565, p-value = 0.2523
```

El p-value ahora es mucho más grande por lo cual se acepta la hipítesis nula de que la variable se distribuye de forma normal.