Actividad Integradora 1

Oscar Gutierrez

2024-08-20

Cargar dataset

```
M= read.csv("food_data_g.csv")
```

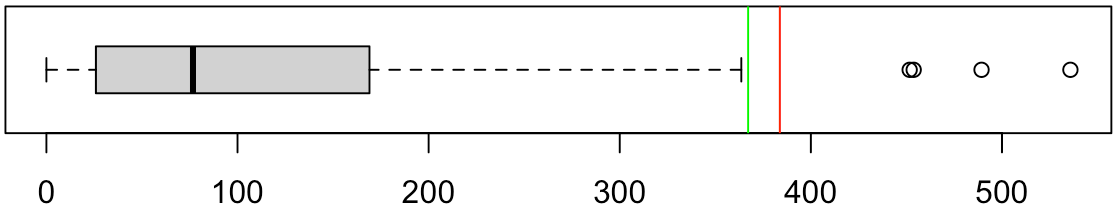
Seleccionar la variable

```
agua = M$Water
```

Analizar datos atípicos

```
q1=quantile(agua, 0.25)
q3 = quantile(agua, 0.75)
ri=IQR(agua)    #Rango intercuartílico de X
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(agua,horizontal=TRUE)   #y1=min en la escala del eje Y, y2=máx en la escala del eje Y
abline(v=q3+1.5*ri, col="red")  #linea vertical en el límite de los datos atipicos o extremos
abline(v= mean(agua)+ 3*sd(agua), col="green") # linea vertical a 3 sd de la media
abline(v=q3+3*ri, col="blue") # linea vertival a 3 ri
summary(agua)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    25.9    76.7   101.7   169.1   535.8
```



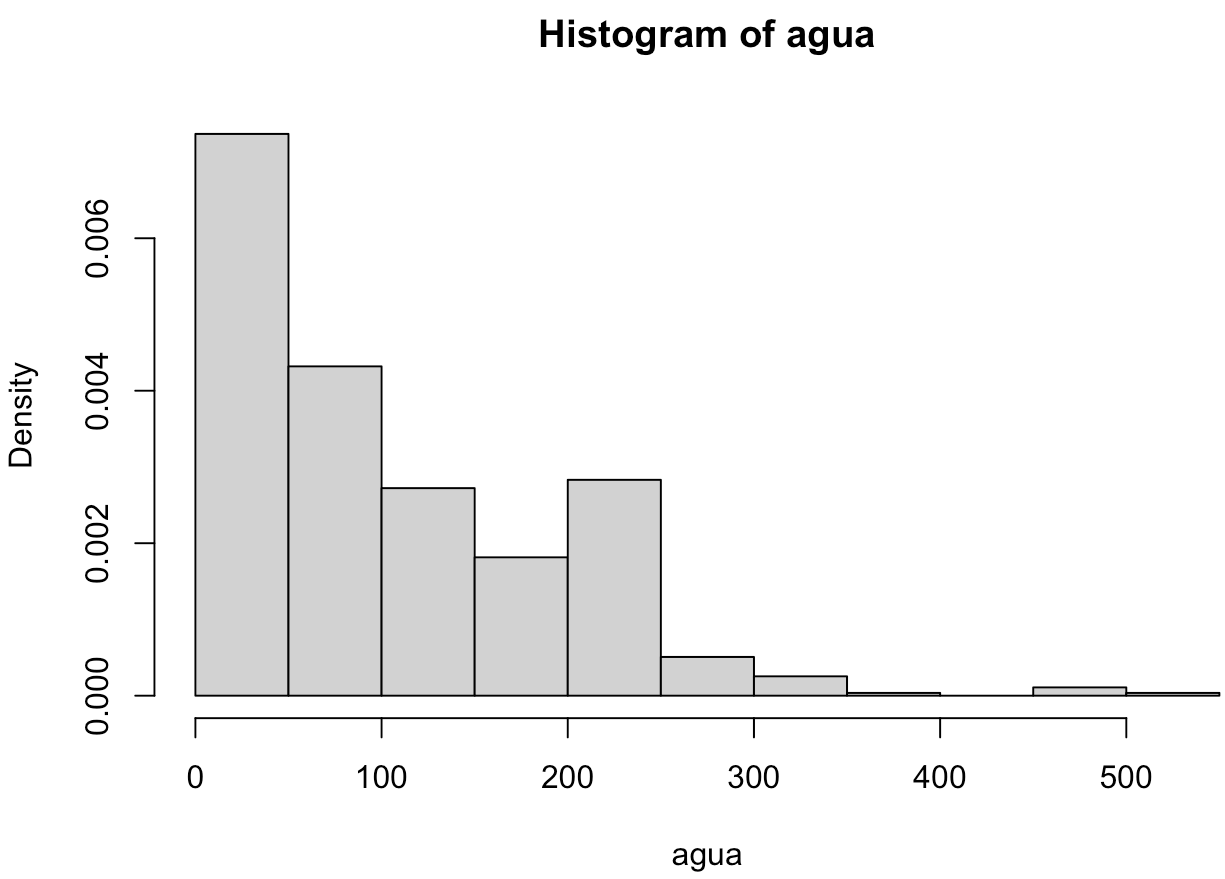

```
DA = M[M$Water > mean(agua)+ 3*sd(agua), ]
DA
```

```
##       X Unnamed..0                         food Caloric.Value  Fat
## 67   66         66              kung pao chicken           779 42.2
## 158 157        157             chicken chow mein           513 16.9
## 199 198        198 chicken mushroom chowder soup           431 23.7
## 248 247        247                 escarole soup            61  4.0
##     Saturated.Fats Monounsaturated.Fats Polyunsaturated.Fats Carbohydrates
## 67             8.2                 13.1                18.2          41.5
## 158            3.0                  3.7                 7.4          50.1
## 199            6.2                  4.5                 9.4          38.3
## 248            1.2                  1.8                 0.8           4.0
##     Sugars Protein Dietary.Fiber Cholesterol Sodium Water Vitamin.A Vitamin.B1
## 67    18.3    59.0           9.1       157.0    2.4 451.7     0.000        0.2
## 158   10.5    40.8           6.0        96.6    1.9 489.3     0.093        0.2
## 199    0.0    16.2           7.5        32.3    1.8 453.8     0.000        0.0
## 248    0.0     3.4           0.0         5.5    2.3 535.8     0.000        0.2
##     Vitamin.B11 Vitamin.B12 Vitamin.B2 Vitamin.B3 Vitamin.B5 Vitamin.B6
## 67        0.000       0.034        0.3       16.7        3.0        1.5
## 158       0.000       0.092        0.1        8.9        1.6        1.1
## 199       0.000       0.000        0.0        0.0        0.0        0.0
## 248       0.058       0.060        0.1        5.1        0.4        0.5
##     Vitamin.C Vitamin.D Vitamin.E Vitamin.K Calcium Copper Iron Magnesium
## 67       42.9         0       6.2     0.012   120.8    0.4  4.6     145.0
## 158      12.1         0       2.6     0.100   126.8    0.2  4.0      66.4
## 199      10.8         0       0.0     0.000     0.0    0.0  2.6       0.0
## 248      10.0         0       0.0     0.000    71.9    0.8  1.7      11.1
##     Manganese Phosphorus Potassium Selenium Zinc Nutrition.Density
## 67        1.5      567.8    1316.7    0.097  4.5           320.100
## 158       0.6      326.2     749.0    0.086  1.9           256.797
## 199       0.0        0.0       0.0    0.000  0.0            99.100
## 248       2.8      177.0     591.7    0.000  5.0            95.000
```

En la gráfica de caja y bigote se puede observar una linea verde, la cual corresponde a la cota de 1.5 ri, y una linea roja la cual corresponde a 3 desviaciones estándar, la cota de 3 ri no se alcanza a observar ya que se sale del límite derecho. De acuerdo con ambos criterios, existen 4 datos atípicos correspondientes a alimentos como sopas. # Normalidad

Histograma

```
hist(agua,freq=FALSE)
```



Los datos no se distribuyen normalmente.

Pruebas de normalidad Anderson Darling y Jarque Bera

h0: Los datos siguen una distribucion normal h1: Los datos no siguen una distribucion normal

```
library(nortest)
ad.test(agua)
```

```
##
##  Anderson-Darling normality test
##
## data:  agua
## A = 15.968, p-value < 2.2e-16
```

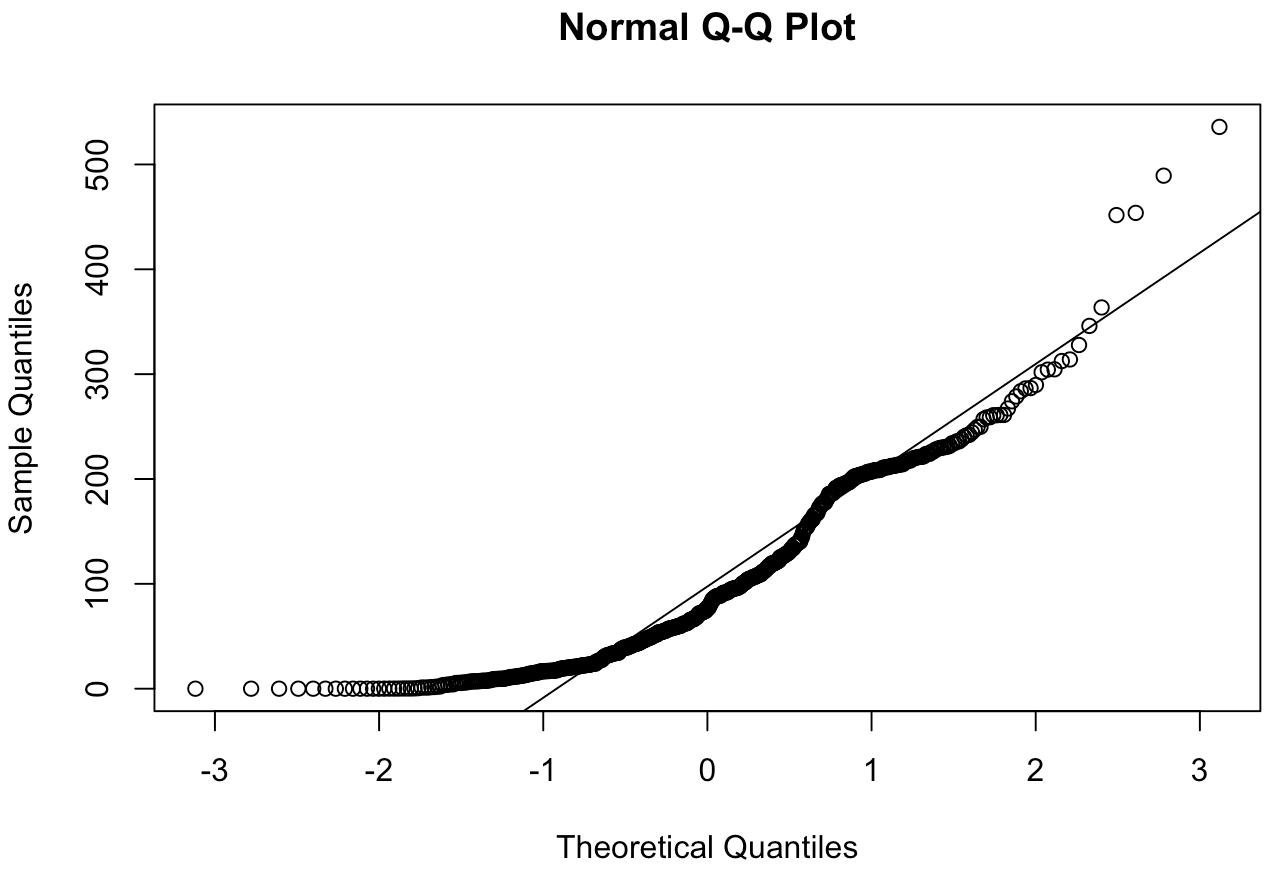
```
library(moments)
jarque.test(agua)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  agua
## JB = 153.58, p-value < 2.2e-16
## alternative hypothesis: greater
```

Ambas pruebas rechazan h0.

QQ plot

```
qqnorm(agua)
qqline(agua)
```



Sesgo y Curtosis

```
cat("sesgo= ",skewness(agua), "\ncurtosis=",kurtosis(agua))
```

```
## sesgo=  1.083794
## curtosis= 4.411058
```

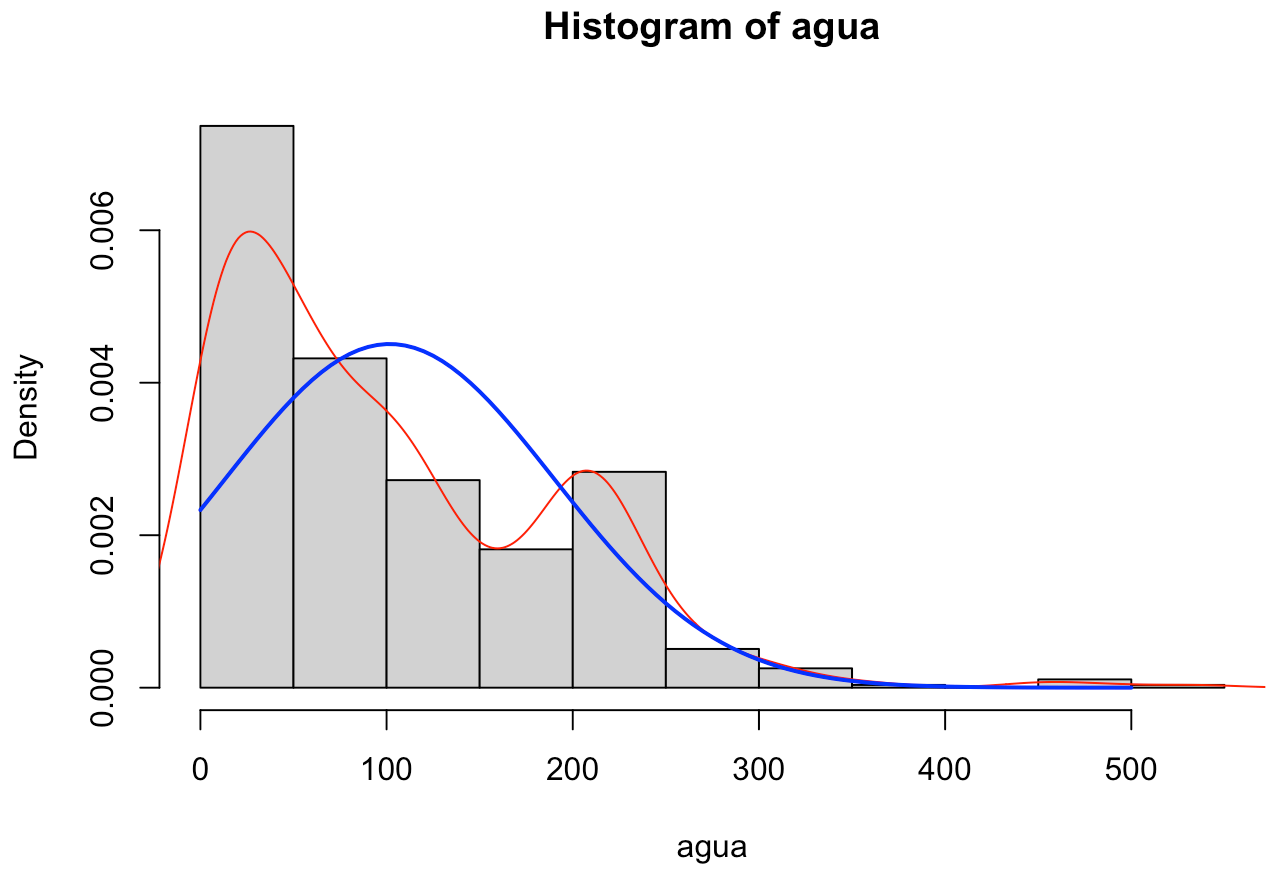
Media, mediana y rango medio

```
cat("media=", mean(agua), "\nmediana=",median(agua), "\nrango medio=", (max(agua)-min(ag
ua))/2)
```

```
## media= 101.6587
## mediana= 76.7
## rango medio= 267.9
```

Grafico de densidad empirica y teorica

```
hist(agua, freq = FALSE)
lines(density(agua), col = "red")
curve(dnorm(x, mean = mean(agua), sd = sd(agua)), from = 0, to = 500, add = TRUE, col =
"blue", lwd = 2)
```



Las pruebas de normalidad de Anderson Darling y Jarque Bera rechazan h0, además, los gráficos como el QQ plot y la comparación de densidad empírica y teórica indican que no hay presencia de normalidad. Otro factor a considerar es que los coeficientes de sesgo y curtosis estan lejos de ser valores de una distribución normal ya que deberían ser 0 y 3 respectivamente. Además, los valores de media, mediana y rango medio son iguales en una distribución normal, mientras que en los datos no se tiene esta característica.

Transformacion a normalidad

Transformacion

```
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

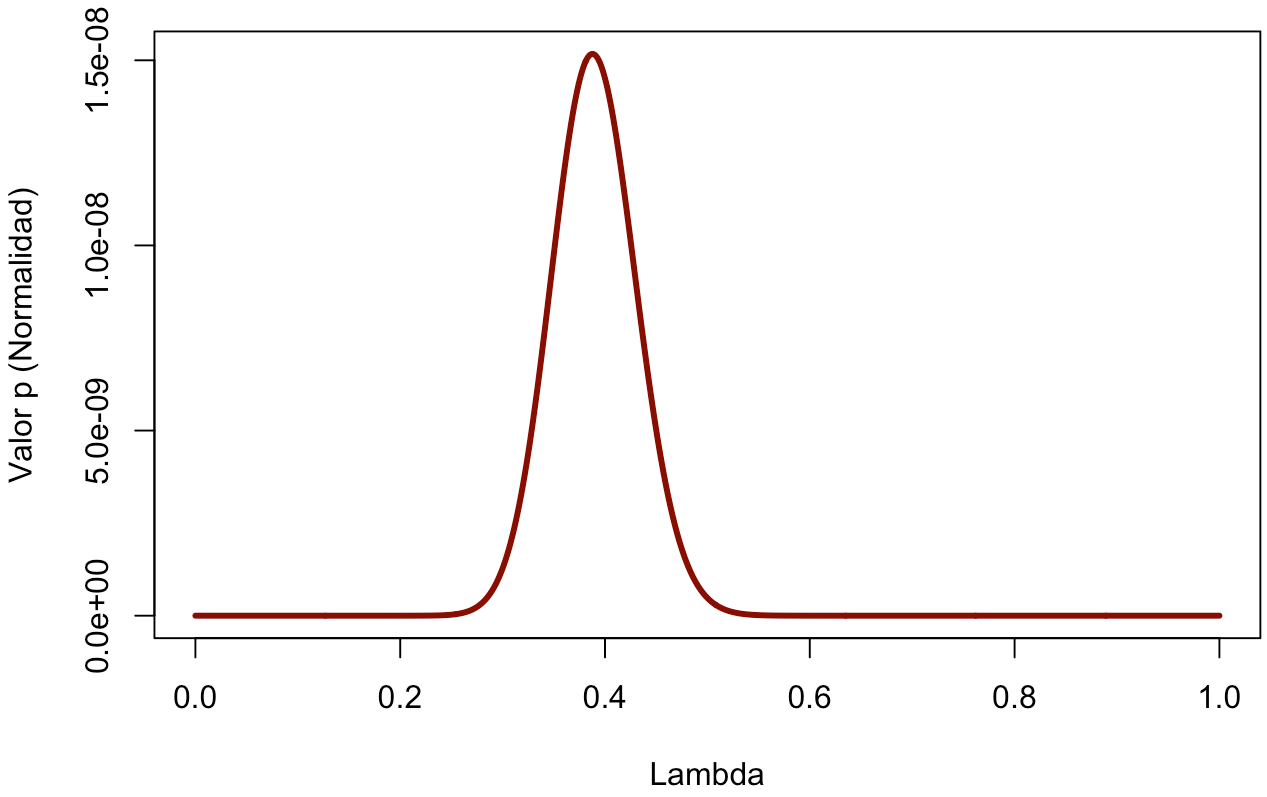
```
lp <- seq(0,1,0.001) # Valores de lambda propuestos
nlp <- length(lp)
n=length(agua)
D <- matrix(as.numeric(NA), ncol=2, nrow=nlp)
d <- NA

for (i in 1:nlp) {
  d = yeo.johnson(agua, lambda = lp[i])
  p = ad.test(d)
  D[i,] = c(lp[i], p$p.value)
}

# Convert matrix to data frame and name the columns
N <- as.data.frame(D)
colnames(N) <- c("Lambda", "Valor-p")

# Remove any rows with NA or infinite values
N <- N[is.finite(N$`Lambda`) & is.finite(N$`Valor-p`), ]

# Now, plot the data
plot(N$Lambda, N$`Valor-p`, type="l", col="darkred", lwd=3, xlab="Lambda", ylab="Valor p
(Normalidad)")
```

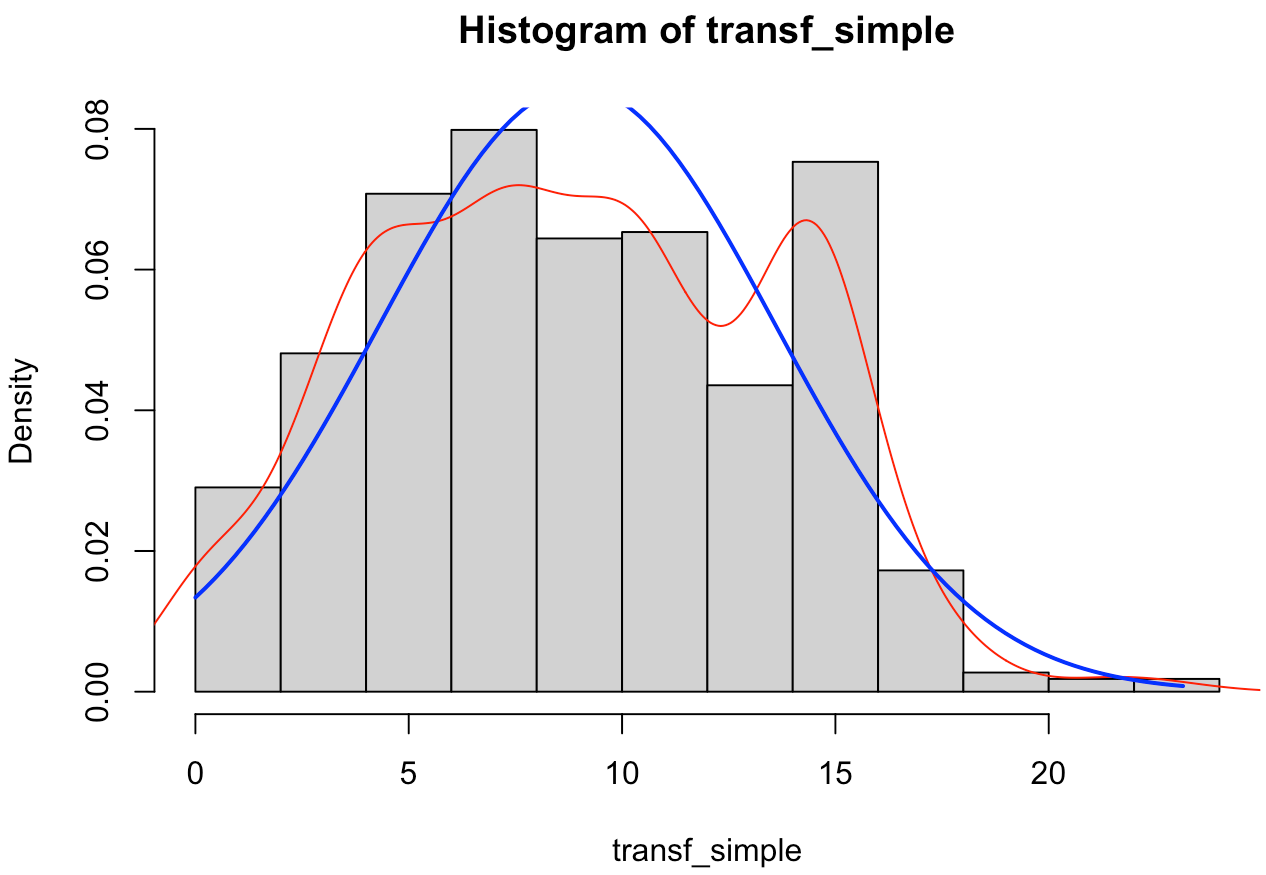



```
G=data.frame(subset(N,N$`Valor-p`==max(N$`Valor-p`)))
l = G$Lambda
```

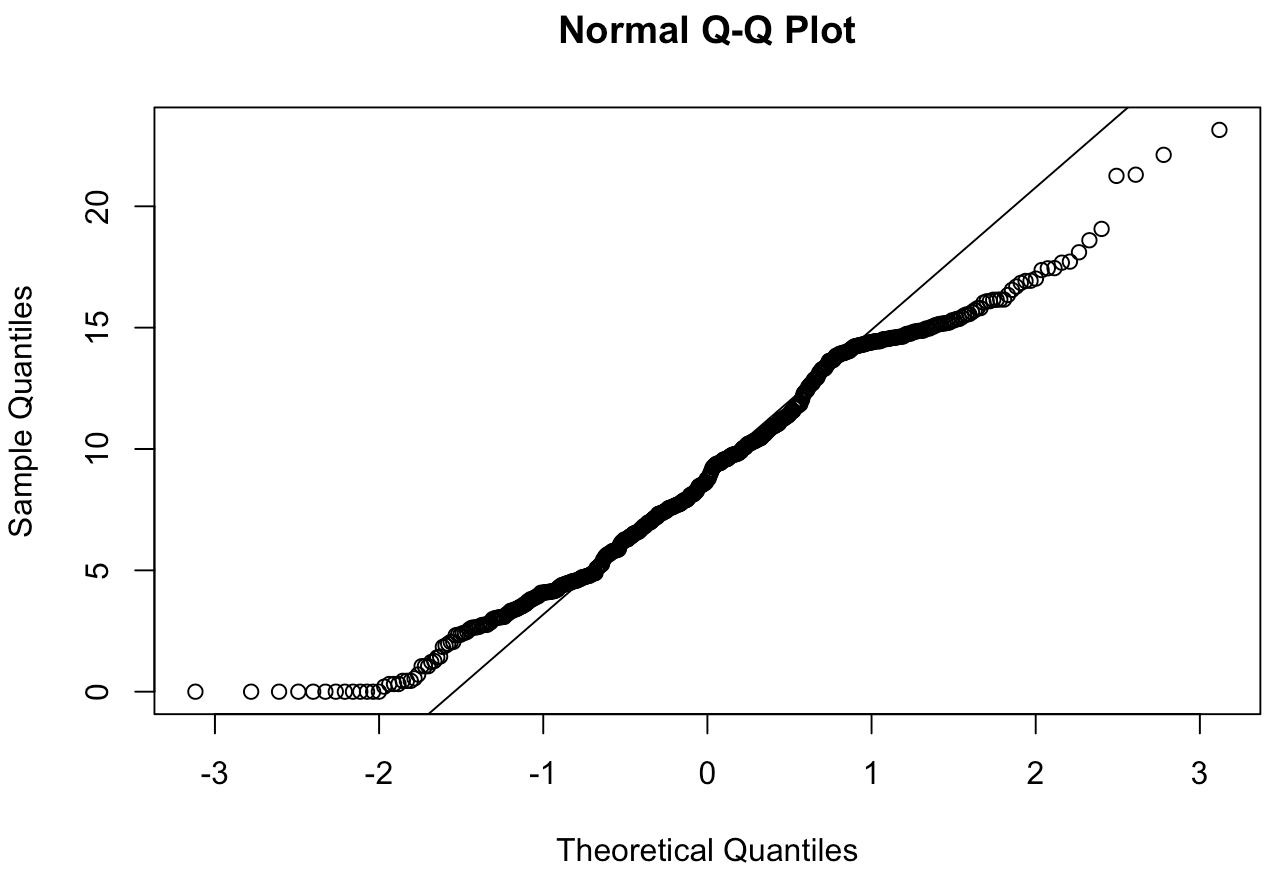
Transformacion simple

La ecuacion para la transformacion simple es $\sqrt{x}$

```
transf_simple = sqrt(agua)
hist(transf_simple,freq=FALSE)
lines(density(transf_simple), col = "red")
curve(dnorm(x, mean = mean(transf_simple), sd = sd(transf_simple)), from = min(transf_si
mple), to = max(transf_simple), add = TRUE, col = "blue", lwd = 2)
```



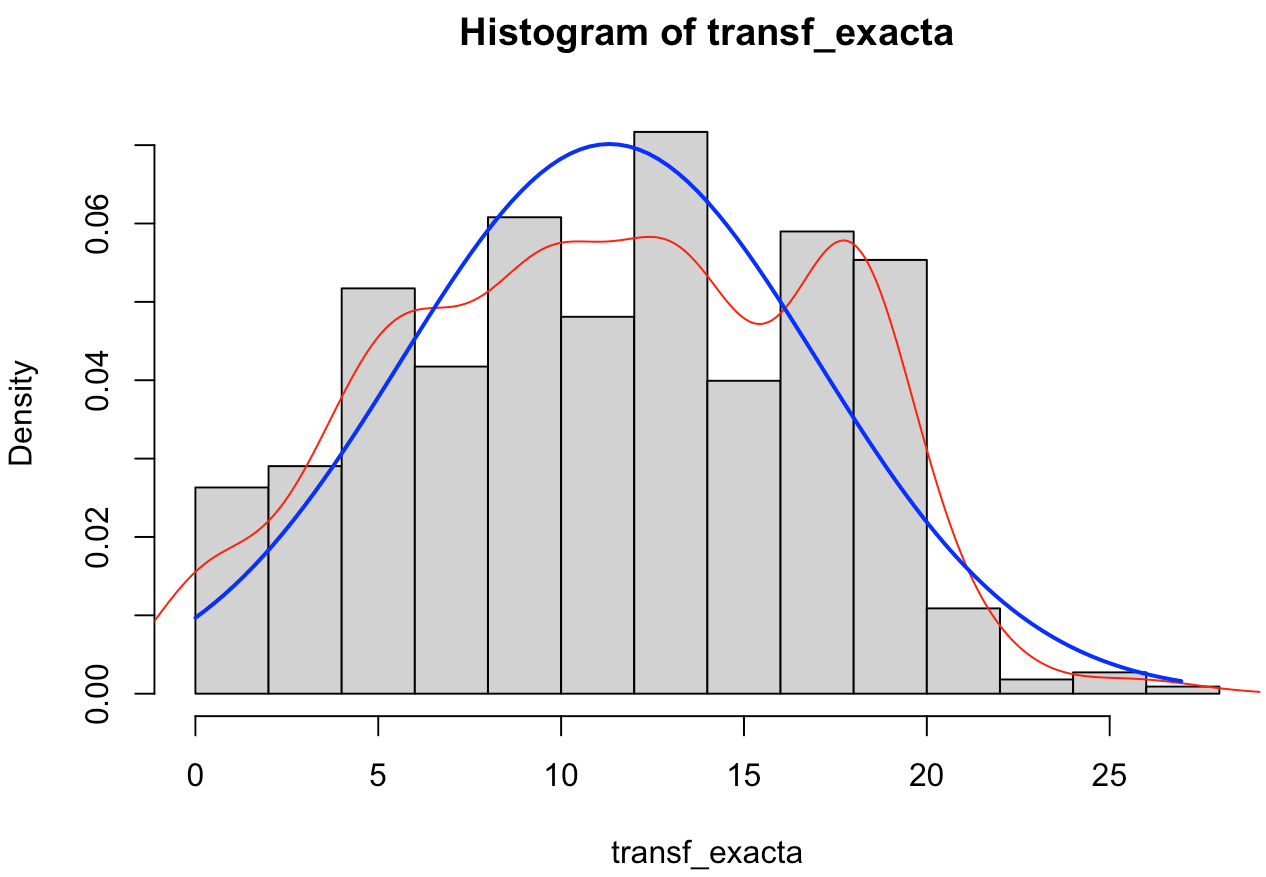

```
qqnorm(transf_simple)
qqline(transf_simple)
```



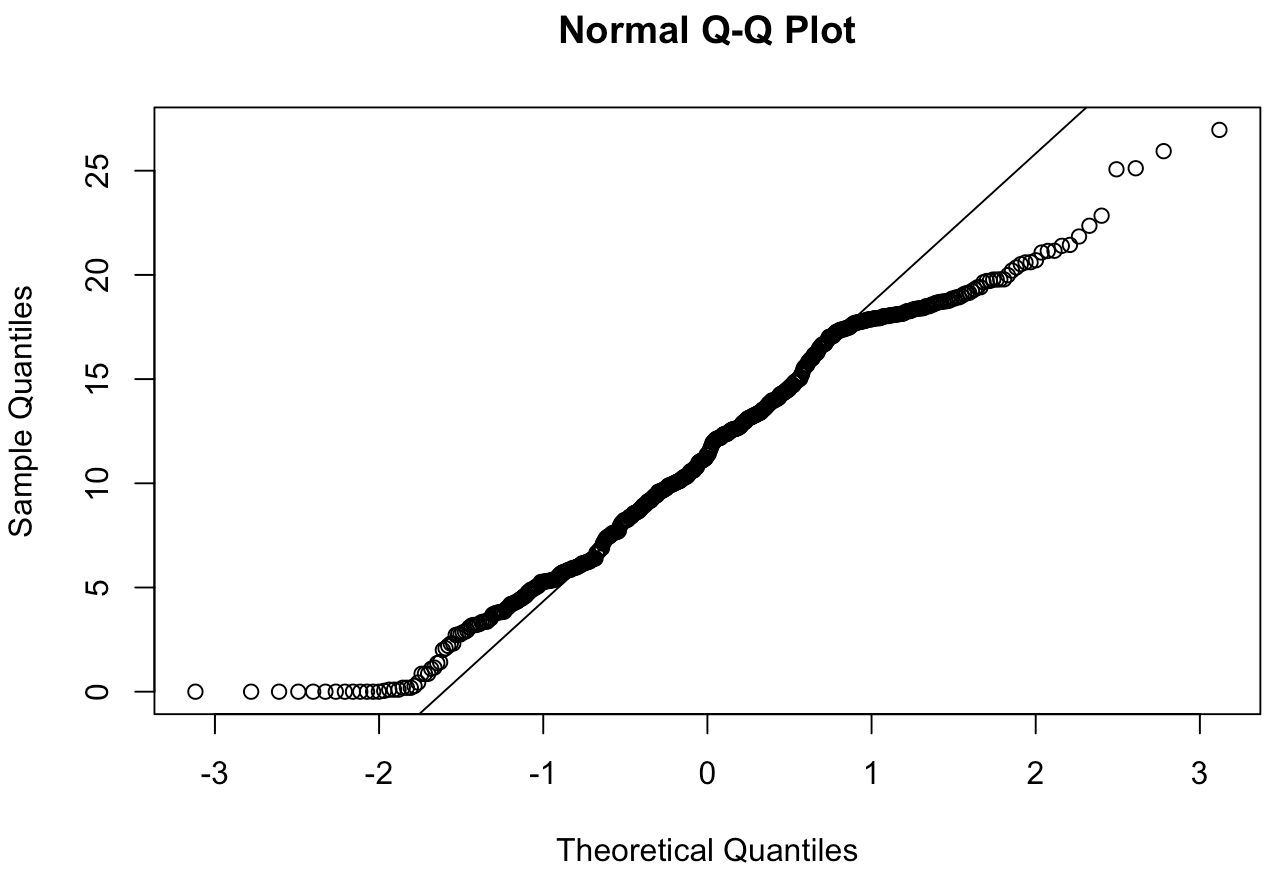
Transformacion exacta

La ecuacion para la transformacion exacta es $\frac{x^{\lambda}+1}{\lambda}$

```
transf_exacta = ((agua + 1)^l - 1)/l
hist(transf_exacta,freq=FALSE)
lines(density(transf_exacta), col = "red")
curve(dnorm(x, mean = mean(transf_exacta), sd = sd(transf_exacta)), from = min(transf_ex
acta), to = max(transf_exacta), add = TRUE, col = "blue", lwd = 2)
```




```
qqnorm(transf_exacta)
qqline(transf_exacta)
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0
5
10
15
20
25
-3
-2
-1
0
1
2
3

Resultados

```
library(nortest)
library(moments)
D0=ad.test(agua)
D1=ad.test(transf_simple)
D2=ad.test(transf_exacta)

P0=jarque.test(agua)
P1=jarque.test(transf_simple)
P2=jarque.test(transf_exacta)

m0=round(c(as.numeric(summary(agua)),kurtosis(agua),skewness(agua),D0$p.value, P0$p.valu
e),3)
m1=round(c(as.numeric(summary(transf_simple)),kurtosis(transf_simple),skewness(transf_si
mple),D1$p.value, P1$p.value),3)
m2=round(c(as.numeric(summary(transf_exacta)),kurtosis(transf_exacta),skewness(transf_ex
acta),D2$p.value, P2$p.value),3)

m<-as.data.frame(rbind(m0,m1,m2))
row.names(m)=c("Original","Primer modelo","Segundo Modelo")
names(m)=c("Minimo","Q1","Mediana","Media","Q3","Máximo","Curtosis","Sesgo","Valor p A
D", "Valor p JB")

m
```

```
##                Minimo     Q1 Mediana   Media      Q3  Máximo Curtosis  Sesgo
## Original            0 25.900  76.700 101.659 169.050 535.800    4.411  1.084
## Primer modelo       0  5.089   8.758   8.952  13.002  23.147    2.292  0.091
## Segundo Modelo      0  6.668  11.375  11.319  16.330  26.958    2.221 -0.079
##                Valor p AD Valor p JB
## Original                0      0.000
## Primer modelo           0      0.002
## Segundo Modelo          0      0.001
```

En el caso de esta variable, no hay anomalías, los ceros y valores atípicos corresponden a comidas que son válidas dentro del conjunto de datos.

Los valores de la media y mediana de ambas transformaciones mejoran considerablemente respecto a los valores originales. Los valores de sesgo y curtosis también se acercan a los valores de una distribución normal, sin embargo, de acuerdo con ambas pruebas de normalidad no se alcanza una distribución normal.

El mejor modelo seria la transformación simple debido a que obtiene un valor p mayor en la prueba de Jarque Bera.