Actividad Integradora 1

Oskar Arturo Gamboa Reyes

2024-08-20

Punto 1.

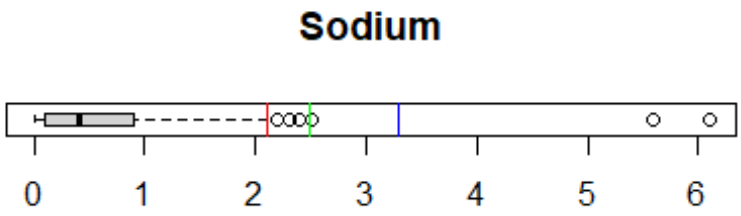
Datos Atípicos

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

  sod = M$Sodium
```

Diagrama de caja y bigote

```
q3=quantile(sod, 0.75)
ri = IQR(sod)
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(sod,horizontal=TRUE, main="Sodium")

abline(v=q3+1.5*ri, col="red")  #linea vertical en el límite de los datos
atípicos o extremos
abline(v=mean(sod)+3*sd(sod), col="green")
abline(v=q3+3*ri, col="blue")

cat("Rojo = 1.5 Rangos intercuartílicos,", "Verde = 3 Desviaciones estandar",
"Azul = 3 Rangos intercuartílicos")
```

```
## Rojo = 1.5 Rangos intercuartílicos, Verde = 3 Desviaciones estandar Azul =
3 Rangos intercuartílicos
```



Datos principales

```
summary(sod)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1000  0.4000  0.5732  0.9000  6.1000
```

```
cat("Rango intercuartílico =", IQR(sod))
```

```
## Rango intercuartílico = 0.8
```

Análisis de datos átipicos

Arriba de 1.5 rangos intercuartilicos podemos encontrar 6 datos, 3 desviaciones estandar tiene 3 datos y 3 rangos intercuartilicos tiene solamente 2 datos. Lo que puedo notar es que los datos más extremos son mariscos, los que suelen tener una alta cantidad de sodio. A parte de estos dos casos extremos los datos atípicos son sopas que suelen tener concentraciones altas de sodio. También leyendo los datos que tienen 0 sodio creo que puedo notar que son datos erroneos ya que son alimentos que en realidad tienen alguna cantidad de sodio.

Normalidad

Pruebas de normalidad

```
library(nortest)
library(moments)
ad.test(sod)
```

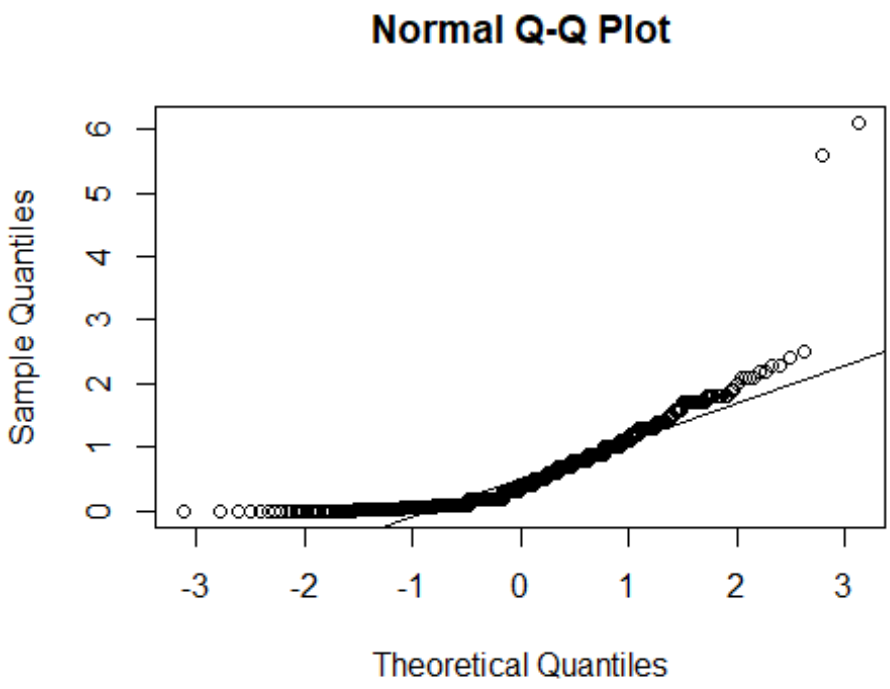
```
##
##  Anderson-Darling normality test
##
## data:  sod
## A = 24.827, p-value < 2.2e-16
```

```
jarque.test(sod)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  sod
## JB = 6834.2, p-value < 2.2e-16
## alternative hypothesis: greater
```

Gráficas

```
qqnorm(sod)
qqline(sod)
```



Curtosis y sesgo

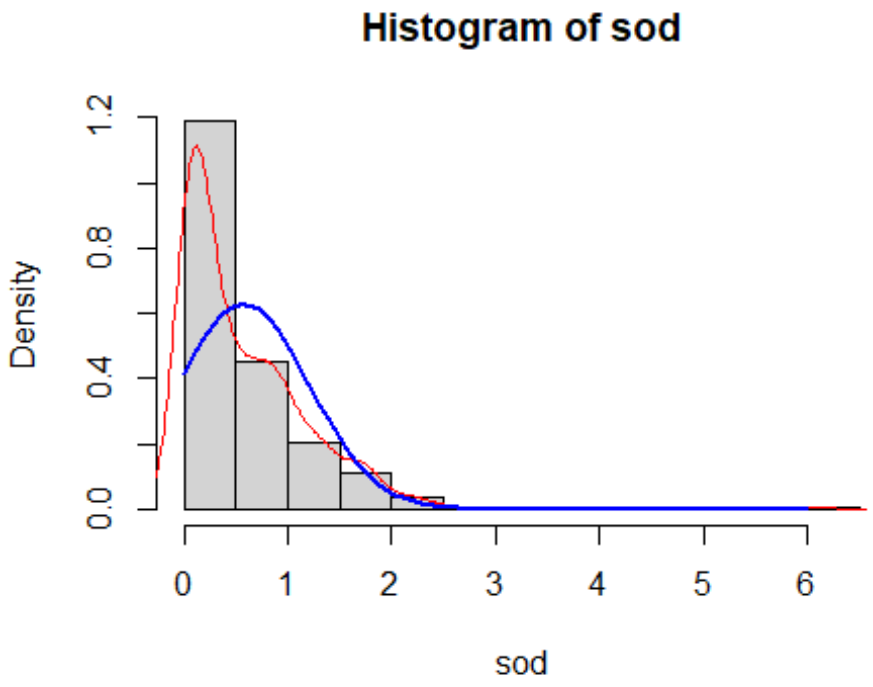
```
library(e1071)
```

```
##
## Attaching package: 'e1071'
```

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

print("Skewness Sodium")

## [1] "Skewness Sodium"

skewness(sod)

## [1] 2.728554

print("Kurtosis Sodium")

## [1] "Kurtosis Sodium"

kurtosis(sod)

## [1] 16.29239
```

Gráfico de densidad empírica

```
hist(sod,freq=FALSE)
lines(density(sod),col="red")
curve(dnorm(x,mean=mean(sod),sd=sd(sod)), from=0, to=6, add=TRUE,
col="blue",lwd=2)
```



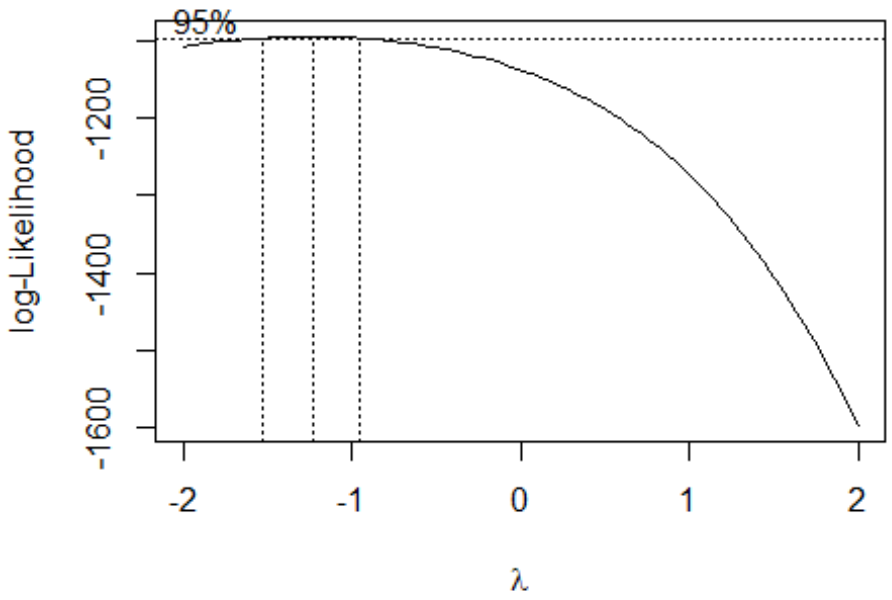
Análisis de datos normalidad

Esta variable no sigue una curva normal, esto lo podemos ver de varias maneras, primero, las pruebas de normalidad indican un p-value muy bajo, el qqplot podemos ver que no se ajusta a la linea normal, tiene muchos datos que se desvian en las puntas, además el sesgo a la derecha es demasiado alto (por lo que podemos ver en el histograma y en el cálculo de sesgo) y finalmente la curtosis es demasiada alta lo que indica una curva demasiado concentrada en un rango de datos (en este caso de 0-0.5).

Transformación a la normalidad

Transformación inicial

```
library(MASS)
bc<-boxcox((sod+1)~1)
```




```
l=bc$x[which.max(bc$y)]
print(paste("lambda: ", l))

## [1] "lambda:  -1.23232323232323"
```

Modelos sugeridos

Aproximado = $\frac{1}{x}$ Exacto = $\frac{(x+1)^{-1.23}-1}{-1.23}$

```
sodAprox = 1/(sod+1)
sodExacto = (((sod+1)^l)-1)/l
```

Comparación de variables

```
print("Datos Original")
```

```
## [1] "Datos Original"
```

```
print(summary(sod))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1000  0.4000  0.5732  0.9000  6.1000
```

```
print("Skewness")
```

```
## [1] "Skewness"
```

```
skewness(sod)
```

```
## [1] 2.728554
```

```
print("Kurtosis")
```

```
## [1] "Kurtosis"
```

```
kurtosis(sod)
```

```
## [1] 16.29239
```

```
print("Datos Aproximada")
```

```
## [1] "Datos Aproximada"
```

```
print(summary(sodAprox))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1408  0.5263  0.7143  0.7120  0.9091  1.0000
```

```
print("Skewness")
```

```
## [1] "Skewness"
```

```
skewness(sodAprox)
```

```
## [1] -0.2667636
```

```
print("Kurtosis")
```

```
## [1] "Kurtosis"
```

```
kurtosis(sodAprox)
```

```
## [1] -1.190872
```

```
print("Datos Exacta")
```

```
## [1] "Datos Exacta"
```

```
print(summary(sodExacto))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.08993 0.27543 0.27054 0.44355 0.73899
```

```
print("Skewness")
```

```
## [1] "Skewness"
```

```
skewness(sodExacto)
```

```
## [1] 0.1822511
```

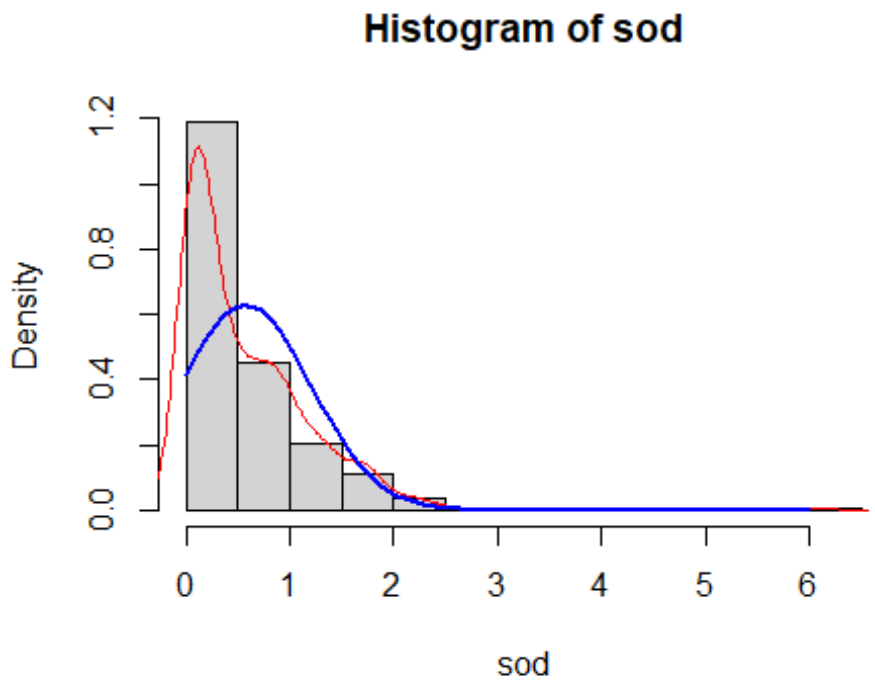
```
print("Kurtosis")
```

```
## [1] "Kurtosis"
```

```
kurtosis(sodExacto)
```

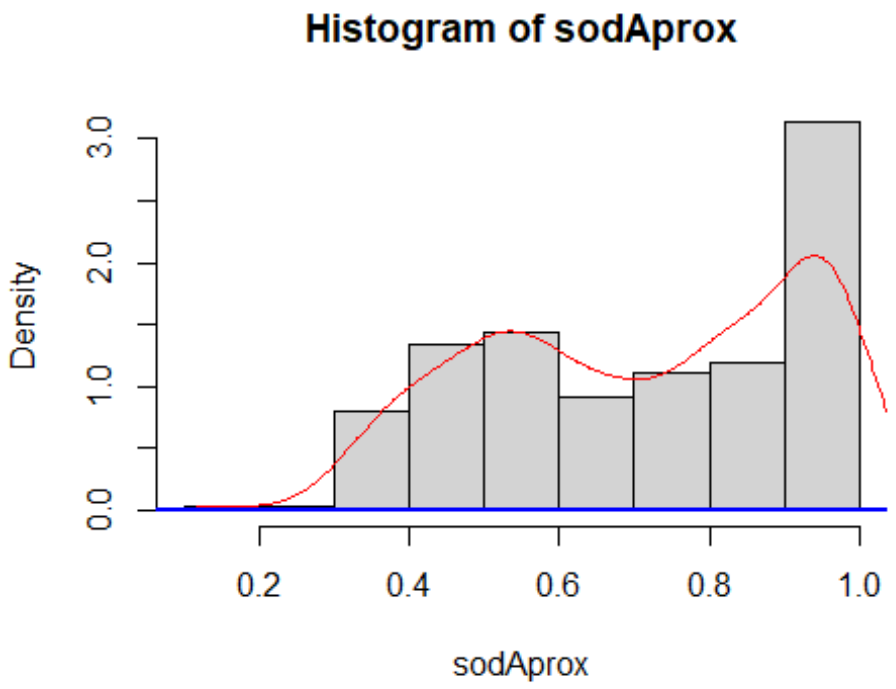
```
## [1] -1.29244
```

```
hist(sod,freq=FALSE)
lines(density(sod),col="red")
curve(dnorm(x,mean=mean(sod),sd=sd(sod)), from=0, to=6, add=TRUE,
col="blue",lwd=2)
```

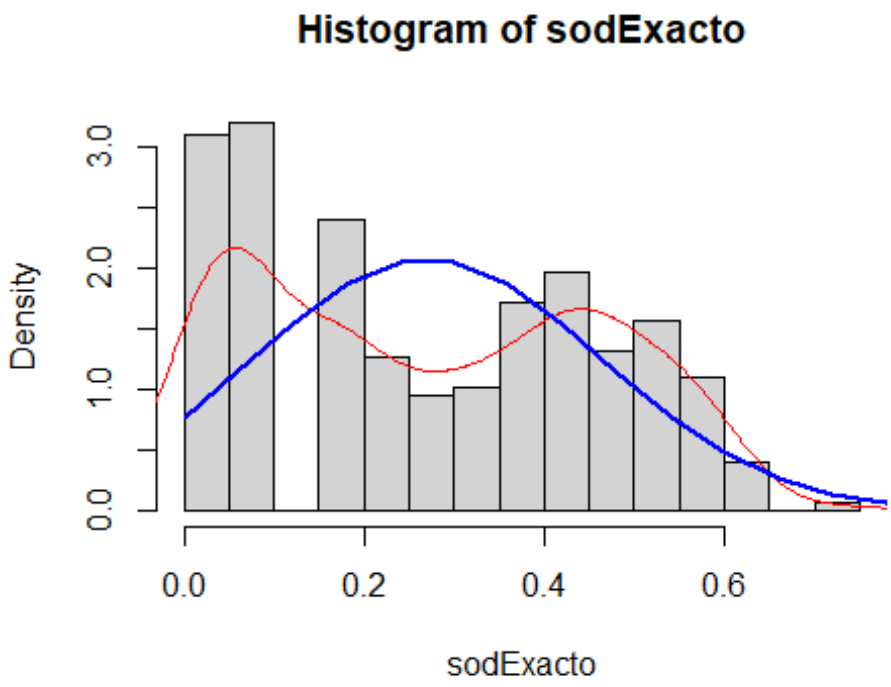



```
hist(sodAprox,freq=FALSE)
lines(density(sodAprox),col="red")
```

```r
curve(dnorm(x,mean=mean(sodAprox),sd=sd(sodAprox)), from=0, to=1000,
add=TRUE, col="blue",lwd=2)
```




```r
hist(sodExacto,freq=FALSE)
lines(density(sodExacto),col="red")
curve(dnorm(x,mean=mean(sodExacto),sd=sd(sodExacto)), from=0, to=6, add=TRUE,
col="blue",lwd=2)
```




```
print("Original")
```

```
## [1] "Original"
```

```
ad.test(sod)
```

```
##
##  Anderson-Darling normality test
##
## data:  sod
## A = 24.827, p-value < 2.2e-16
```

```
jarque.test(sod)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  sod
## JB = 6834.2, p-value < 2.2e-16
## alternative hypothesis: greater
```

```
print("Aproximado")
```

```
## [1] "Aproximado"
```

```
ad.test(sodAprox)
```

```
##
##  Anderson-Darling normality test
```

```
##
## data:  sodAprox
## A = 12.714, p-value < 2.2e-16
```

```
jarque.test(sodAprox)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  sodAprox
## JB = 38.771, p-value = 3.811e-09
## alternative hypothesis: greater
```

```
print("Exacto")
```

```
## [1] "Exacto"
```

```
ad.test(sodExacto)
```

```
##
##  Anderson-Darling normality test
##
## data:  sodExacto
## A = 12.799, p-value < 2.2e-16
```

```
jarque.test(sodExacto)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  sodExacto
## JB = 41.049, p-value = 1.22e-09
## alternative hypothesis: greater
```

Detectando anomalías

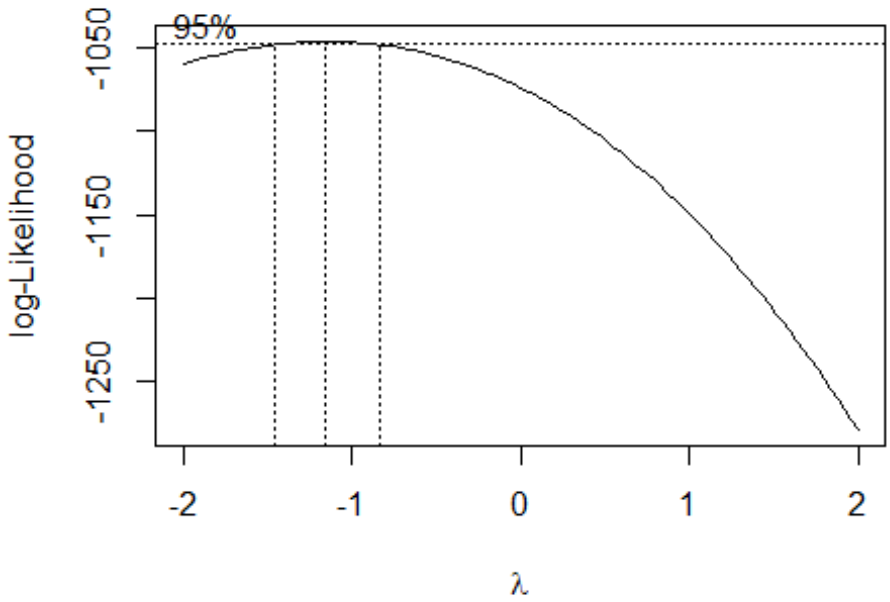
Como comenté en el análisis anterior de datos atítpicos, las comidas que tienen 0 son datos erroneos, ya que no muestra la cantidad de sodio que realmente tienen y los datos extremos son comidas del mar que suelen tener una cantidad enorme de sodio, como esto no define al menú entero, así que las voy a eliminar.

```
sod0=subset(sod,sod>0 & sod<3)
```

Resultados transformación final

Modelos sugeridos

```
library(MASS)
bc<-boxcox((sod0+1)~1)
```




```
l=bc$x[which.max(bc$y)]
print(paste("lambda: ", l))

## [1] "lambda:  -1.15151515151515"
```

$$\text{Aproximado} = \frac{1}{x} \quad \text{Exacto} = \frac{(x+1)^{-1.15}-1}{-1.15}$$

```
sod0Aprox = 1/(sod0+1)
sod0Exacto = (((sod0+1)^l)-1)/l
```

Comparaciones

```
print("Datos Original")

## [1] "Datos Original"

print(summary(sod0))

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0010  0.1000  0.4000  0.5685  0.9000  2.5000

print("Skewness")

## [1] "Skewness"

skewness(sod0)

## [1] 1.073652
```

```r
print("Kurtosis")
```

```
## [1] "Kurtosis"
```

```r
kurtosis(sod0)
```

```
## [1] 0.4346818
```

```r
print("Datos Aproximada")
```

```
## [1] "Datos Aproximada"
```

```r
print(summary(sod0Aprox))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2857  0.5263  0.7143  0.7066  0.9091  0.9990
```

```r
print("Skewness")
```

```
## [1] "Skewness"
```

```r
skewness(sod0Aprox)
```

```
## [1] -0.2102716
```

```r
print("Kurtosis")
```

```
## [1] "Kurtosis"
```

```r
kurtosis(sod0Aprox)
```

```
## [1] -1.289636
```

```r
print("Datos Exacta")
```

```
## [1] "Datos Exacta"
```

```r
print(summary(sod0Exacto))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0009989 0.0902662 0.2789511 0.2817655 0.4537141 0.6631972
```

```r
print("Skewness")
```

```
## [1] "Skewness"
```

```r
skewness(sod0Exacto)
```

```
## [1] 0.1621736
```

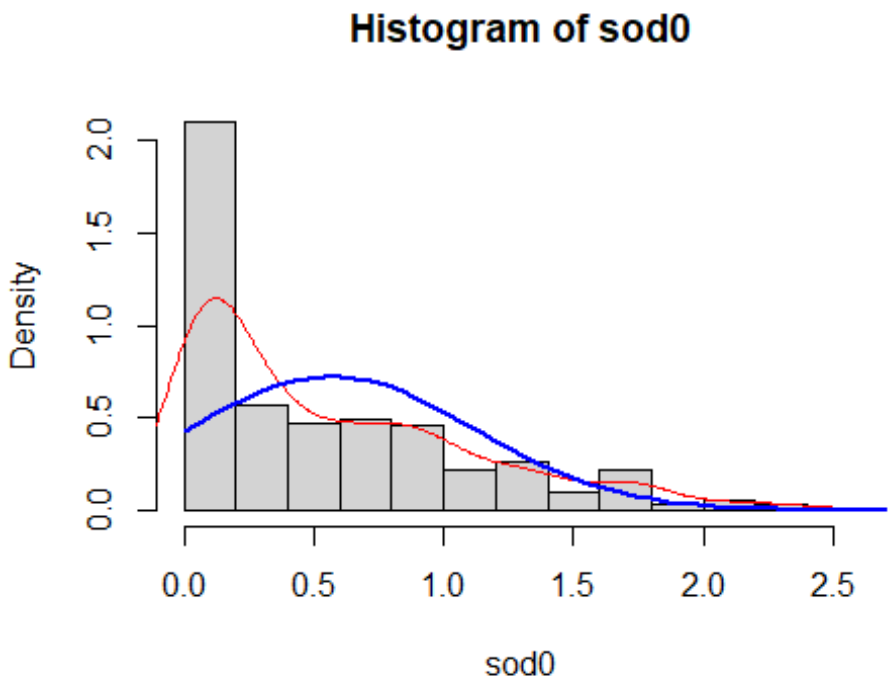
```r
print("Kurtosis")
```

```
## [1] "Kurtosis"
```

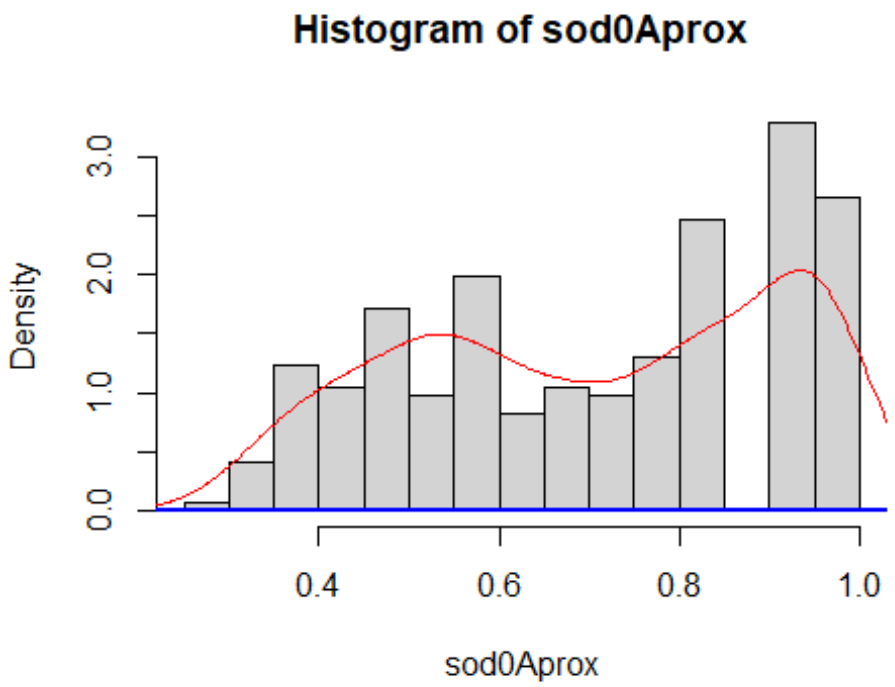
```r
kurtosis(sod0Exacto)
```

```
## [1] -1.328711
```

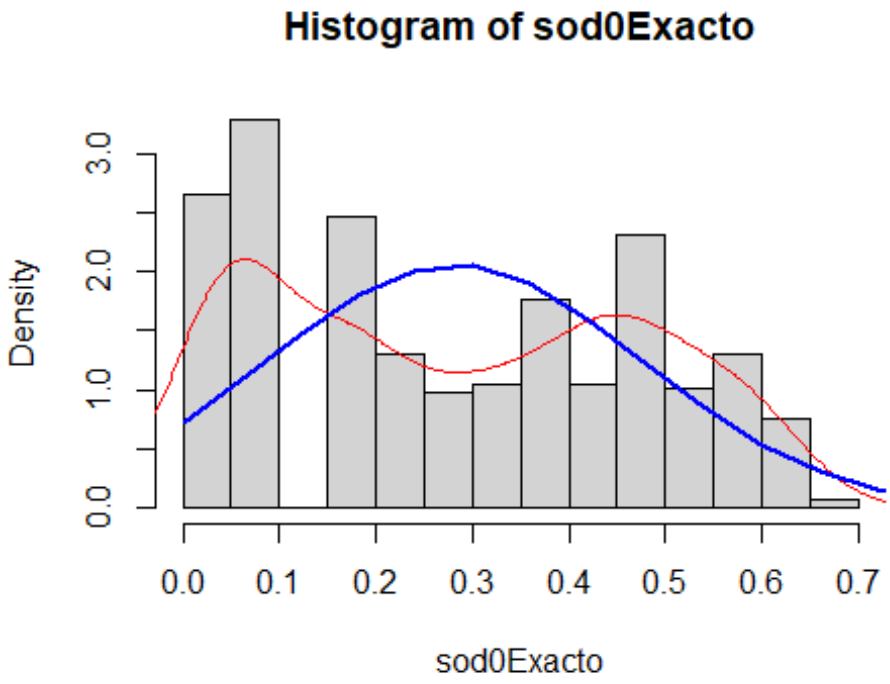
```
hist(sod0,freq=FALSE)
lines(density(sod0),col="red")
curve(dnorm(x,mean=mean(sod0),sd=sd(sod0)), from=0, to=6, add=TRUE,
col="blue",lwd=2)
```




```
hist(sod0Aprox,freq=FALSE)
lines(density(sod0Aprox),col="red")
curve(dnorm(x,mean=mean(sod0Aprox),sd=sd(sod0Aprox)), from=0, to=1000,
add=TRUE, col="blue",lwd=2)
```




```
hist(sod0Exacto,freq=FALSE)
lines(density(sod0Exacto),col="red")
curve(dnorm(x,mean=mean(sod0Exacto),sd=sd(sod0Exacto)), from=0, to=6,
add=TRUE, col="blue",lwd=2)
```




```
print("Original")
```

```
## [1] "Original"
```

```
ad.test(sod0)
```

```
##
##  Anderson-Darling normality test
##
## data:  sod0
## A = 22.223, p-value < 2.2e-16
```

```
jarque.test(sod0)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  sod0
## JB = 107.83, p-value < 2.2e-16
## alternative hypothesis: greater
```

```
print("Aproximado")
```

```
## [1] "Aproximado"
```

```
ad.test(sod0Aprox)
```

```
##
##  Anderson-Darling normality test
```

```
##
## data:  sod0Aprox
## A = 12.291, p-value < 2.2e-16
```

```
jarque.test(sod0Aprox)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  sod0Aprox
## JB = 40.671, p-value = 1.473e-09
## alternative hypothesis: greater
```

```
print("Exacto")
```

```
## [1] "Exacto"
```

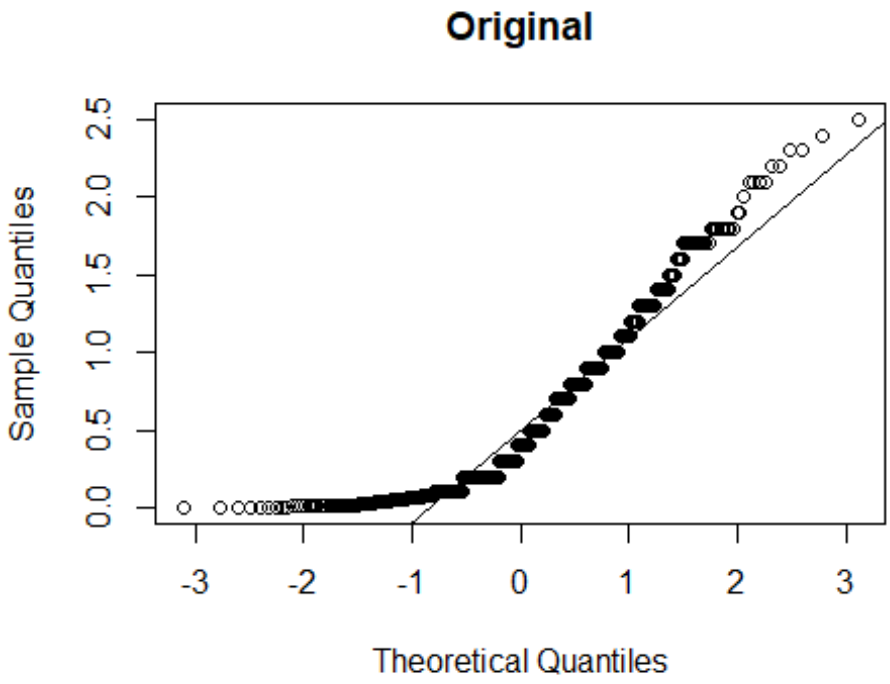
```
ad.test(sod0Exacto)
```

```
##
##  Anderson-Darling normality test
##
## data:  sod0Exacto
## A = 12.335, p-value < 2.2e-16
```

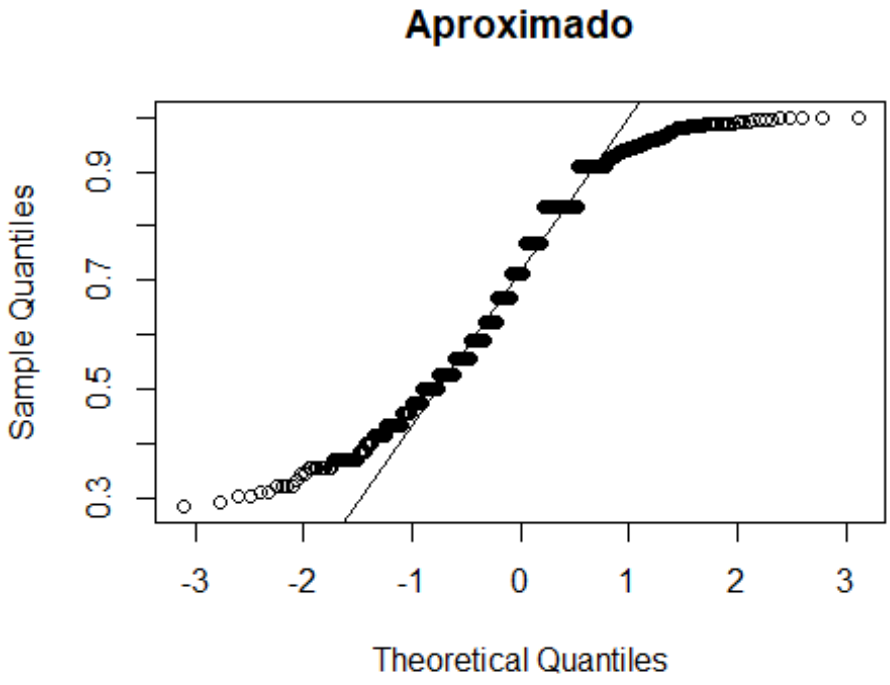
```
jarque.test(sod0Exacto)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  sod0Exacto
## JB = 41.343, p-value = 1.053e-09
## alternative hypothesis: greater
```

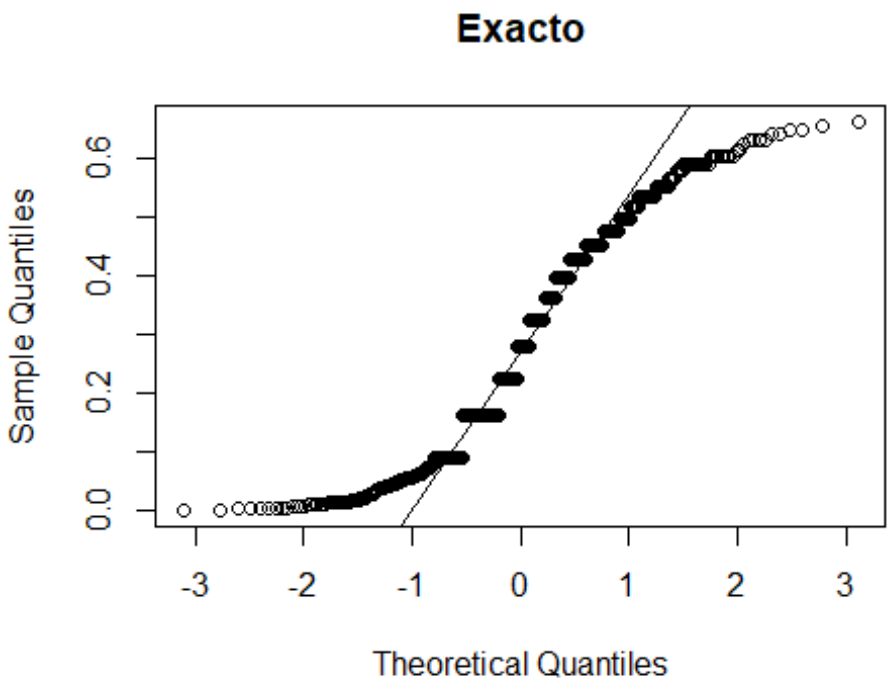
```
qqnorm(sod0, main = "Original")
qqline(sod0)
```




```r
qqnorm(sod0Aprox, main = "Aproximado")
qqline(sod0Aprox)
```



```r
qqnorm(sod0Exacto, main = "Exacto")
qqline(sod0Exacto)
```



Conclusiones

Ningúno de los métdodos logró hacer una curva normal, ningún test resultó en un p-value suficientemente grande. Podemos ver que la variable tiene demasiados datos atípicos, esto se nota perfectamente en el qqplot, las colas de la gráfica se desvian mucho de la linea normal. tiene demasiados datos menores a 0.1, mientras que también tienen datos mas grandes que no definen el menú. Aunque no logramos los resultados esperados, pudimos mejorar el sesgo y la curtosis, logrando resultados que describirian una curva normal (sesgo menor a 0.5 y curtosis más cercana a 3).