act_integradora

Rogelio Lizárraga

2024-08-20

Análisis descriptivo de la variable Sodium

```
M=read.csv("food_data_g.csv")
head(M$Sodium)
```

```
## [1] 0.016 0.300 0.000 0.017 0.046 0.100
```

```
summary(M$Sodium)
```

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

```
IQR(M$Sodium)
```

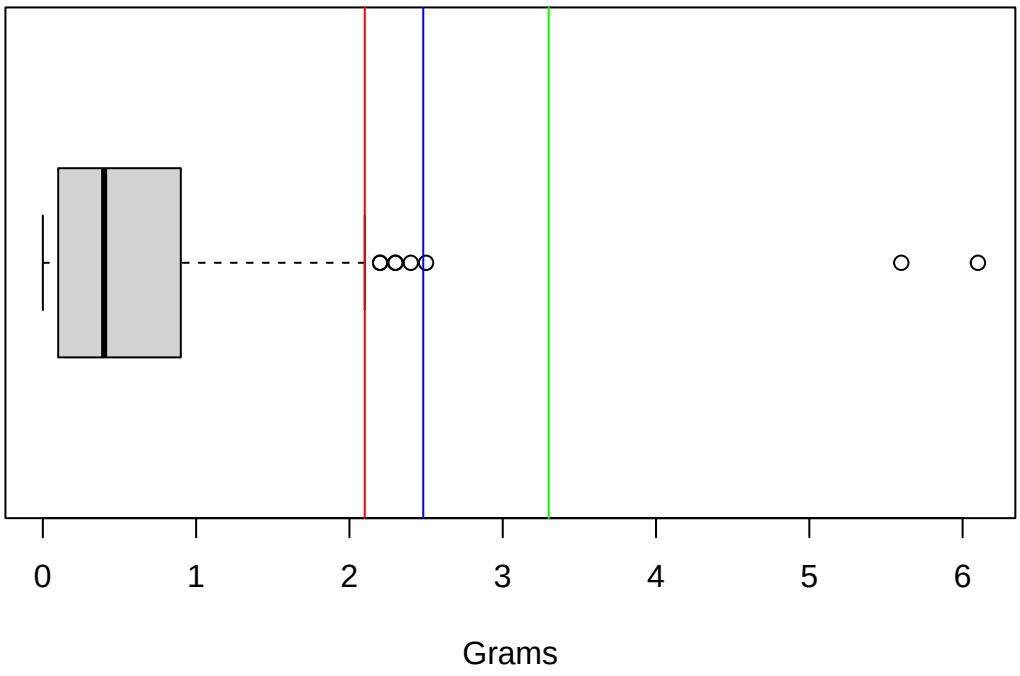
```
## [1] 0.8
```

Análisis de datos atípicos y normalidad de Sodium

Datos atípicos

```
q1 = quantile(M$Sodium, 0.25)  #Cuantil 1 de la variable X
q3 = quantile(M$Sodium, 0.75)
ri= IQR(M$Sodium)  #Rango intercuartílico de X


g_caja = boxplot(M$Sodium,horizontal=TRUE,ylim=c(min(M$Sodium), max(M$Sodium)), main = "", xlab = "Grams
abline(v=q3+1.5*ri,col="red")
abline(v= mean(M$Sodium) + 3*sd(M$Sodium),col="blue")
abline(v=q3+3*ri,col="green")
```



Eliminamos datos atípicos

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
M2 <- M %>%
filter (M$Sodium!=0)
head(M2)
```

```
##   X Unnamed..0                 food Caloric.Value  Fat Saturated.Fats
## 1 0          0         cream cheese            51  5.0            2.9
## 2 1          1    neufchatel cheese           215 19.4           10.9
## 3 3          3       ricotta cheese            30  2.0            1.3
## 4 4          4 cream cheese low fat            30  2.3            1.4
## 5 5          5 cream cheese fat free           19  0.2            0.1
```

```
## 6 6          6        gruyere cheese           116  9.1           5.3
##   Monounsaturated.Fats Polyunsaturated.Fats Carbohydrates Sugars Protein
## 1                1.300                0.200           0.8  0.500     0.9
## 2                4.900                0.800           3.1  2.700     7.8
## 3                0.500                0.002           1.5  0.091     1.5
## 4                0.600                0.042           1.2  0.900     1.2
## 5                0.091                0.075           1.4  1.000     2.8
## 6                2.800                0.500           0.1  0.100     8.3
##   Dietary.Fiber Cholesterol Sodium Water Vitamin.A Vitamin.B1 Vitamin.B11
## 1             0        14.6  0.016   7.6     0.200      0.033       0.064
## 2             0        62.9  0.300  53.6     0.200      0.099       0.079
## 3             0         9.8  0.017  14.7     0.075      0.019       0.079
## 4             0         8.1  0.046  10.0     0.016      0.080       0.062
## 5             0         2.2  0.100  12.9     0.063      0.020       0.089
## 6             0        30.8  0.200   9.3     0.061      0.021       0.072
##   Vitamin.B12 Vitamin.B2 Vitamin.B3 Vitamin.B5 Vitamin.B6 Vitamin.C Vitamin.D
## 1       0.092      0.097      0.084      0.052      0.096     0.004     0.000
## 2       0.090      0.100      0.200      0.500      0.078     0.000     0.000
## 3       0.091      0.027      0.041      0.016      0.007     0.006     0.000
## 4       0.049      0.026      0.080      0.100      0.003     0.000     0.036
## 5       0.092      0.021      0.025      0.200      0.038     0.000     0.000
## 6       0.078      0.004      0.043      0.200      0.051     0.000     0.034
##   Vitamin.E Vitamin.K Calcium Copper  Iron Magnesium Manganese Phosphorus
## 1     0.000     0.100   0.008 14.100 0.082     0.027     1.300      0.091
## 2     0.300     0.045  99.500  0.034 0.100     8.500     0.088    117.300
## 3     0.001     0.011   0.097 41.200 0.097     0.096     4.000      0.024
## 4     0.009     0.019  22.200  0.072 0.008     1.200     0.098     22.800
## 5     0.049     0.059  63.200  0.039 0.053     4.000     0.028     94.100
## 6     0.035     0.048 283.100  0.033 0.094    10.100     0.002    169.400
##   Potassium Selenium  Zinc Nutrition.Density
## 1      15.5   19.100 0.039             7.070
## 2     129.2    0.054 0.700           130.100
## 3      30.8   43.800 0.035             5.196
## 4      37.1    0.034 0.053            27.007
## 5      50.0    0.013 0.300            67.679
## 6      22.7    0.079 1.100           300.694
```

```
Sodium = M2$Sodium
```

Normalidad

```
summary(M$Sodium)
```

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

Pruebas de hipótesis de Sodium H_0 = el conjunto de datos tiene una asimetría y una curtosis que coincide con una distribución normal. H_1 = el conjunto de datos tiene una asimetría y una curtosis que no coincide con una distribución normal

```r
library(moments)
skewness(M$Sodium)
```

```
## [1] 2.735999
```

```r
kurtosis(M$Sodium)
```

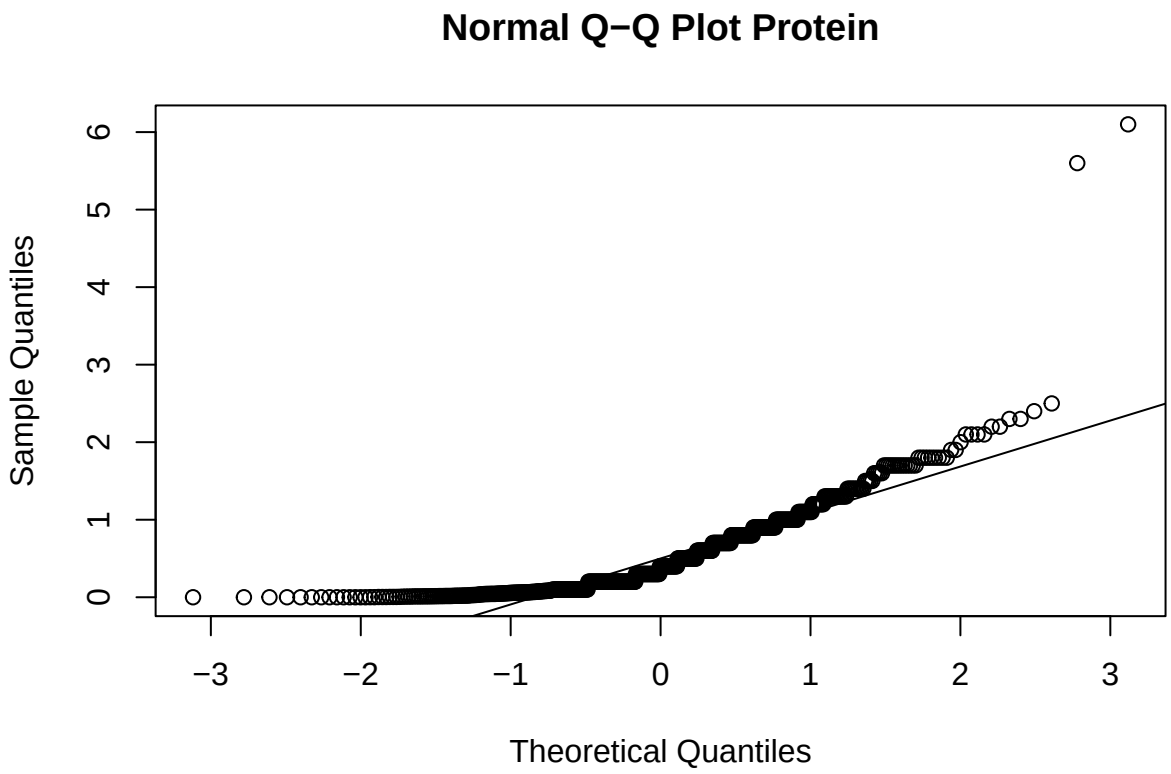
```
## [1] 19.3626
```

```r
jarque.test(M$Sodium)
```

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

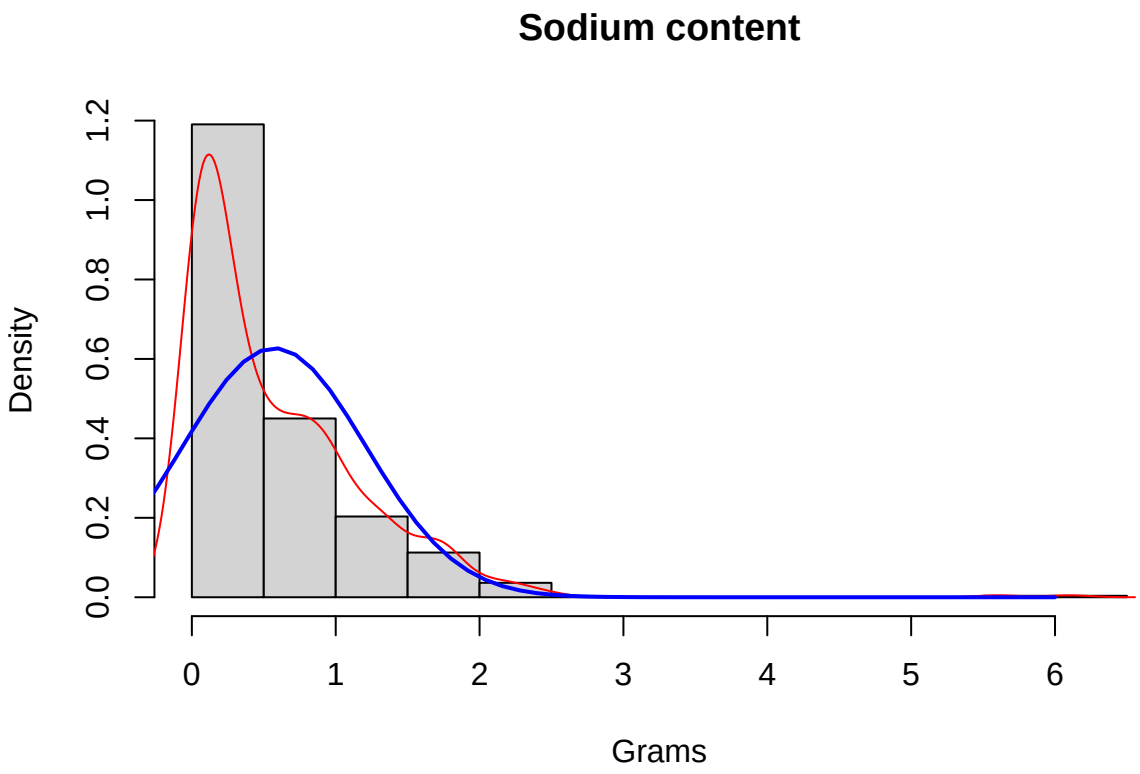
Como nuestro valor $p < 0.05$, se rechaza H_0, por lo que el conjunto de datos tiene una asimetría y una curtosis que NO coincide con una distribución normal. Además, observamos que tenemos un sesgo de 2.73 (bastante alto) y una curtosis de 19.36 (la cual es extremadamente elevada), por lo que estos NO coinciden con una distribución normal.

```r
qqnorm(M$Sodium, main = 'Normal Q-Q Plot Protein')
qqline(M$Sodium)
```



Histograma

```
hist(M$Sodium,freq=FALSE,  main = "Sodium content", xlab = "Grams")
lines(density(M$Sodium),col="red")
curve(dnorm(x,mean=mean(M$Sodium),sd=sd(M$Sodium)), from=-6, to=6, add=TRUE, col="blue",lwd=2)
```



Conclusión parte 1:

Como podemos observar en los datos de la variable **Sodium**, la mediana y la media están bastante lejos, así como el rango medio y la mediana, por lo que pareciera que los datos no se distribuyen de manera normal. Por otro lado, se observa en la gráfica de bigote:

Cota roja (1.5 rangos intercuartílicos) = Hay una cantidad significativa de datos que sobrepasan esta cota (seis datos). Cota azul (3 σ) = Hay una cantidad de datos que sobrepasan esta cota (tres datos). Cota verde (3 rangos intercuartílicos) = Hay datos extremos que sobrepasan esta cota (dos datos).

Sin embargo, no eliminaremos ninguno de estos datos (los que sobrepasan las cota), pues son parte de la población del menú de alimentos. Es decir, son datos correctos, pues es común que, el kung pao, por ejemplo, tenga cantidades altas de sodio.

Por otro lado, eliminamos los ceros, pues no es posible que el queso, por ejemplo, no contenga sodio, cuando es parte de su composición, o un sandwich de chick-fil-a. Este nuevo dataset corregido se utilizará en las transformaciones.

Observando los gráficos QQ, nos damos cuenta que cuentan con colas muy largas, por lo que no se distribuye como una normal. Además, en el histograma se observa que los datos tienen un sesgo hacia la derecha y una curtosis demasiada alta, pero la mayoría de los datos deberían de estar en $(\mu \pm \sigma)$, por lo que no se distribuyen como una normal.

Finalmente, vemos que se rechaza H_0, al ser el valor p demasiado pequeño, por lo que el conjunto de datos tiene una asimetría y una curtosis que NO coincide con una distribución normal. Además, observamos que el sesgo y la curtosis son demasiado altos, por lo que estos NO coinciden con una distribución normal.

Transformación a normalidad con los datos ya corregidos

Box-Cox para encontrar lambda

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

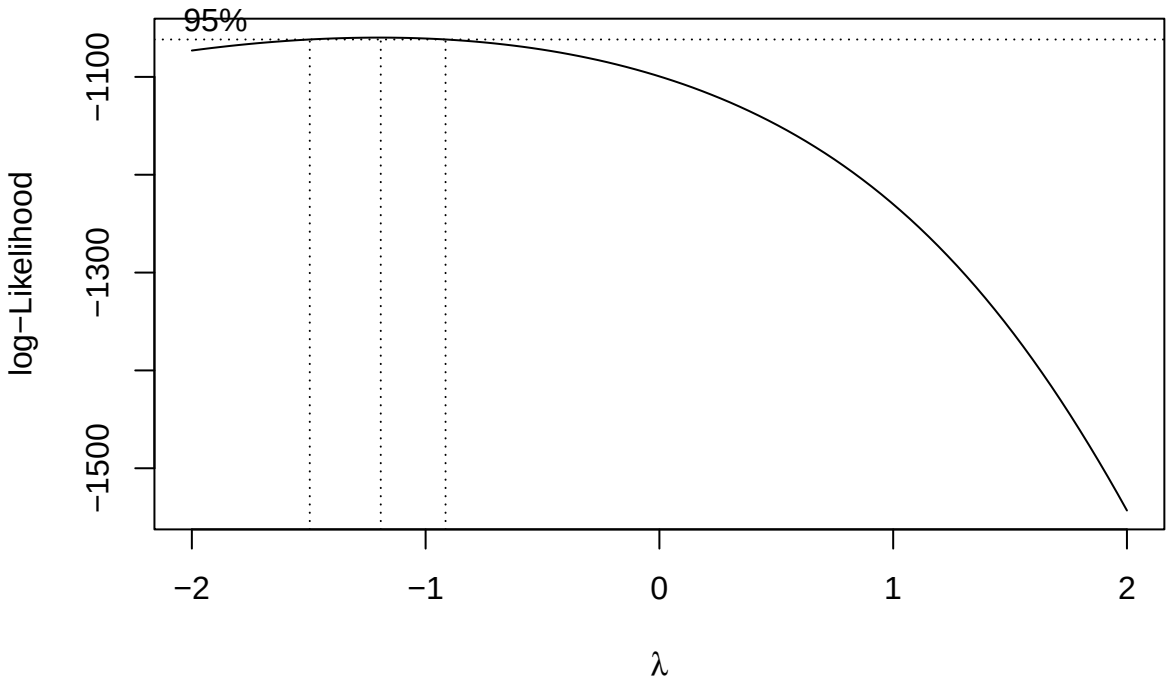
```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(VGAM)
```

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

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

```
exacto<-boxcox((Sodium + 1)~1)
```



```r
l_exacto =exacto$x[which.max(exacto$y)]

l_exacto
```

```
## [1] -1.191919
```

```r
aproximado = 1/(Sodium + 1)
sodium_exacto = ((Sodium + 1)^l_exacto-1)/l_exacto
```

Modelo aproximado =

$$\frac{1}{x+1}$$

Modelo exacto =

$$\frac{(x+1)^{-1.19}-1}{-1.19}$$

Antes que nada, veremos si es normal para ver si es buena idea utilizar este modelo: H_0 = el conjunto de datos se distribuye de manera normal. H_1 = el conjunto de datos NO se distribuye de manera normal.

```r
library(nortest)
cox_1 = ad.test(aproximado)$p.value
cox_2 = ad.test(sodium_exacto)$p.value
cox_3 = ad.test(M$Sodium)$p.value
cat('Valor p modelo aproximado con Anderson-Darling:', cox_1, '\n')
```

```
## Valor p modelo aproximado con Anderson-Darling: 3.7e-24
```

```r
cat('Valor p modelo Box-Cox exacto con Anderson-Darling:', cox_2, '\n')
```

```
## Valor p modelo Box-Cox exacto con Anderson-Darling: 3.7e-24
```

```r
cat('Valor p datos originales con Anderson-Darling:', cox_3, '\n')
```

```
## Valor p datos originales con Anderson-Darling: 3.7e-24
```

```r
cox_4 = jarque.test(aproximado)$p.value
cox_5 = jarque.test(sodium_exacto)$p.value
cox_6 = jarque.test(M$Sodium)$p.value
cat('Valor p modelo Box-Cox aproximado con Jarque-Bera:', cox_4, '\n')
```

```
## Valor p modelo Box-Cox aproximado con Jarque-Bera: 1.066347e-08
```

```r
cat('Valor p modelo Box-Cox exacto con Jarque-Bera:', cox_5, '\n')
```

```
## Valor p modelo Box-Cox exacto con Jarque-Bera: 3.917205e-09
```

```r
cat('Valor p datos originales con Jarque-Bera:', cox_6, '\n')
```

```
## Valor p datos originales con Jarque-Bera: 0
```

Como todos los valores $p < 0.05$, los datos no se distribuyen de manera normal, por lo que no se planteará utilizar este modelo.

Yeo-Johnson para encontrar mejores resultados

```
# Sumamos unos para evitar divisiones entre cero y trabajar con valores positivos en el aproximado
sodium_1<- 1/(Sodium + 1)
sodium_2<- yeo.johnson(Sodium + 1, lambda = l_exacto)
```

Modelo aproximado =

$$\frac{1}{x+1}$$

Modelo exacto =

$$\frac{x^{-1.19}-1}{-1.19}$$

Normalidad y análisis con Yeo-Johnson

```
library(e1071)
```

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

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

```
cat('Datos originales \n')
```

```
## Datos originales
```

```
summary(M$Sodium)
```

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

```
print("Curtosis")
```

```
## [1] "Curtosis"
```

```
kurtosis(M$Sodium)
```

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

```
print("Sesgo")
```

```
## [1] "Sesgo"
```

```r
skewness(M$Sodium)
```

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

```r
cat("Rango medio:", IQR(M$Sodium))
```

```
## Rango medio: 0.8
```

```r
cat('\nModelo aproximado \n')
```

```
##
## Modelo aproximado
```

```r
summary(sodium_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1408  0.5263  0.7143  0.7045  0.9091  0.9990
```

```r
print("Curtosis")
```

```
## [1] "Curtosis"
```

```r
kurtosis(sodium_1)
```

```
## [1] -1.190013
```

```r
print("Sesgo")
```

```
## [1] "Sesgo"
```

```r
skewness(sodium_1)
```

```
## [1] -0.2446265
```

```r
cat("Rango medio:", IQR(sodium_1))
```

```
## Rango medio: 0.3827751
```

```r
cat('\nModelo exacto \n')
```

```
##
## Modelo exacto
```

```r
summary(sodium_2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4720  0.4925  0.5435  0.5529  0.6031  0.7697
```

```r
print("Curtosis")
```

```
## [1] "Curtosis"
```

```r
kurtosis(sodium_2)
```

```
## [1] -0.7478948
```

```r
print("Sesgo")
```

```
## [1] "Sesgo"
```

```r
skewness(sodium_2)
```

```
## [1] 0.5128746
```

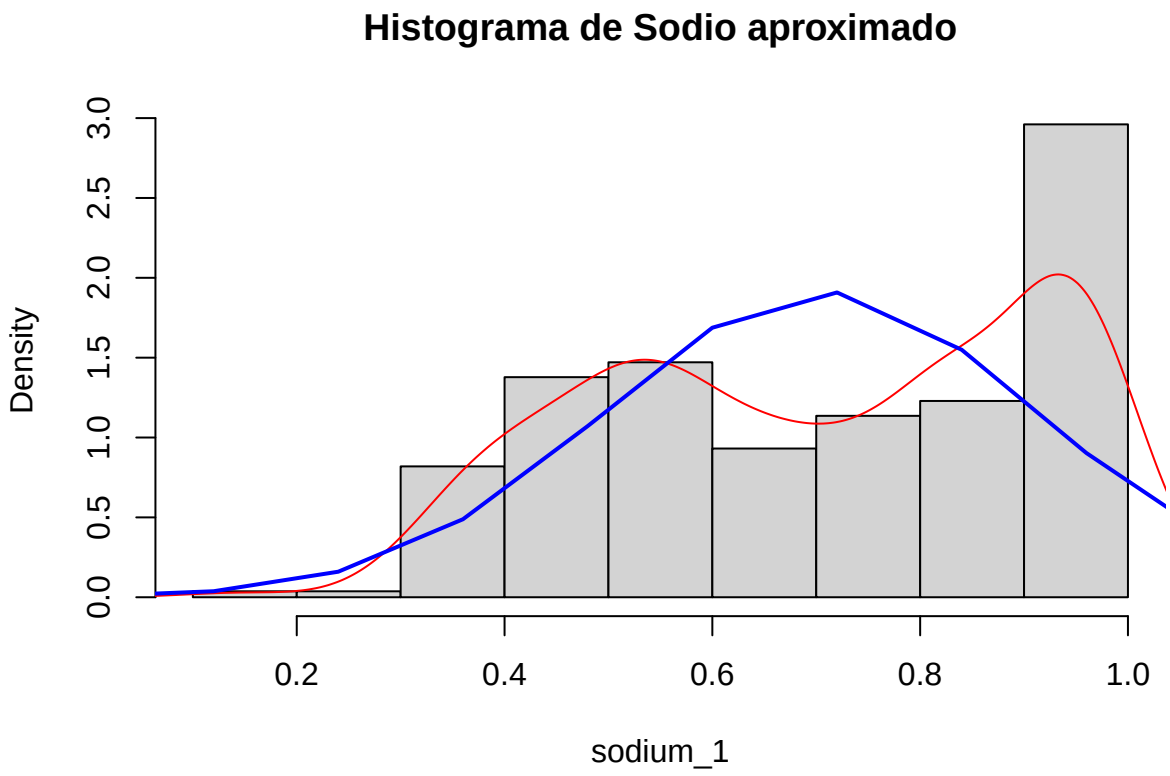
```r
cat("Rango medio:", IQR(sodium_2))
```

```
## Rango medio: 0.1106554
```

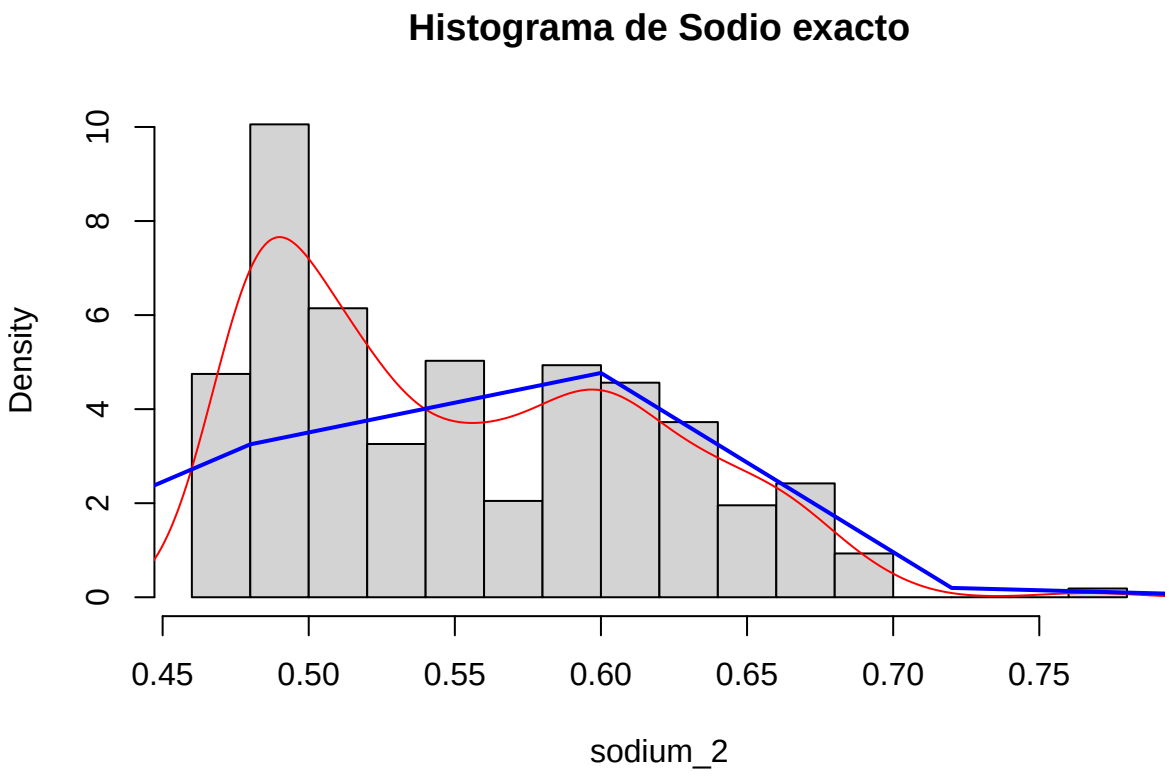
Para los datos originales, la media, la mediana y el rango medio están muy alejados, por lo que pareciera que no es normal. Se observa una curtosis muy elevada y un sesgo alto. Para el modelo aproximado, observamos que la media y la mediana están cercanas, pero el rango medio no. Observamos que la curtosis está muy elevada y el sesgo está ligeramente elevado, por lo que pareciera que no es normal.

Para el modelo exacto, observamos que la media, la mediana están cercanas, pero el rango medio está muy alejado. El máximo se encuentra relativamente alejado del tercer cuartil. Observamos que el sesgo es alto, pero la curtosis está más elevada, por lo que pareciera que no es normal.

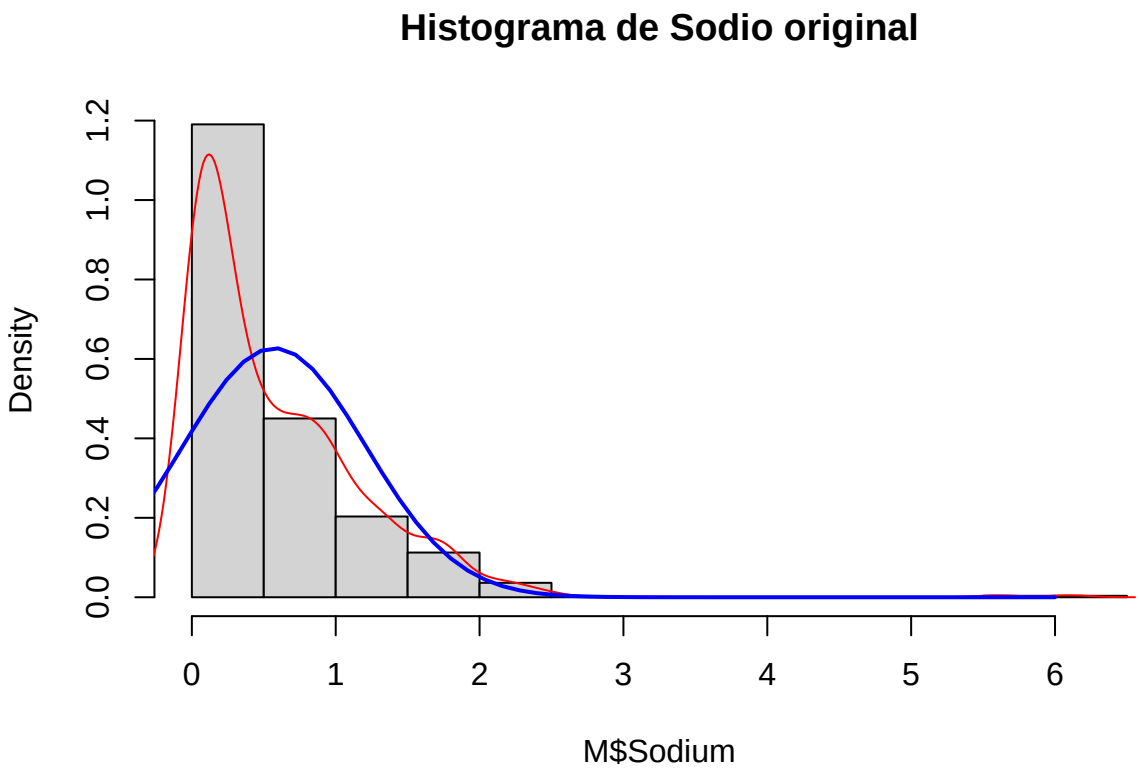
```r
hist(sodium_1,freq = FALSE,main="Histograma de Sodio aproximado")
lines(density(sodium_1),col="red")
curve(dnorm(x,mean=mean(sodium_1),sd=sd(sodium_1)), from=-6, to=6, add=TRUE, col="blue",lwd=2)
```




```
hist(sodium_2,freq =FALSE,main="Histograma de Sodio exacto")
lines(density(sodium_2),col="red")
curve(dnorm(x,mean=mean(sodium_2),sd=sd(sodium_2)), from=-6, to=6, add=TRUE, col="blue",lwd=2)
```




```
hist(M$Sodium,freq=FALSE,  main = "Histograma de Sodio original")
lines(density(M$Sodium),col="red")
curve(dnorm(x,mean=mean(M$Sodium),sd=sd(M$Sodium)), from=-6, to=6, add=TRUE, col="blue",lwd=2)
```



Observamos que el histograma de sodio aproximado tiene un sesgo a la izquierda con una curtosis demasiado alta, por lo que no es normal. El histograma se sodio con el modelo exacto tiene un sesgo a la derecha con una curtosis también elevada y vemos que los datos no se distribuyen de manera normal. Como se había dicho anteriormente, el histograma de sodio original no se distribuye de manera normal, pues tiene un sesgo a la derecha y una curtosis extremadamente alta.

H_0 = el conjunto de datos se distribuye de manera normal. H_1 = el conjunto de datos NO se distribuye de manera normal.

```
library(nortest)
yeo_1 = ad.test(sodium_1)$p.value
yeo_2 = ad.test(sodium_2)$p.value
yeo_3 = ad.test(M$Sodium)$p.value
cat('Valor p modelo Yeo aproximado con Anderson-Darling:', yeo_1, '\n')
```

```
## Valor p modelo Yeo aproximado con Anderson-Darling: 3.7e-24
```

```
cat('Valor p modelo Yeo exacto con Anderson-Darling:', yeo_2, '\n')
```

```
## Valor p modelo Yeo exacto con Anderson-Darling: 3.7e-24
```

```
cat('Valor p datos originales con Anderson-Darling:', yeo_3, '\n')
```

```
## Valor p datos originales con Anderson-Darling: 3.7e-24
```

```r
yeo_4 = jarque.test(sodium_1)$p.value
yeo_5 = jarque.test(sodium_2)$p.value
yeo_6 = jarque.test(M$Sodium)$p.value
cat('Valor p modelo Yeo aproximado con Jarque-Bera:', yeo_4, '\n')
```

```
## Valor p modelo Yeo aproximado con Jarque-Bera: 1.066347e-08
```

```r
cat('Valor p modelo Yeo xacto con Jarque-Bera:', yeo_5, '\n')
```

```
## Valor p modelo Yeo xacto con Jarque-Bera: 1.593437e-08
```

```r
cat('Valor p datos originales con Jarque-Bera:', yeo_6, '\n')
```

```
## Valor p datos originales con Jarque-Bera: 0
```

Como todos los valores $p < 0.05$, se rechaza H_0 para todos los conjuntos de datos, con las pruebas de Anderson-Darling y Jarque-Bera, por lo que estos NO se distribuyen de manera normal.

La razón por lo que los datos son atípicos es debido a las cantidades de sodio que se reportan, las cuales probablemente sean erróneas (en muchos datos). Para corregir esto, sería necesario verificar las cantidades de sodio de nuevo con la información nutrimental verificado. Sin embargo, si los datos reportados de sodio son correctos, no es raro que los datos no se distribuyan de una manera normal, pues es la población de la distribución, por lo que realmente ese es su comportamiento.

Conclusión parte 2:

En los modelos realizados, no se encontraron buenos resultados, pues ambos dieron distribuciones muy alejadas de la normalidad. Es decir, los conjuntos de datos no se distribuyeron de manera normal. Debido a lo anterior, realmente no hay un modelo que sobresalga ante los dos, pues los valores p fueron tan pequeños que no son de relevancia las transformaciones.

Sin embargo, si escogieramos uno, diría que la transformación de Yeo-Johnson es la mejor, pues hace transformaciones para una cantidad de datos negativos y positivos, por lo que es buena idea implementarlos. Además, no es tan complejo y es fácil de implementar.