Actividad Integradora

Erika Martínez Meneses

2024-08-20

Lectura de Datos

```
file.choose()
```

```
## [1] "C:\\Users\\erika\\Documents\\Agos-
Dic2024\\Estadística\\food_data_g.csv"
```

```
library(readr)
data <- read_csv("C:\\Users\\erika\\Documents\\Agos-
Dic2024\\Estadística\\food_data_g.csv")
```

```
## New names:
## Rows: 551 Columns: 37
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter:
"," chr
## (1): food dbl (36): ...1, Unnamed: 0, Caloric Value, Fat, Saturated
Fats,
## Monounsatura...
## ℹ Use `spec()` to retrieve the full column specification for this
data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `` -> `...1`
```

Análisis de Grasas saturadas

```
head(data)
```

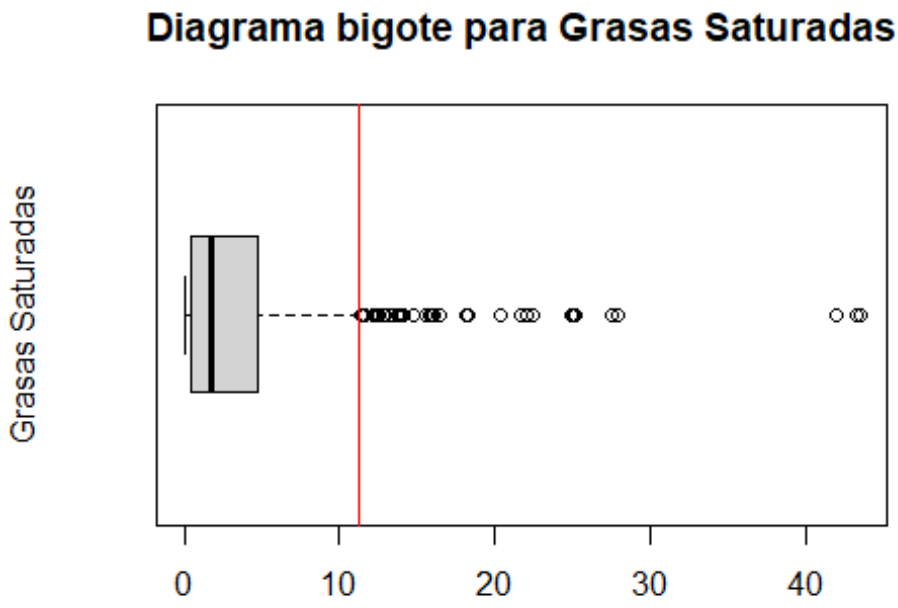
```
## # A tibble: 6 × 37
##    ...1 `Unnamed: 0` food                `Caloric Value`   Fat
`Saturated Fats`
##   <dbl>        <dbl> <chr>                         <dbl> <dbl>
<dbl>
## 1     0            0 cream cheese                     51   5
2.9
## 2     1            1 neufchatel cheese               215  19.4
10.9
## 3     2            2 requeijao cremoso l…             49   3.6
2.3
## 4     3            3 ricotta cheese                   30   2
1.3
```

```
## 5     4            4 cream cheese low fat              30   2.3
1.4
## 6     5            5 cream cheese fat fr…              19   0.2
0.1
## # ℹ 31 more variables: `Monounsaturated Fats` <dbl>,
## #   `Polyunsaturated Fats` <dbl>, Carbohydrates <dbl>, Sugars <dbl>,
## #   Protein <dbl>, `Dietary Fiber` <dbl>, Cholesterol <dbl>, Sodium
<dbl>,
## #   Water <dbl>, `Vitamin A` <dbl>, `Vitamin B1` <dbl>, `Vitamin B11`
<dbl>,
## #   `Vitamin B12` <dbl>, `Vitamin B2` <dbl>, `Vitamin B3` <dbl>,
## #   `Vitamin B5` <dbl>, `Vitamin B6` <dbl>, `Vitamin C` <dbl>,
## #   `Vitamin D` <dbl>, `Vitamin E` <dbl>, `Vitamin K` <dbl>, Calcium
<dbl>, …
```

```
saturadas <- data$`Saturated Fats`
```

Graficar el diagrama de caja y bigote

```
q1=quantile(saturadas,0.25)  #Cuantil 1 de la variable X
ri=IQR(saturadas) # ri= q3-q1   o   ri=IQR(X)    #Rango intercuartílico
de X
q3 = ri + q1

boxplot(saturadas, horizontal = TRUE, main="Diagrama bigote para Grasas
Saturadas", ylab="Grasas Saturadas")
abline(v=q3+1.5*ri,col="red")
```



Se puede observar en el diagrama de caja que para la variable **Grasas Saturadas** existen datos atípicos.

Calcula las principales medidas que te ayuden a identificar datos atípicos (utilizar summary te puede abreviar el cálculo): Cuartil 1, Cuartil 3, Media, Cuartil 2, Rango intercuartílico y Desviación estándar

```
summary(saturadas)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.500   1.800   3.723   4.800  43.500
```

```
print("Desviación estándar")
```

```
## [1] "Desviación estándar"
```

```
sd(saturadas)
```

```
## [1] 5.397021
```

Identifica la cota de 1.5 rangos intercuartílicos para datos atípicos, ¿hay datos atípicos de acuerdo con este criterio? ¿cuántos son?

```
q1 <- quantile(saturadas, 0.25)
print("Q1")
```

```
## [1] "Q1"
```

```
q1
```

```
## 25%
## 0.5
```

```
q3 <- quantile(saturadas, 0.75)
print("Q3")
```

```
## [1] "Q3"
```

```
q3
```

```
## 75%
## 4.8
```

```
iqr <- q3 - q1
print("IQR")
```

```
## [1] "IQR"
```

```
iqr
```

```
## 75%
## 4.3
```

```r
# Límite inferior y superior para datos atípicos (1.5 IQR)
lim_inf <- q1 - 1.5 * iqr
print("Límite inferior")
```

```
## [1] "Límite inferior"
```

```r
lim_inf
```

```
##   25%
## -5.95
```

```r
lim_sup <- q3 + 1.5 * iqr
print("Límite superior")
```

```
## [1] "Límite superior"
```

```r
lim_sup
```

```
##   75%
## 11.25
```

```r
outliers_1.5iqr <- saturadas[saturadas < lim_inf | saturadas > lim_sup]
print("outliers")
```

```
## [1] "outliers"
```

```r
outliers_1.5iqr
```

```
##  [1] 22.0 43.5 20.3 12.8 16.4 16.1 13.3 24.9 25.2 15.8 27.5 13.0 22.5
25.1 43.2
## [16] 11.5 11.4 12.2 14.1 11.4 11.6 18.2 12.4 14.0 11.4 12.6 14.8 13.7
15.6 11.6
## [31] 12.5 15.9 21.6 27.9 13.9 42.0 18.3 21.6 12.3 12.4 12.1 12.4
```

Identifica la cota de 3 desviaciones estándar alrededor de la media, ¿hay datos atípicos de acuerdo con este criterio? ¿cuántos son?

```r
# Límite inferior y superior para datos atípicos (3 desviaciones
estándar)
mean <- mean(saturadas)
print("Media")
```

```
## [1] "Media"
```

```r
mean
```

```
## [1] 3.722715
```

```r
sd <- sd(saturadas)
print("Desviación estándar")
```

```
## [1] "Desviación estándar"
```

```r
sd
```

```
## [1] 5.397021
```

```
outliers_3sd <- saturadas[saturadas < mean - 3 * sd | saturadas > mean +
3 * sd]
print("Outliers")
```

```
## [1] "Outliers"
```

```
outliers_3sd
```

```
##  [1] 22.0 43.5 20.3 24.9 25.2 27.5 22.5 25.1 43.2 21.6 27.9 42.0 21.6
```

Identifica la cota de 3 rangos intercuartílicos para datos extremos, ¿hay datos extremos de acuerdo con este criterio? ¿cuántos son?

```
# cota de 3 rangos intercuartílicos
lim_inf_3 <- q1 - 3 * iqr
print("Límite inferior")
```

```
## [1] "Límite inferior"
```

```
lim_inf_3
```

```
##   25%
## -12.4
```

```
lim_sup_3 <- q3 + 3 * iqr
print("Límite superior")
```

```
## [1] "Límite superior"
```

```
lim_sup_3
```

```
##  75%
## 17.7
```

```
outliers_3iqr <- saturadas[saturadas < lim_inf_3 | saturadas > lim_sup_3]
print("outliers")
```

```
## [1] "outliers"
```

```
outliers_3iqr
```

```
##  [1] 22.0 43.5 20.3 24.9 25.2 27.5 22.5 25.1 43.2 18.2 21.6 27.9 42.0
18.3 21.6
```

Interpreta los resultados obtenidos y argumenta sobre el comportamiento de los datos atípicos y extremos en la variable seleccionada

Existen múltiples datos atípicos para el caso de las Grasas Saturadas, sobre todo en el análisis de la cota de 1.5 rangos intercuartílicos

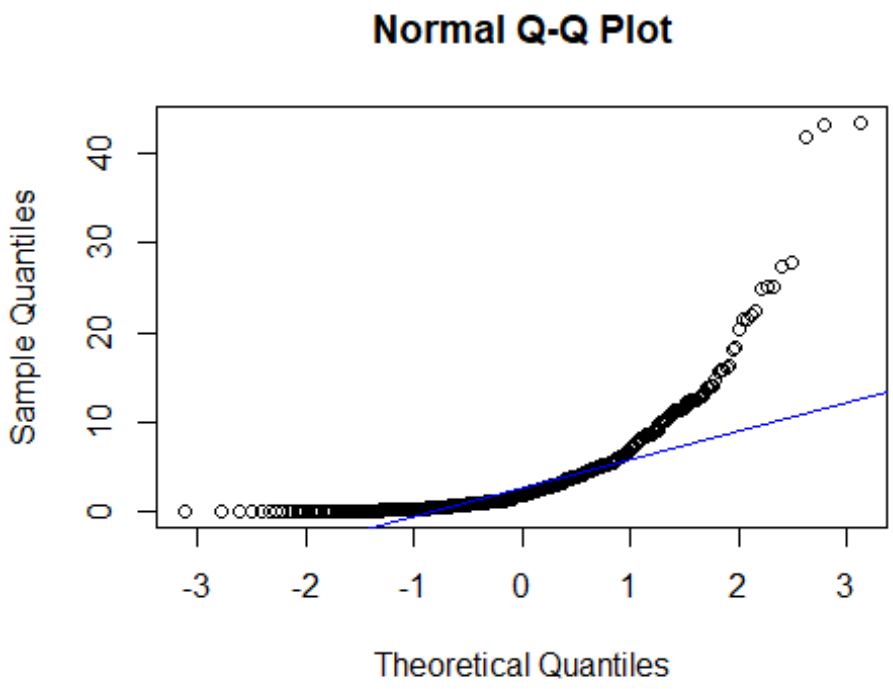
Realiza pruebas de normalidad univariada para la variable (utiliza las pruebas de Anderson-Darling y de Jarque Bera). No olvides incluir H0 y H1 para la prueba de normalidad.

H_{0} = La muestra proviene de una distribución normal h_{1} = La muestra no proviene de una distribución normal

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

##
##  Anderson-Darling normality test
##
## data:  saturadas
## A = 50.094, p-value < 2.2e-16

library(moments)
jarque.test(saturadas)

##
##  Jarque-Bera Normality Test
##
## data:  saturadas
## JB = 7694.1, p-value < 2.2e-16
## alternative hypothesis: greater
```

Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos)

```
qqnorm(saturadas)
qqline(saturadas, col = "blue")
```



Calcula el coeficiente de sesgo y el coeficiente de curtosis

Sesgo: Indica la asimetría de la distribución. Un sesgo cercano a 0 indica simetría; valores positivos indican una cola larga a la derecha, y negativos a la izquierda.

Curtosis: Mide la "agudeza" de la distribución. Un valor de curtosis cercano a 3 indica una distribución normal (mesocúrtica), valores mayores indican distribuciones con colas más pesadas (leptocúrtica), y menores indican colas ligeras (platicúrtica).

```
sesgo <- skewness(saturadas)
print("Sesgo")
```

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

```
sesgo
```

```
## [1] 3.428631
```

```
curtosis <- kurtosis(saturadas)
print("Curtosis")
```

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

```
curtosis
```

```
## [1] 19.97384
```

Compara las medidas de media, mediana y rango medio de cada variable

```
summary(saturadas)
```

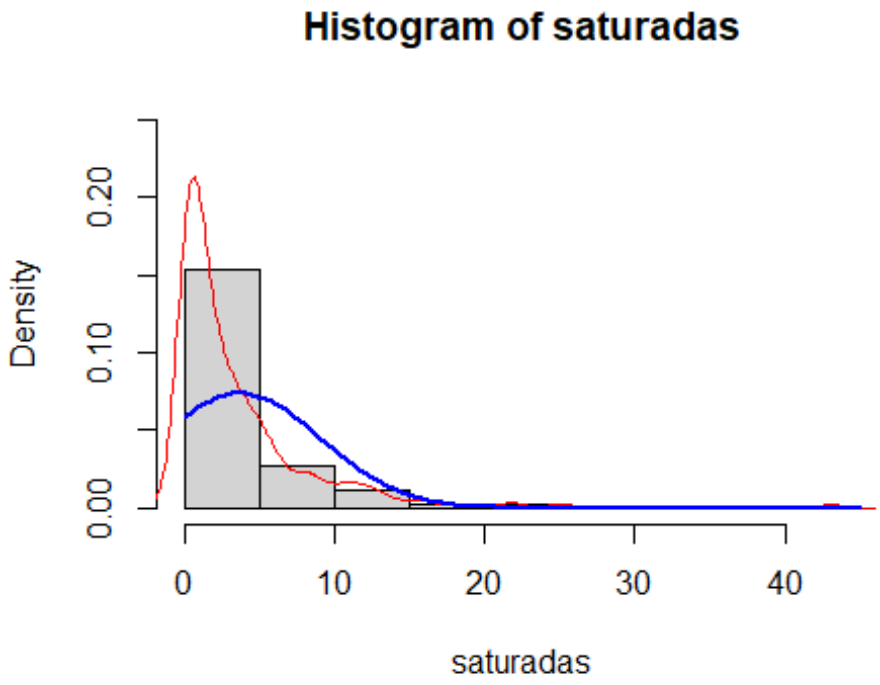
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.500   1.800   3.723   4.800  43.500
```

```
rango_med = (max(saturadas) - min(saturadas))/2 # Rango medio
cat("Rango Medio: ", rango_med)
```

```
## Rango Medio:  21.75
```

Realiza el gráfico de densidad empírica y teórica suponiendo normalidad en la variable. *

```
hist(saturadas,freq=FALSE, ylim = c(0,.25))
lines(density(saturadas),col="red")
curve(dnorm(x,mean=mean(saturadas),sd=sd(saturadas)), add=TRUE,
col="blue",lwd=2)
```



Interpreta los gráficos y los resultados obtenidos en cada punto con vías a indicar si hay normalidad de los datos

Cuando los puntos en el QQ plot siguen aproximadamente la línea recta, los datos son aproximadamente normales, en este caso vemos que en las colas se alejan bastante de la línea lo que nos indica falta de normalidad. Esto también lo podemos observar en las pruebas de normalidad tanto en la prueba de Anderson-Darling como en la de Jarque Bera se rechaza la H_0 ya que nuestros valores p son extremadamente

pequeños, mucho menor que cualquier umbral típico, por lo tanto, la hipótesis nula de normalidad se rechaza con un alto grado de confianza.

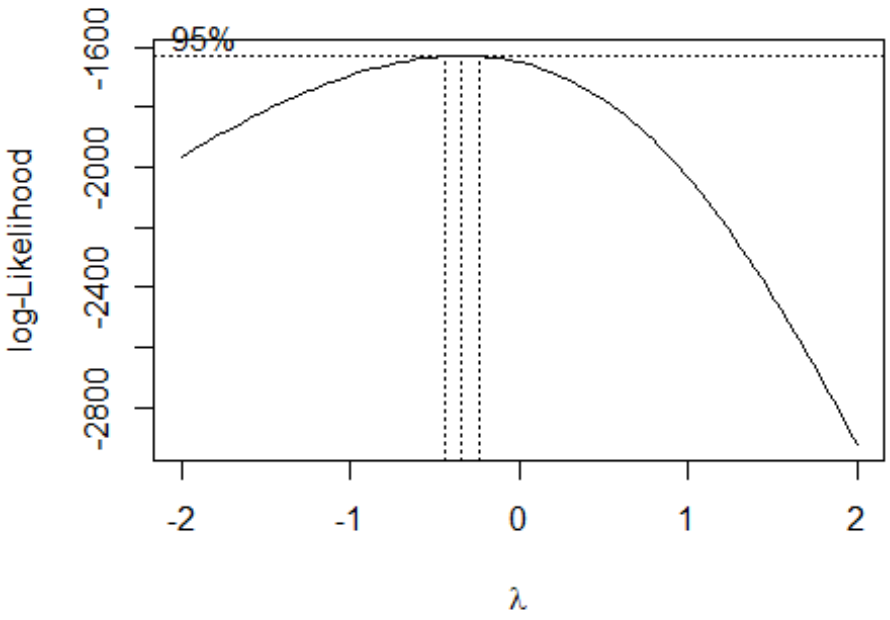
Respecto al sesgo y curtosis. Un sesgo de 3.428631 indica una asimetría positiva significativa en la distribución de tus datos. Esto significa que la cola derecha (valores altos) es más larga o tiene más peso que la cola izquierda y un valor de curtosis de 19.97384 es extremadamente alto, lo que indica que la distribución tiene colas muy pesadas y un pico muy alto.

Transformación a normalidad

Encuentra la mejor transformación de los datos para lograr normalidad. Puedes hacer uso de la transformación Box-Cox o de Yeo Johnson o el comando powerTransform para encontrar la mejor lambda para la transformación. Utiliza el modelo exacto y el aproximado de acuerdo con las sugerencias de Box y Cox para la transformación.

Box-Cox

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




```
l=bc$x[which.max(bc$y)]
l
```

```
## [1] -0.3434343
```

Escribe las ecuaciones de los modelos de transformación encontrados.

Original

```
library(e1071)

##
## Attaching package: 'e1071'

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

summary(saturadas)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.500   1.800   3.723   4.800  43.500

print("Curtosis")

## [1] "Curtosis"

kurtosis(saturadas)

## [1] 16.90141

print("Sesgo")

## [1] "Sesgo"

skewness(saturadas)

## [1] 3.419301
```

Aproximado

$$\frac{1}{\sqrt{x}}$$

```
sat1=1/(sqrt(saturadas+1))
summary(sat1)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1499  0.4152  0.5976  0.6125  0.8165  1.0000

print("Curtosis")

## [1] "Curtosis"

kurtosis(sat1)

## [1] -1.161463
```

```
print("Sesgo")

## [1] "Sesgo"

skewness(sat1)

## [1] 0.0781703
```

Exacto

$$\frac{(x+1)^{-0.3434}-1}{-0.3434}$$

```
sat2=((saturadas+1)^l-1)/l
summary(sat2)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3785  0.8673  0.8670  1.3197  2.1210

print("Curtosis")

## [1] "Curtosis"

kurtosis(sat2)

## [1] -1.084878

print("Sesgo")

## [1] "Sesgo"

skewness(sat2)

## [1] 0.09373237
```

Analiza la normalidad de las transformaciones obtenidas con los datos originales. Utiliza como argumento de normalidad:

Compara las medidas: Mínimo, máximo, media, mediana, cuartil 1 y cuartil 3, sesgo y curtosis.

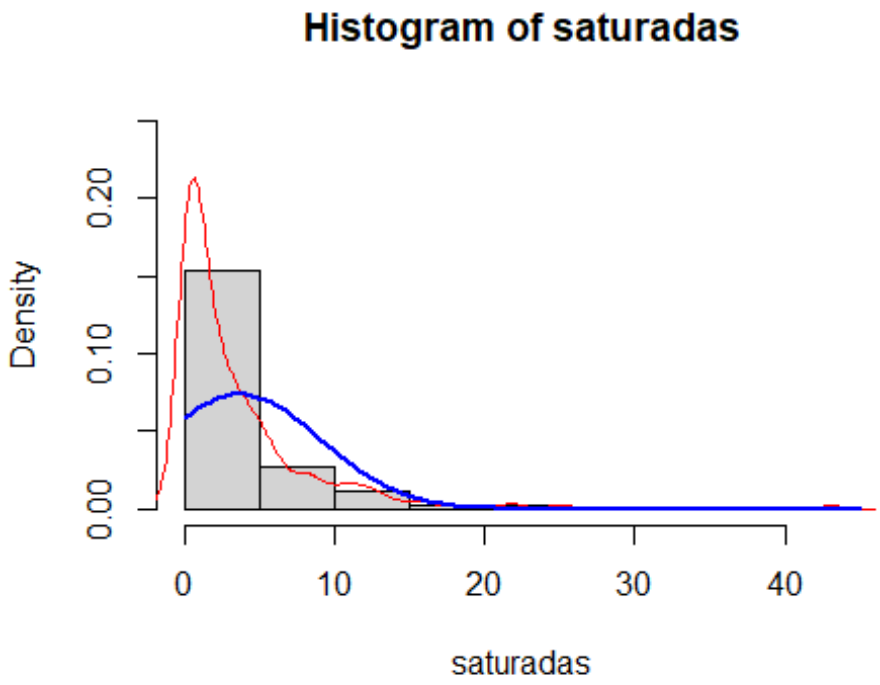
```
m0 = round(c(as.numeric(summary(saturadas)), kurtosis(saturadas),
skewness(saturadas)),3)
m1 = round(c(as.numeric(summary(sat1)), kurtosis(sat1),
skewness(sat1)),3)
m2 = round(c(as.numeric(summary(sat2)), kurtosis(sat2),
skewness(sat2)),3)
m <- as.data.frame(rbind(m0,m1,m2))
row.names(m)=c("Original", "Aproximado", "Exacto")
names(m) = c( "Mínimo", "Q1", "Mediana", "Media", "Q3", "Max",
"Curtosis", "Sesgo")
m
```

```
##            Mínimo    Q1 Mediana Media    Q3    Max Curtosis Sesgo
## Original     0.00 0.500   1.800 3.723 4.800 43.500   16.901 3.419
## Aproximado   0.15 0.415   0.598 0.612 0.816  1.000   -1.161 0.078
## Exacto       0.00 0.379   0.867 0.867 1.320  2.121   -1.085 0.094
```

Grafica las funciones de densidad empírica y teórica de los 2 modelos obtenidos (exacto y aproximado) y los datos originales.

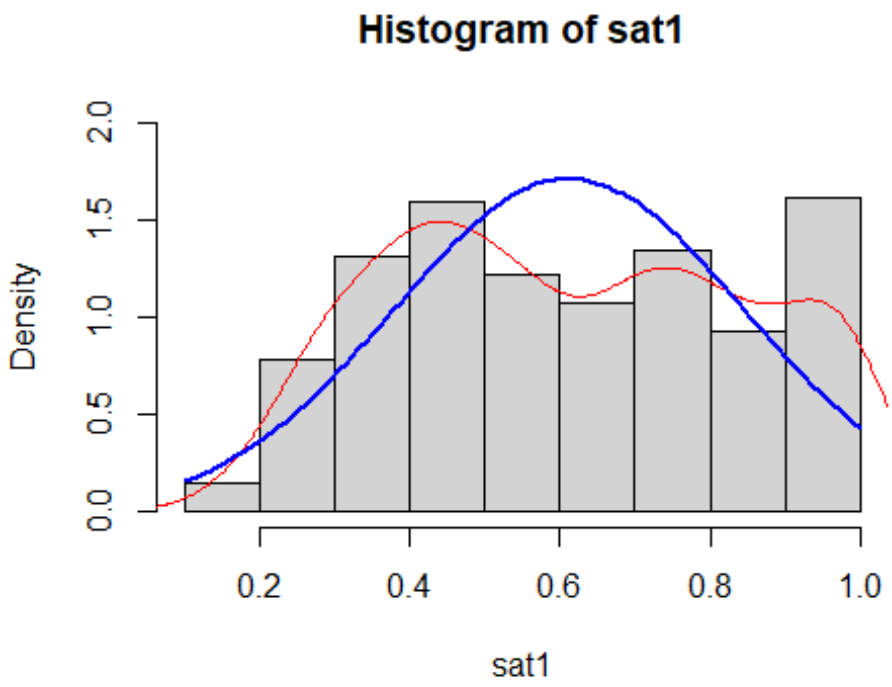
Original

```
hist(saturadas,freq=FALSE, ylim = c(0,.25))
lines(density(saturadas),col="red")
curve(dnorm(x,mean=mean(saturadas),sd=sd(saturadas)), add=TRUE,
col="blue",lwd=2)
```



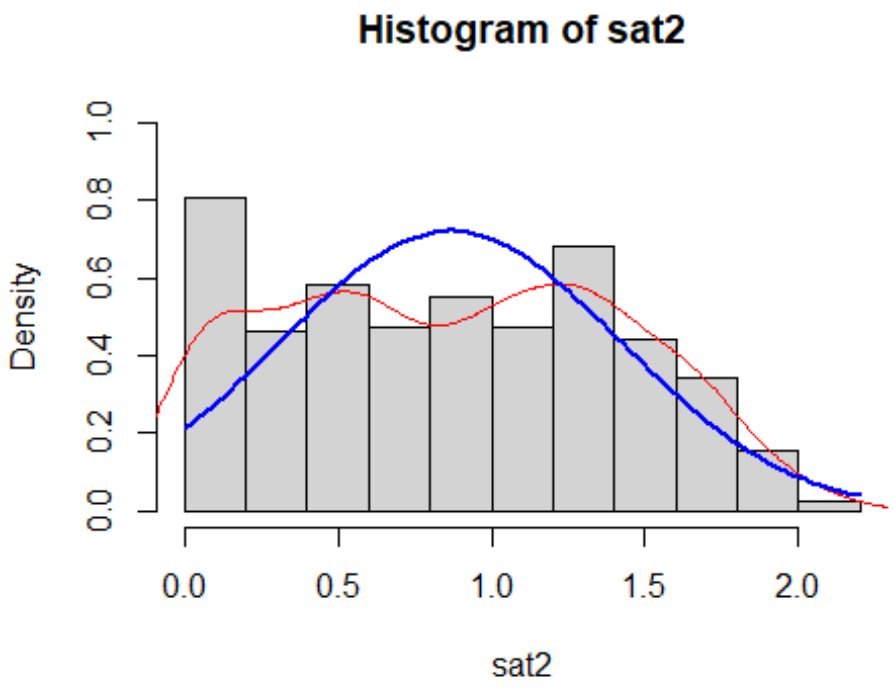
Aproximada

```
hist(sat1,freq=FALSE, ylim = c(0,2))
lines(density(sat1),col="red")
curve(dnorm(x,mean=mean(sat1),sd=sd(sat1)), add=TRUE, col="blue",lwd=2)
```



Aproximada

```
hist(sat2,freq=FALSE, ylim = c(0,1))
lines(density(sat2),col="red")
curve(dnorm(x,mean=mean(sat2),sd=sd(sat2)), add=TRUE, col="blue",lwd=2)
```



Realiza la prueba de normalidad de Anderson-Darling y de Jarque Bera para los datos transformados y los originales

Prueba de Normalidad

H_{0} = La muestra proviene de una distribución normal h_{1} = La muestra no proviene de una distribución normal

Original

```
D=ad.test(saturadas)
D$p.value
```

```
## [1] 3.7e-24
```

```
jarque.test(saturadas)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  saturadas
## JB = 7694.1, p-value < 2.2e-16
## alternative hypothesis: greater
```

Aproximado

```
D=ad.test(sat1)
D$p.value
```

```
## [1] 1.22141e-15
```

```
jarque.test(sat1)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  sat1
## JB = 31.179, p-value = 1.696e-07
## alternative hypothesis: greater
```

Exacto

```
D=ad.test(sat2)
D$p.value
```

```
## [1] 1.762123e-13
```

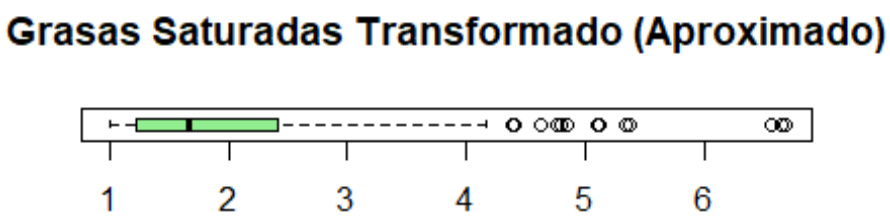
```
jarque.test(sat2)
```

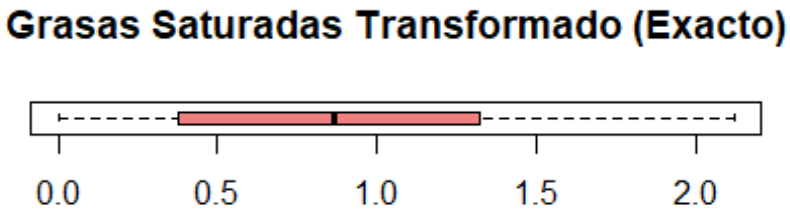
```
##
##  Jarque-Bera Normality Test
##
## data:  sat2
## JB = 27.486, p-value = 1.075e-06
## alternative hypothesis: greater
```

Detecta anomalías y corrige tu base de datos (datos atípicos, ceros anámalos, etc).

Anomalías

```
# Datos transformados (Aproximado y Exacto)
data_clean_approx <- subset(data, sqrt(saturadas + 1) > 0)
data_clean_exact <- subset(data, ((saturadas + 1)^l - 1) / l > 0)

# Boxplot para los datos transformados
par(mfrow = c(2, 1))
boxplot(sqrt(saturadas + 1), horizontal = TRUE, col = "lightgreen", main
= "Grasas Saturadas Transformado (Aproximado)")
boxplot(((saturadas + 1)^l - 1) / l, horizontal = TRUE, col =
"lightcoral", main = "Grasas Saturadas Transformado (Exacto)")
```





Elimino los datos que estén a 3 rangos intercuartílicos

```
saturadas2 <- saturadas[saturadas>lim_inf_3 & saturadas < lim_sup_3]
```

Comenta la normalidad de las transformaciones obtenidas. Utiliza como argumento de normalidad:

Compara las medidas: Mínimo, máximo, media, mediana, cuartil 1 y cuartil 3, sesgo y curtosis.

Original

```
summary(saturadas2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.500   1.700   3.074   4.500  16.400
```

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

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

```
kurtosis(saturadas2)
```

```
## [1] 2.017047
```

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

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

```
skewness(saturadas2)
```

```
## [1] 1.581477
```

Aproximado

$$\sqrt{x+1}$$

```
sat1_1=1/(sqrt(saturadas2+1))
summary(sat1_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2397  0.4264  0.6086  0.6241  0.8165  1.0000
```

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

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

```
kurtosis(sat1_1)
```

```
## [1] -1.216091
```

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

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

```
skewness(sat1_1)
```

```
## [1] 0.1136315
```

Exacto

$$\frac{(x+1)^{-0.3434}-1}{-0.3434}$$

```
sat2_1=((saturadas2+1)^l-1)/l
summary(sat2_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3785  0.8416  0.8362  1.2904  1.8201
```

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

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

```
kurtosis(sat2_1)
```

```
## [1] -1.198279
```

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

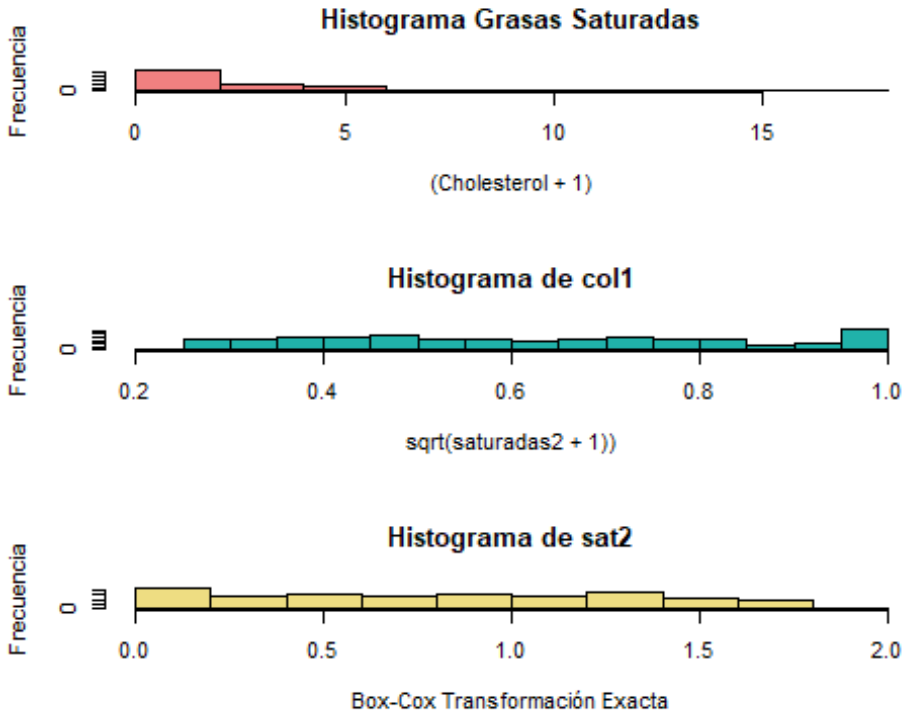
```
## [1] "Sesgo"
```

```
skewness(sat2_1)
```

```
## [1] 0.02815822
```

Obten el histograma de los 2 modelos obtenidos (exacto y aproximado) y de los datos originales.

```
# Histogramas de las transformaciones
par(mfrow = c(3, 1))  # Organizar en una matriz de 3x1

# Histograma Grasas Saturadas
hist(saturadas2, col = "lightcoral",
     main = "Histograma Grasas Saturadas",
     xlab = "(Cholesterol + 1)", ylab = "Frecuencia")

# Histograma aplicado a sat1
hist(sat1_1, col = "lightseagreen",
     main = "Histograma de col1",
     xlab = "sqrt(saturadas2 + 1))", ylab = "Frecuencia")

# Histograma aplicado a sat2
hist(sat2_1, col = "lightgoldenrod",
     main = "Histograma de sat2",
     xlab = "Box-Cox Transformación Exacta", ylab = "Frecuencia")
```



Interpreta la prueba de normalidad de Anderson-Darling y Jarque Bera para los datos transformados y los originales

H_{0} = La muestra proviene de una distribución normal h_{1} = La muestra no proviene de una distribución normal

Original

```
D=ad.test(saturadas2)
D$p.value

## [1] 3.7e-24

jarque.test(saturadas2)

##
##  Jarque-Bera Normality Test
##
## data:  saturadas2
## JB = 317.25, p-value < 2.2e-16
## alternative hypothesis: greater
```

Aproximado

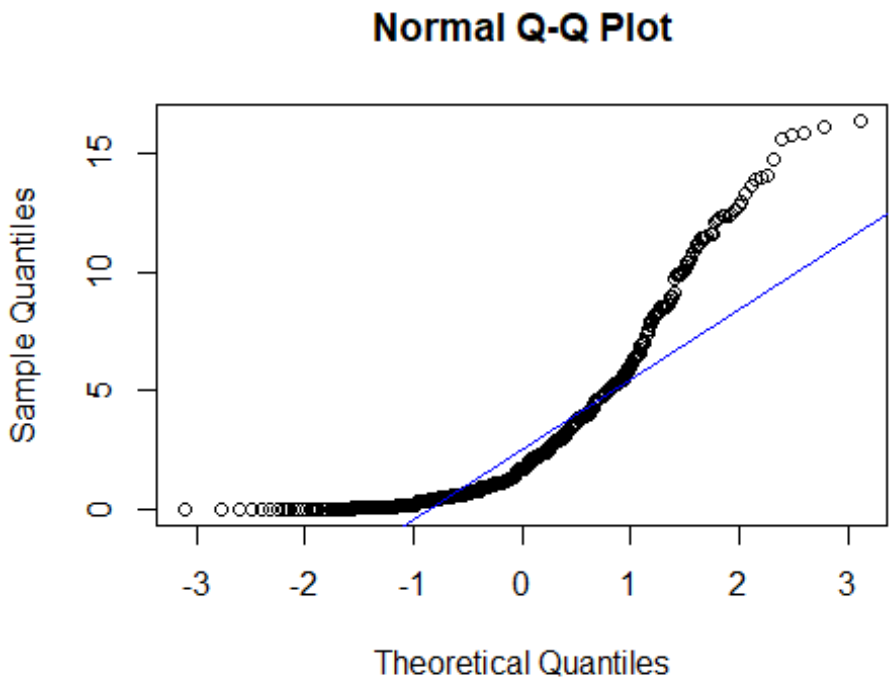
```
D=ad.test(sat1_1)
D$p.value

## [1] 1.502936e-17

jarque.test(sat1_1)

##
##  Jarque-Bera Normality Test
##
## data:  sat1_1
## JB = 33.827, p-value = 4.515e-08
## alternative hypothesis: greater
```

Exacto

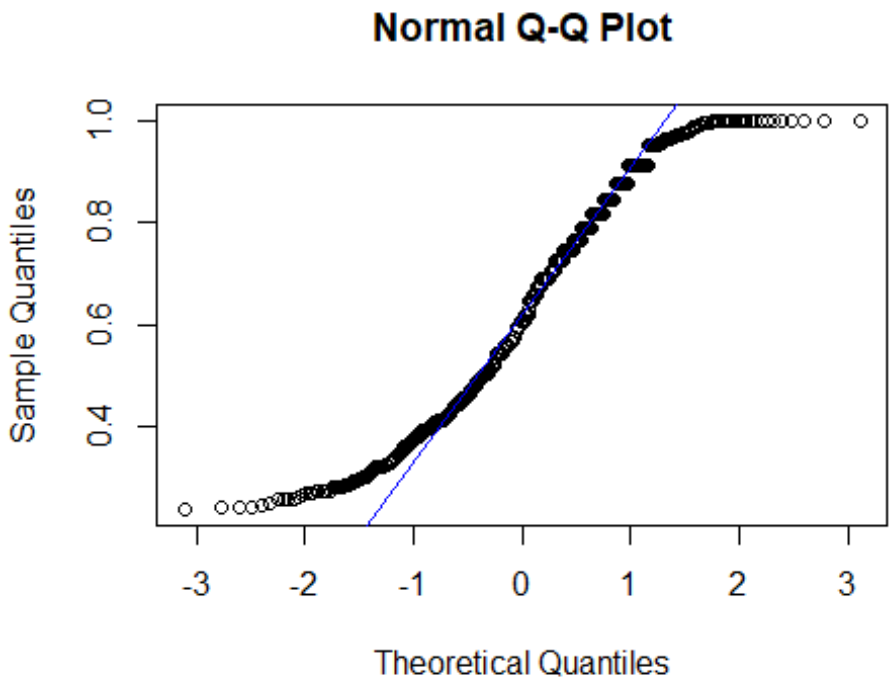
```
D=ad.test(sat2_1)
D$p.value

## [1] 1.310796e-15

jarque.test(sat2_1)

##
##  Jarque-Bera Normality Test
##
## data:  sat2_1
## JB = 31.779, p-value = 1.257e-07
## alternative hypothesis: greater
```

Indica posibilidades de motivos de alejamiento de normalidad (sesgo, curtosis, datos atípicos, etc)

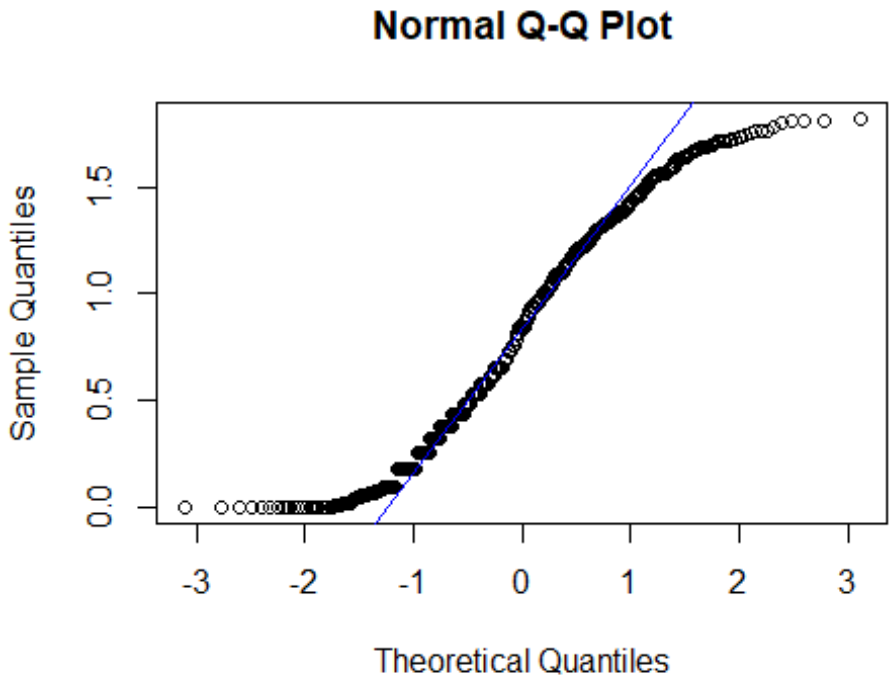
```
qqnorm(saturadas2)
qqline(saturadas2, col = "blue")
```




```
qqnorm(sat1_1)
qqline(sat1_1, col = "blue")
```




```r
qqnorm(sat2_1)
qqline(sat2_1, col = "blue")
```



Define la mejor transformación de los datos de acuerdo a las características de los modelos que encontraste. Toma en cuenta los criterios del inciso anterior para analizar normalidad y la economía del modelo.

Después de la transformación en génerar los valoresp mejoran acercandose más a la normalidad, sobretodo en el modelo aproximado, es el que más se aproxima a la normalidad cambiando de e-16 a e-08 en la prueba de normalidad de Jarque-Bera , sin embargo no es suficiente para considerar que los datos se distribuyen con normalidad ya que sigue siendo un valor extremadamente pequeña por lo que se sigue rechazando la hipótesis nula. Así mismo, su valor de curtosis de 1.790584 nos indica tiene colas menos pesadas que una distribución normal y el sesgo de 0.1139502 indica que la distribución tiene una asimetría positiva muy leve. Esto lo podemos visualizar en el QQPlot, los datos se desvían significativamente de la línea diagonal en ambos extremos, demostrando la asimetría positiva, dado que la parte inferior de los puntos se curva hacia abajo y la parte superior se curva hacia arriba.Y seguimos viendo datos atípicos. que son los puntos alejados de la línea en los extremos superior e inferior. Se podría probar eliminando más datos atípicos para buscar la normalidad, sin embargo decidí únicamente eliminar los datos que estén a 3 rangos intercuartílicos para no afectar significativamente y modificar la base de datos.

A pesar de que después de la transformación podemos observar una mejora y acercamiento hacia la normalidad no es suficiente la mejoría por lo que se sigue rechazando la hipotesis nula y concluyendo que las Grasas Saturadas no siguen una distribución normal.