# Actividad Integradora 1

Ricardo Salinas

2024-08-20

#Punto 1. Análisis descriptivo de la variable

Analiza una de las siguientes variables en cuanto a sus datos atípicos y normalidad. La variable que te corresponde analizar te será asignada por tu profesora al inicio de la actividad:

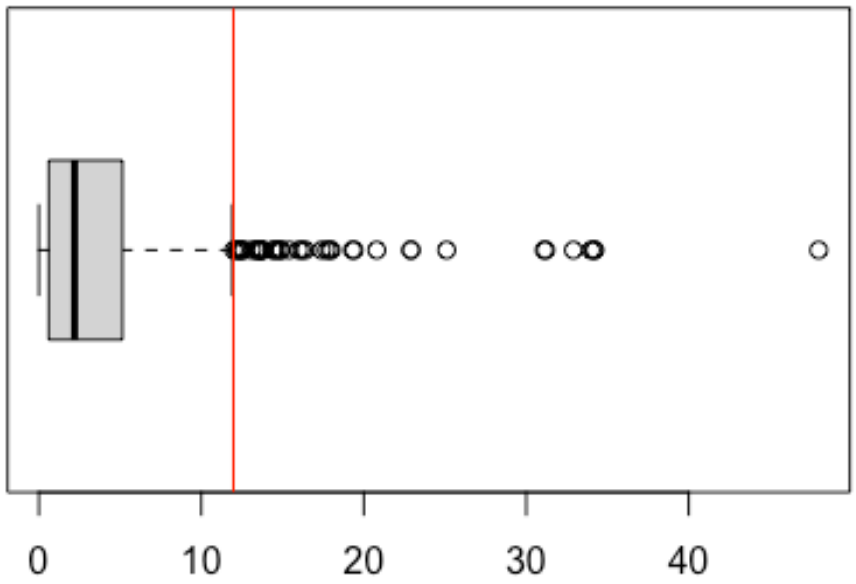
Calorías Grasas saturadas **Grasas monosaturadas** Sodio Agua Sodio Densidad Nutricional

1. Para analizar datos atípicos se te sugiere:

```
#Primeramente asignaremos el dataset a una variable
M = read.csv("Downloads/food_data_g.csv")
M1 = M$Monounsaturated.Fats

#Graficar el diagrama de caja y bigote

boxplot(M1,horizontal=TRUE)


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

summary(M1)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.600   2.200   4.002   5.150  48.000

ri=IQR(M1)
print("El rango intercuartilico de Grasas Monosaturadas es:")

## [1] "El rango intercuartilico de Grasas Monosaturadas es:"

print(ri)

## [1] 4.55

print("La desviacion estandar de Grasas Monosaturadas es:")

## [1] "La desviacion estandar de Grasas Monosaturadas es:"

de = sd(M1)
print(de)
```

```
## [1] 5.540608
```

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

q3=quantile(M1, 0.75)


boxplot(M1,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 atípicos o extremos
```

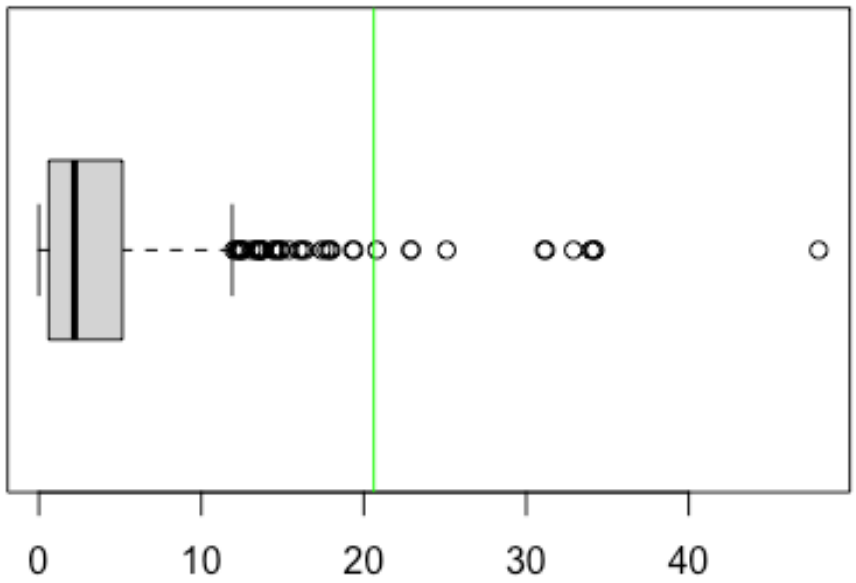



```
v = q3 + 1.5 * ri
da = sum(M1 > v)

print("El numero de variables atipicos con 1.5 rangos intercuartílicos son:")

## [1] "El numero de variables atipicos con 1.5 rangos intercuartílicos son:"

print(da)

## [1] 40
```

```r
print("La variable cuenta con una gran cantidad de datos atipicos, ya que estos cruzan la linea roja, sobrepasando el valor de 1.5 rangos intercuartilicos")
```

```
## [1] "La variable cuenta con una gran cantidad de datos atipicos, ya que estos cruzan la linea roja, sobrepasando el valor de 1.5 rangos intercuartilicos"
```

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

q3=quantile(M1, 0.75)


boxplot(M1,horizontal=TRUE)   #y1=min en la escala del eje Y, y2=máx en la escala del eje
abline(v=mean(M1)+3*sd(M1), col="green")  #linea vertical en el límite de los datos atípicos o extremos
```



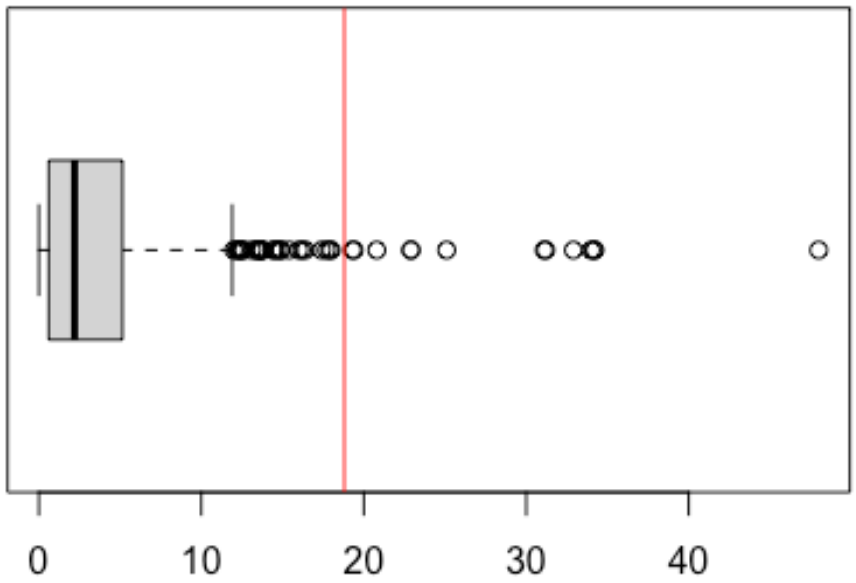

```r
v1 = mean(M1)+3*sd(M1)
da1 = sum(M1 > v1)

print("El numero de variables atipicos con 3 desviaciones estandar son:")
```

```
## [1] "El numero de variables atipicos con 3 desviaciones estandar son:"
```

```
print(da1)
```

```
## [1] 11
```

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

boxplot(M1,horizontal=TRUE)   #y1=min en la escala del eje Y, y2=máx en la escala del eje Y
abline(v=q3+3*ri, col="red")  #linea vertical en el límite de los datos atípicos o extremos
```




```
v2 = q3+3*ri
da2 = sum(M1 > v2)

print("El numero de variables atipicos con 3 rangos intercuartílicos son:")
```

```
## [1] "El numero de variables atipicos con 3 rangos intercuartílicos son:"
```

```
print(da2)
```

```
## [1] 13
```

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

print("Los datos nos demuestran que tenemos datos atipicos un tanto extremos a compraracion de la mayoria de los datos, siendo que se siguen teniendo datos atipicos fuera de los 3 rangos intercuantilicos y tambien con 3 desviaciones estandar, aunque estos no son muchos datos, pueden causar inconsistencias al trabajar con estos.")
```

```
## [1] "Los datos nos demuestran que tenemos datos atipicos un tanto extremos a compraracion de la mayoria de los datos, siendo que se siguen teniendo datos atipicos fuera de los 3 rangos intercuantilicos y tambien con 3 desviaciones estandar, aunque estos no son muchos datos, pueden causar inconsistencias al trabajar con estos."
```

2. Para analizar normalidad se te sugiere:

```r
#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.

library(nortest)
library(moments)

ad.test(M1)
```

```
##
##  Anderson-Darling normality test
##
## data:  M1
## A = 46.499, p-value < 2.2e-16
```

```r
jarque.test(M1)
```

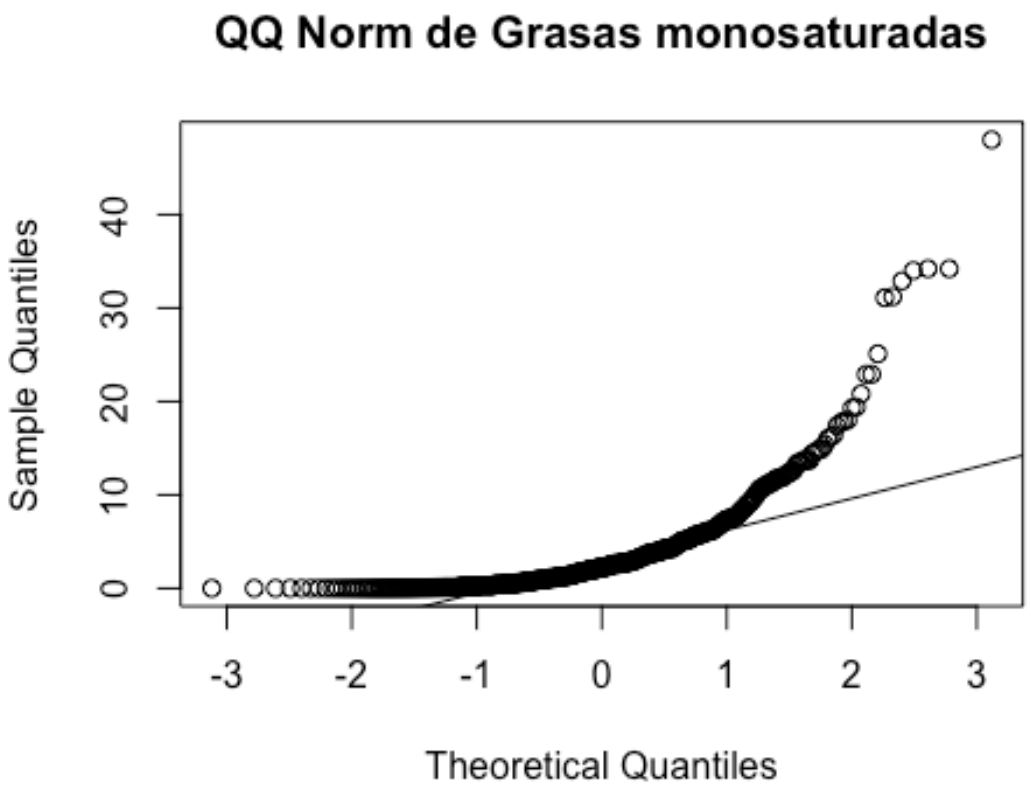
```
##
##  Jarque-Bera Normality Test
##
## data:  M1
## JB = 5884, p-value < 2.2e-16
## alternative hypothesis: greater
```

```r
print("Nuestra variable muestra en ambos tests que no se puede considerar como normal, siendo que tiene una p extremadamente baja, entonces se rechaza H0 y se concluye que la variable no sigue una distribucion normal")
```

```
## [1] "Nuestra variable muestra en ambos tests que no se puede considerar como normal, siendo que tiene una p extremadamente baja, entonces se rechaza H0 y se concluye que la variable no sigue una distribucion normal"
```

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

qqnorm(M1, main = "QQ Norm de Grasas monosaturadas")
```

```r
qqline(M1, main = "QQ Line de Grasas monosaturadas")
```




```r
#Calcula el coeficiente de sesgo y el coeficiente de curtosis

print("Curtosis:")
```

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

```r
kurtosis(M1)
```

```
## [1] 17.66712
```

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

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

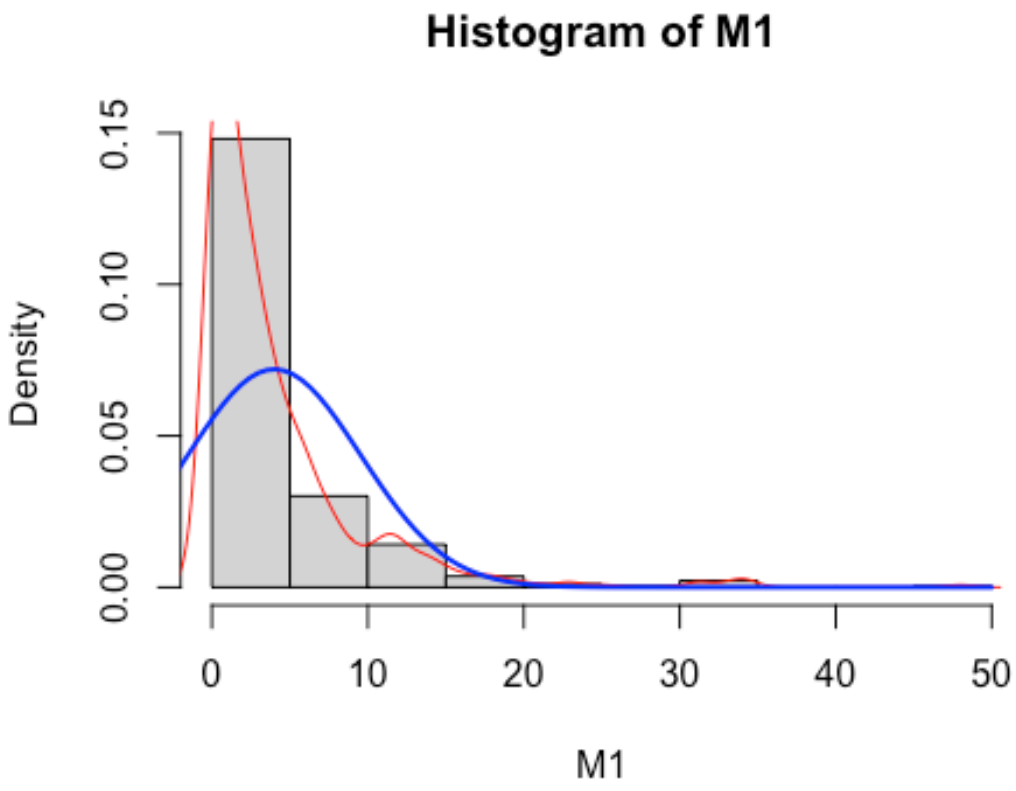
```r
skewness(M1)
```

```
## [1] 3.207982
```

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

summary(M1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.600   2.200   4.002   5.150  48.000
```

```
#Realiza el gráfico de densidad empírica y teórica suponiendo normalidad en
la variable. Adapta el código:

hist(M1,freq=FALSE)
lines(density(M1),col="red")
curve(dnorm(x,mean=mean(M1),sd=sd(M1)), from=-6, to=50, 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


print("Las graficas nos estan demostrando que los datos cuentan con una
cantidad importante de datos atipicos, en las graficas qqplot, podemos ver
como los datos no siguen la linea la cual denota si se cuenta con una
distribucion normal, por lo cual se puede concluir con que no se cuenta con
una distribucion normal.")
```

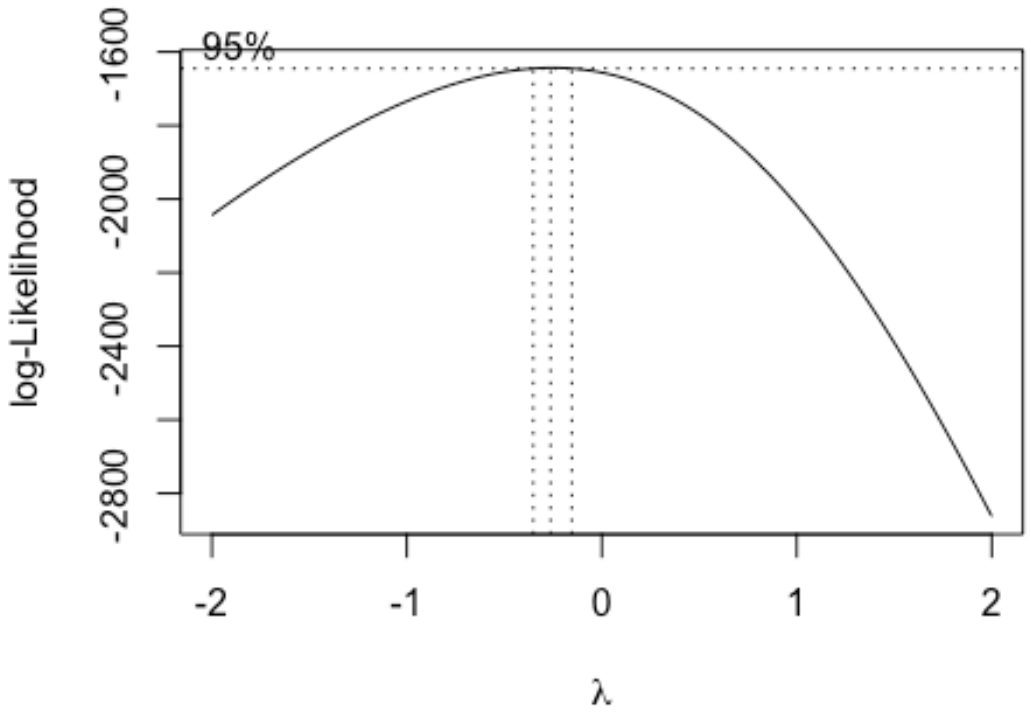
```
## [1] "Las graficas nos estan demostrando que los datos cuentan con una
cantidad importante de datos atipicos, en las graficas qqplot, podemos ver
```

```
como los datos no siguen la linea la cual denota si se cuenta con una
distribucion normal, por lo cual se puede concluir con que no se cuenta con
una distribucion normal."

#Comenta las características encontradas:

print("Ambas pruebas, tanto Anderson-Darling y de Jarque Bera, rechazan h0,
en cuanto al sesgo, nuestro valor de 3.21 nos muestra que tenemos un sesgo
hacia la derecha, y la curtosis nos muestra un valor de 17.66712, lo cual nos
demuestra que tenemos datos atipicos muy alejados de nuestros datos
regulares.")

## [1] "Ambas pruebas, tanto Anderson-Darling y de Jarque Bera, rechazan h0,
en cuanto al sesgo, nuestro valor de 3.21 nos muestra que tenemos un sesgo
hacia la derecha, y la curtosis nos muestra un valor de 17.66712, lo cual nos
demuestra que tenemos datos atipicos muy alejados de nuestros datos
regulares."

#Considera alejamientos de normalidad por simetría, curtosis

print("Se puede concluir facilmente que estos datos no muestran simetria
gracias al sesgo el cual indica hacia la derecha, y la curtosis tiene un
valor muy elevado el cual denota la presencia de datos outliers.")

## [1] "Se puede concluir facilmente que estos datos no muestran simetria
gracias al sesgo el cual indica hacia la derecha, y la curtosis tiene un
valor muy elevado el cual denota la presencia de datos outliers."

#Comenta si hay aparente influencia de los datos atípicos en la normalidad de
los datos

print("Con lo antes dicho, se puede notar facilmente con los datos y con las
representaciones visuales que la distribucion no es normal, dado por sus
altas cantidades de datos atipicos.")

## [1] "Con lo antes dicho, se puede notar facilmente con los datos y con las
representaciones visuales que la distribucion no es normal, dado por sus
altas cantidades de datos atipicos."

#Emite una conclusión sobre la normalidad de los datos. Se debe argumentar en
términos de los 3 puntos analizados: las pruebas de normalidad, los gráficos
y las medidas.

print("Las pruebas de nromalidad nos confirmaron de manera muy clara que se
rechaza h0, ya que el valor de p es muy bajo, los graficos nos ayudan a
visualizar la poca simetria que se tiene y los valores outliers que hay en
los extremos, y las medidad nos clarifican datos que llegan a estar muy lejos
de la media, la media siento 4.002 y nuestro maximo siendo 48.000 ")

## [1] "Las pruebas de nromalidad nos confirmaron de manera muy clara que se
rechaza h0, ya que el valor de p es muy bajo, los graficos nos ayudan a
```

```
visualizar la poca simetria que se tiene y los valores outliers que hay en
los extremos, y las medidad nos clarifican datos que llegan a estar muy lejos
de la media, la media siento 4.002 y nuestro maximo siendo 48.000 "
```

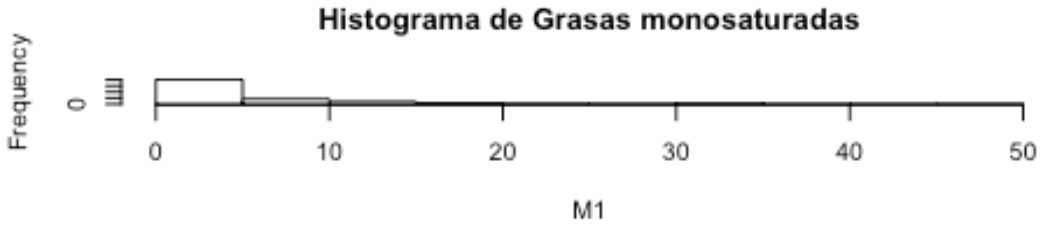
Punto 2. Transformación a normalidad

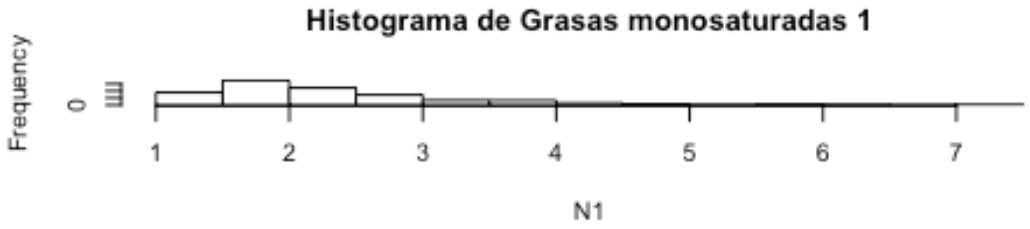
```
#Encuentra la mejor transformación de los datos para lograr normalidad.
Puedes hacer uso de la transformación Box-Cox o de Yeo Johnson o el comando
powerTransform para encontrar la mejor lambda para la transformación. Utiliza
el modelo exacto y el aproximado de acuerdo con las sugerencias de Box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

library(MASS)

#Box Cox
bc <- boxcox((M1 + 1) ~ 1)
```

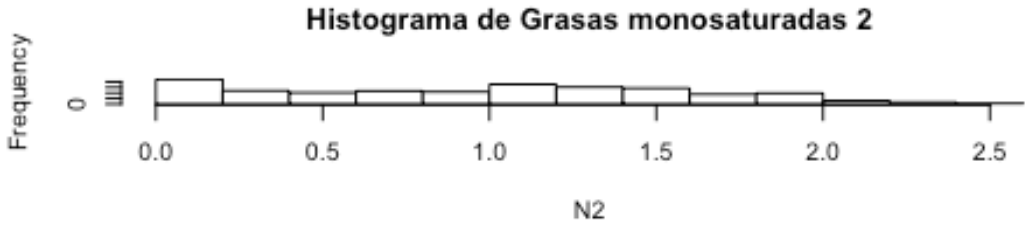



```
l=bc$x[which.max(bc$y)]
print("El valor optimo de lambda es:")

## [1] "El valor optimo de lambda es:"

print(l)
```

```
## [1] -0.2626263

#Histogramas
N1=sqrt((M1 + 1) + 1)
N2=((M1+1)^l-1)/l
par(mfrow=c(3,1))
hist(M1,col=0,main="Histograma de Grasas monosaturadas")
hist(N1,col=0,main="Histograma de Grasas monosaturadas 1")
hist(N2,col=0,main="Histograma de Grasas monosaturadas 2")
```




```
#Escribe las ecuaciones de los modelos de transformación encontrados.
```

x1 = $\sqrt{x+1}$

x2 = $\frac{x^{-}.26-1}{-0.26}$

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.

print("Datos originales:")
```

```
## [1] "Datos originales:"
```

```
summary(M1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.600   2.200   4.002   5.150  48.000
```

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

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

```
kurtosis(M1)
```

```
## [1] 17.66712
```

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

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

```
skewness(M1)
```

```
## [1] 3.207982
```

```
print("Nuevos datos:")
```

```
## [1] "Nuevos datos:"
```

```
summary(N2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.4422  1.0023  0.9658  1.4446  2.4375
```

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

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

```
kurtosis(N2)
```

```
## [1] 2.003913
```

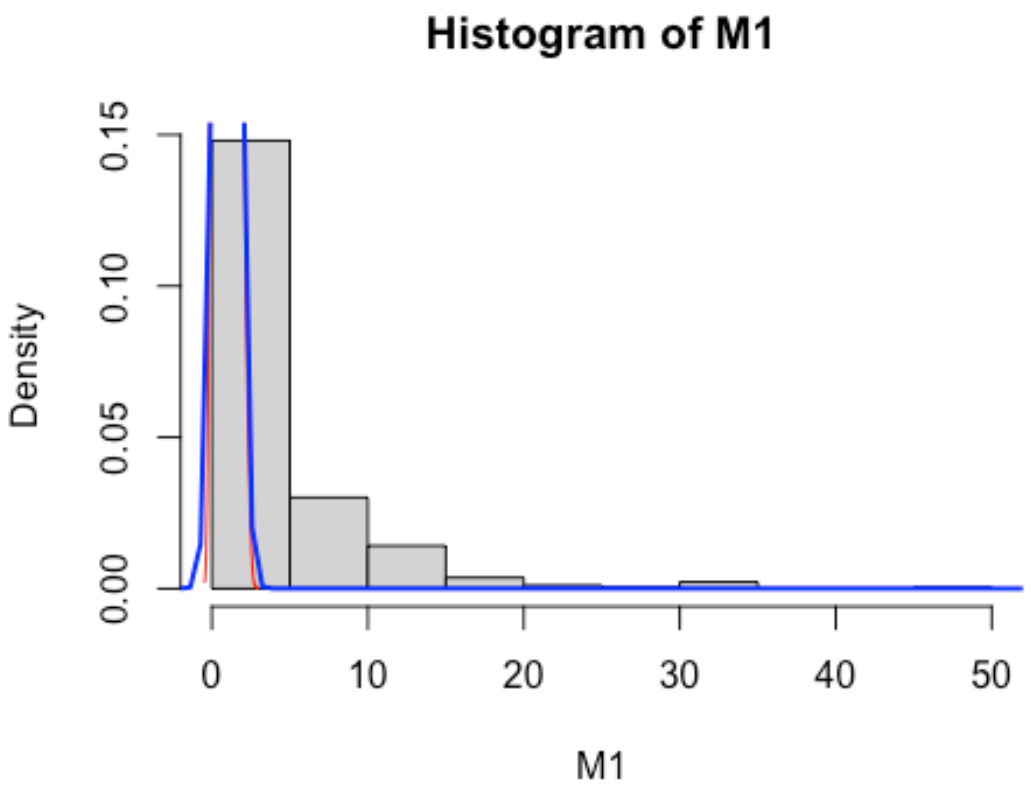
```
print("Sesgo:")
```

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

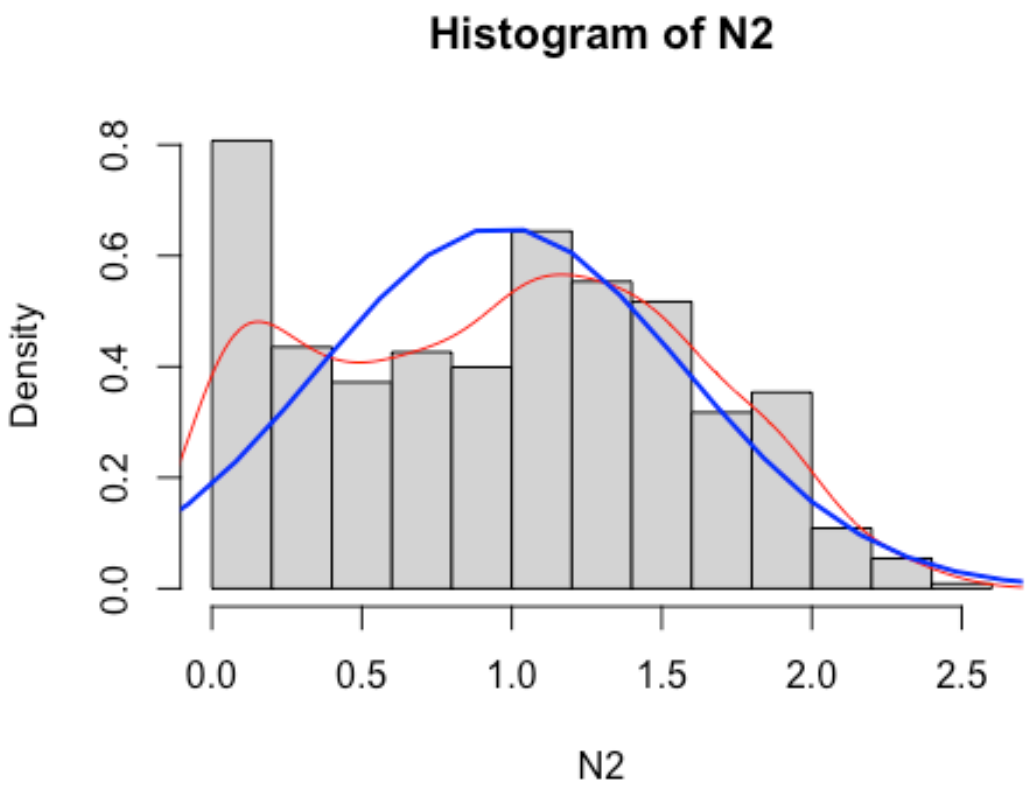
```
skewness(N2)
```

```
## [1] 0.05428609
```

```
#Grafica las funciones de densidad empírica y teórica de los 2 modelos
obtenidos (exacto y aproximado) y los datos originales.
hist(M1,freq=FALSE)
lines(density(N2),col="red")
curve(dnorm(x,mean=mean(N2),sd=sd(N2)), from=-6, to=60, add=TRUE,
col="blue",lwd=2)
```




```
hist(N2,freq=FALSE)
lines(density(N2),col="red")
curve(dnorm(x,mean=mean(N2),sd=sd(N2)), from=-6, to=10, 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

library(nortest)
library(moments)
print("Datos originales:")

## [1] "Datos originales:"

ad.test(M1)

##
##  Anderson-Darling normality test
##
## data:  M1
## A = 46.499, p-value < 2.2e-16

jarque.test(M1)

##
##  Jarque-Bera Normality Test
##
## data:  M1
```

```
## JB = 5884, p-value < 2.2e-16
## alternative hypothesis: greater
```

```
print("Datos nuevos:")
```

```
## [1] "Datos nuevos:"
```

```
ad.test(N2)
```

```
##
##  Anderson-Darling normality test
##
## data:  N2
## A = 4.7415, p-value = 9.375e-12
```

```
jarque.test(N2)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  N2
## JB = 23.05, p-value = 9.882e-06
## alternative hypothesis: greater
```

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

print("Se tienen una gran cantidad de datos en cero, los cuales estan
afectando la distribucion de los datos, asi que creare una nueva variable la
cual elimine estos.")
```

```
## [1] "Se tienen una gran cantidad de datos en cero, los cuales estan
afectando la distribucion de los datos, asi que creare una nueva variable la
cual elimine estos."
```

```
N3 = N2[N2 != 0]
```

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.

print("Datos originales:")
```

```
## [1] "Datos originales:"
```

```
summary(M1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.600   2.200   4.002   5.150  48.000
```

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

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

```
kurtosis(M1)
```

```
## [1] 17.66712
```

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

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

```
skewness(M1)
```

```
## [1] 3.207982
```

```
print("Nuevos datos:")
```

```
## [1] "Nuevos datos:"
```

```
summary(N3)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.002994 0.495314 1.035730 1.000328 1.459498 2.437530
```

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

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

```
kurtosis(N3)
```

```
## [1] 2.052755
```

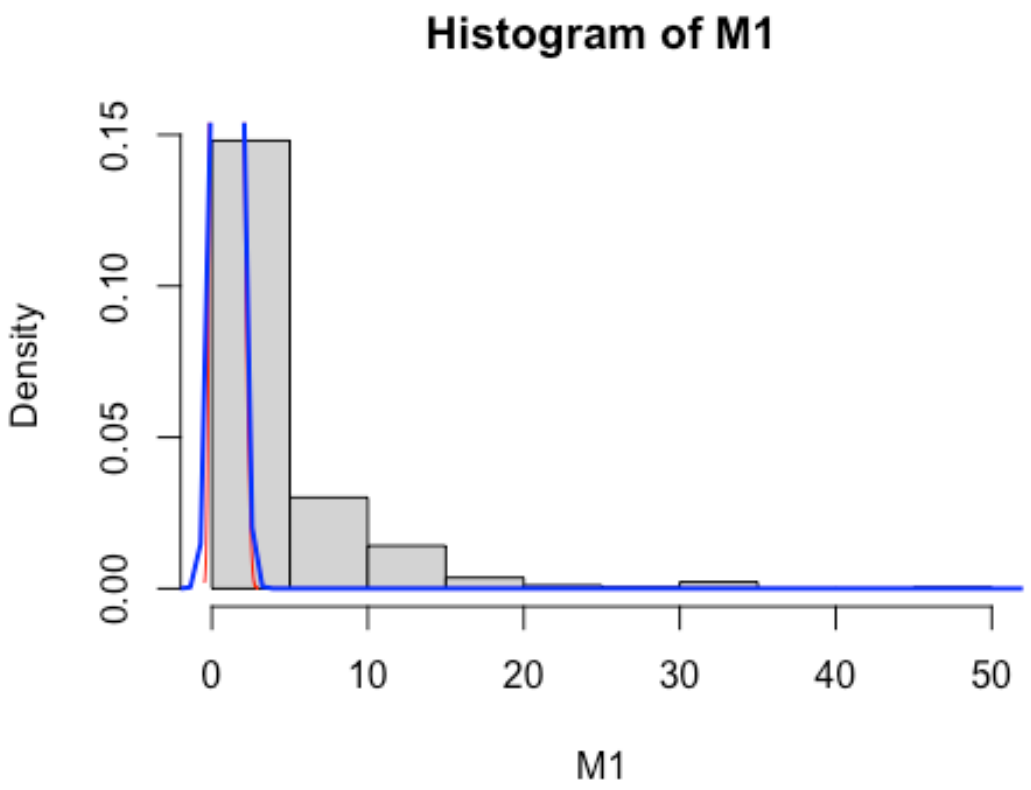
```
print("Sesgo:")
```

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

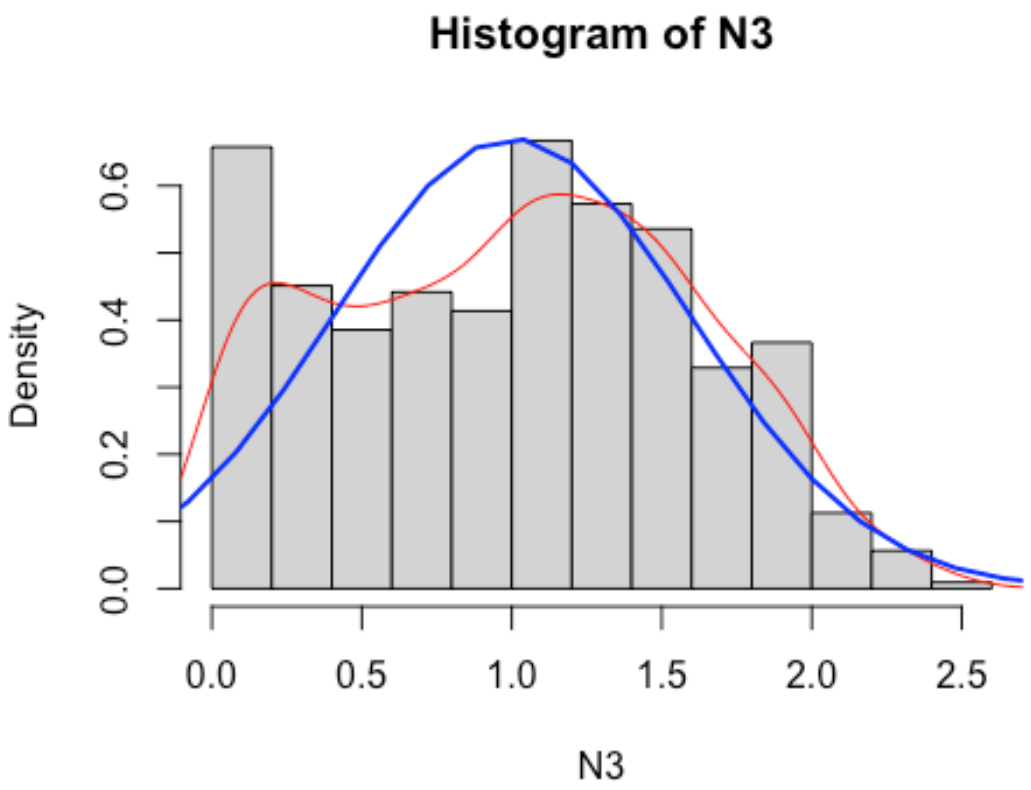
```
skewness(N3)
```

```
## [1] 0.04016749
```

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

hist(M1,freq=FALSE)
lines(density(N2),col="red")
curve(dnorm(x,mean=mean(N2),sd=sd(N2)), from=-6, to=60, add=TRUE,
col="blue",lwd=2)
```




```
hist(N3,freq=FALSE)
lines(density(N3),col="red")
curve(dnorm(x,mean=mean(N3),sd=sd(N3)), from=-6, to=10, add=TRUE,
col="blue",lwd=2)
```




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


print("Datos originales:")

## [1] "Datos originales:"

ad.test(M1)

##
##  Anderson-Darling normality test
##
## data:  M1
## A = 46.499, p-value < 2.2e-16

jarque.test(M1)

##
##  Jarque-Bera Normality Test
##
## data:  M1
## JB = 5884, p-value < 2.2e-16
## alternative hypothesis: greater
```

```r
print("Datos nuevos:")
```

```
## [1] "Datos nuevos:"
```

```r
ad.test(N3)
```

```
##
##  Anderson-Darling normality test
##
## data:  N3
## A = 3.7189, p-value = 2.756e-09
```

```r
jarque.test(N3)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  N3
## JB = 20.033, p-value = 4.467e-05
## alternative hypothesis: greater
```

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

print("Los datos siguen teniendo una distribucion asimetrica, ya que se tiene que respetar la integridad de lso datos, y estos originalmente estaban muy dispersos, se logro eliminar datos atipicos los cuales causaban un sesgo muy alto, y ahora con nuestra nueva variable tenemos un valor mucho mas pequeño.")
```

```
## [1] "Los datos siguen teniendo una distribucion asimetrica, ya que se tiene que respetar la integridad de lso datos, y estos originalmente estaban muy dispersos, se logro eliminar datos atipicos los cuales causaban un sesgo muy alto, y ahora con nuestra nueva variable tenemos un valor mucho mas pequeño."
```

```r
#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.

print("Box cox resulta muy util al momento de querer tranformar y normalizar los datos, pero tambien es buena opcion analizar los datos de manera perosnal para encontrar datos que puedan estar afectando los analisis, ya que este dataset contaba con una buena cantidad de ceros los cuales estaban afectando las pruebas, pero como se considera que estos datos estaban cerca de la media, se mantuvieron y no se vieron afectados, por lo cual decidi eliminarlos, lo que resulto en nuestra ultima variable N3")
```

```
## [1] "Box cox resulta muy util al momento de querer tranformar y normalizar los datos, pero tambien es buena opcion analizar los datos de manera perosnal para encontrar datos que puedan estar afectando los analisis, ya que este dataset contaba con una buena cantidad de ceros los cuales estaban afectando
```

las pruebas, pero como se considera que estos datos estaban cerca de la media, se mantuvieron y no se vieron afectados, por lo cual decidi eliminarlos, lo que resulto en nuestra ultima variable N3"