Actividad 3: Transformaciones

Samantha Daniela Guanipa Ugas A01703936

2023-08-18

Para esta actividad se seleccionó la variable “sugars”

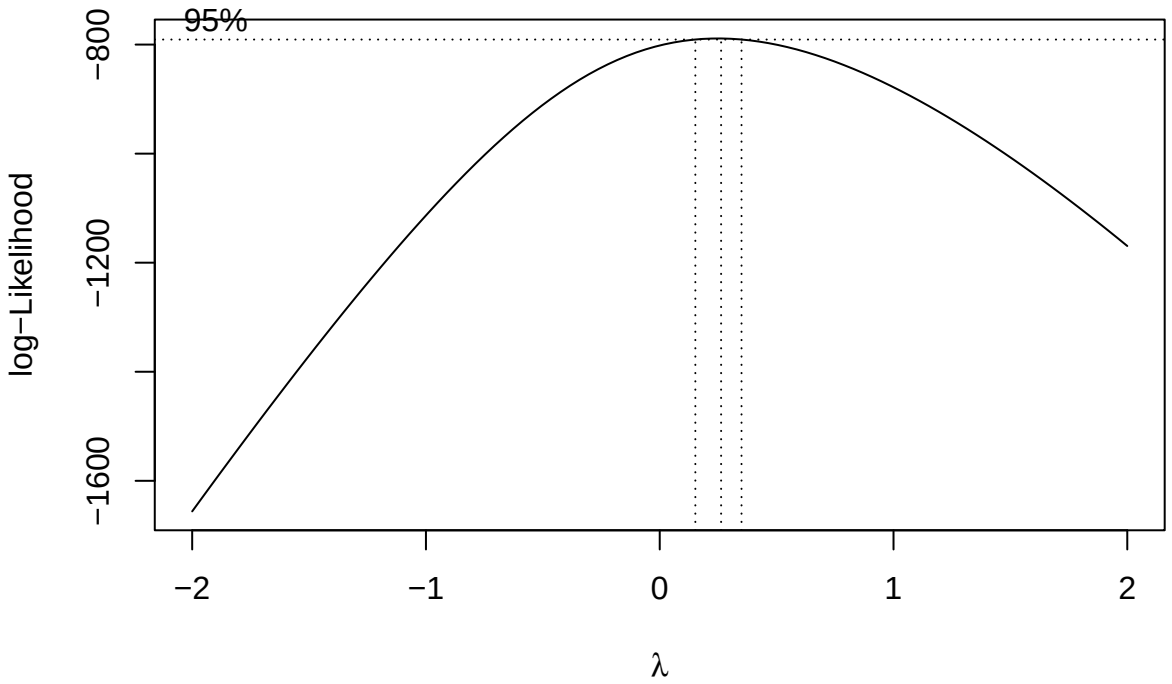
Transformación Box-Cox

1. Utiliza la transformación Box-Cox. Utiliza el modelo exacto y el aproximado de acuerdo con las sugerencias de Box y Cox para la transformación

```
M1=read.csv("mc-donalds-menu-act2.csv") #leer la base de datos
```

a) Se calcula el valor de lambda que maximiza la función de verosimilitud es 0.2626263

```
library(MASS)
bc<-boxcox((M1[ ,19]+1)~1)
```




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

```
## [1] 0.2626263
```

2. Escribe las ecuaciones de los modelos encontrados.

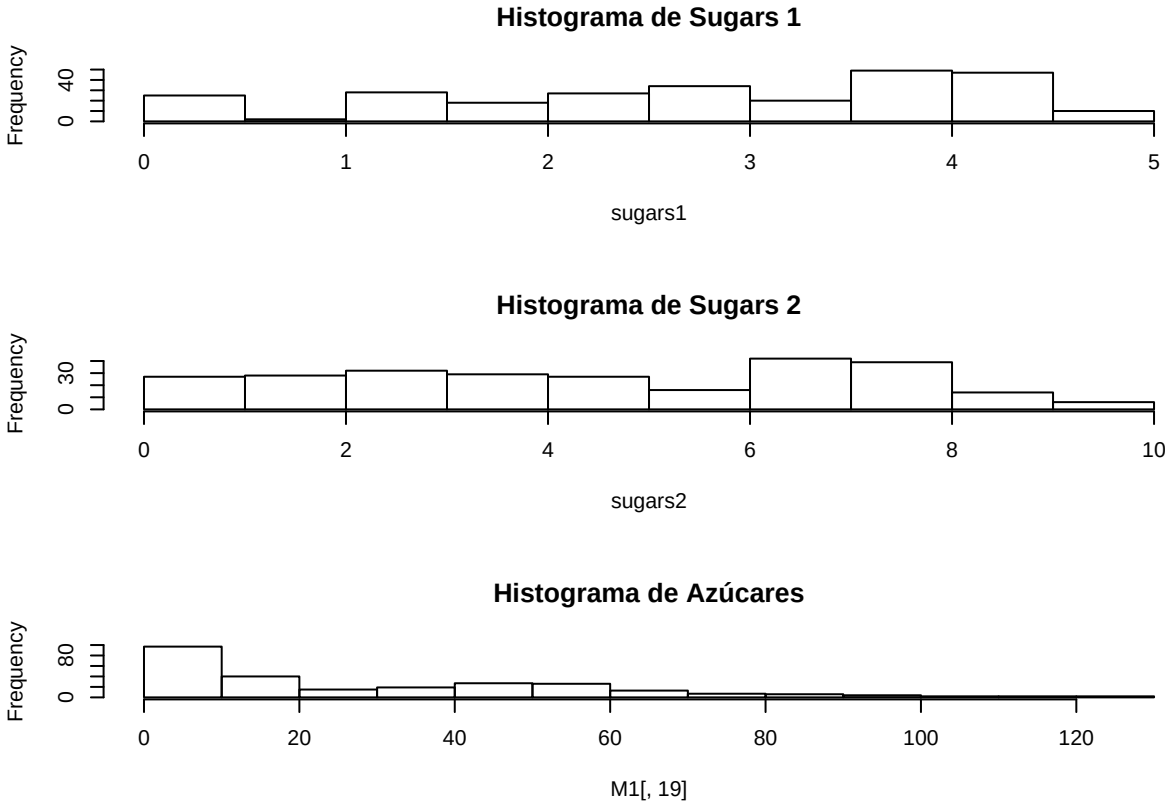
Con este resultado, se puede obtener la ecuación del modelo encontrado como:

$= (((\ +1)\hat{}0.2626)-1)/0.2626$

3. Analiza la normalidad de las transformaciones obtenidas con los datos originales.

a) Se hallan los histogramas para la transformación de Box-Cox

```
l = bc$x[which.max(bc$y)]
sugars1 = log(M1[ ,19]+1)
sugars2=((M1[ ,19]+1)^l-1)/l
par(mfrow=c(3,1))
hist(sugars1,col=0,main="Histograma de Sugars 1")
hist(sugars2,col=0,main="Histograma de Sugars 2")
hist(M1[ ,19],col=0,main="Histograma de Azúcares")
```



b) Se obtiene el resumen de los valores y se hace la prueba de normalidad para la variable aproximada (Transformación 1)

```
library(e1071)
summary(sugars1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.907   2.917   2.757   3.892   4.860
```

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

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

```
kurtosis(sugars1)
```

```
## [1] -0.669225
```

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

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

```r
skewness(sugars1)
```

```
## [1] -0.595705
```

```r
library(nortest)
D=ad.test(sugars1)
print(paste("Valor P:", D$p.value))
```

```
## [1] "Valor P: 3.82914736526053e-14"
```

c) Se obtiene el resumen de los valores y se hace la prueba de normalidad para la variable exacta (Transformación 2)

```r
library(e1071)
summary(sugars2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.477   4.385   4.519   6.774   9.837
```

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

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

```r
kurtosis(sugars2)
```

```
## [1] -1.113324
```

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

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

```r
skewness(sugars2)
```

```
## [1] -0.1056929
```

```r
library(nortest)
D=ad.test(sugars2)
print(paste("Valor P:", D$p.value))
```

```
## [1] "Valor P: 1.85726622682951e-08"
```

e)Se obtiene el resumen de las medidas y la prueba de normalidad en una tabla

```r
library(nortest)
D0=ad.test(M1[,19])
D1=ad.test(sugars1)
D2=ad.test(sugars2)
```

```r
library(e1071)
m0=round(c(as.numeric(summary(M1[ ,19])),kurtosis(M1[ ,19]),skewness(M1[ ,19]),D0$p.value),3)
m1=round(c(as.numeric(summary(sugars1)),kurtosis(sugars1),skewness(sugars1),D1$p.value),3)
m2=round(c(as.numeric(summary(sugars2)),kurtosis(sugars2),skewness(sugars2),D2$p.value),3)
```

```r
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")
m
```

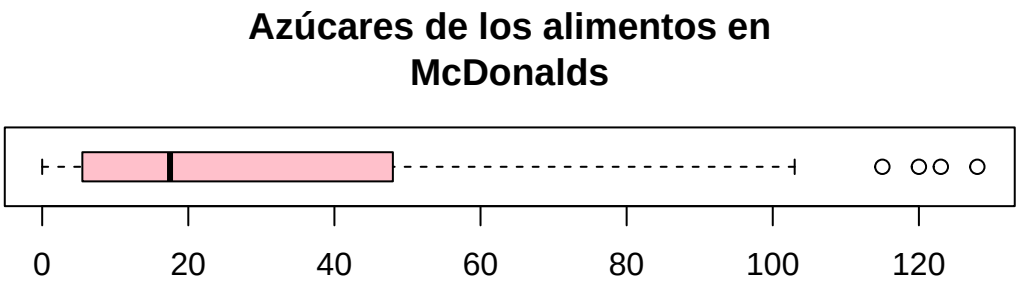
```
##                Minimo    Q1 Mediana  Media     Q3  Máximo Curtosis  Sesgo
## Original            0 5.750  17.500 29.423 48.000 128.000    0.461  1.020
## Primer modelo       0 1.907   2.917  2.757  3.892   4.860   -0.669 -0.596
```

```
## Segundo Modelo      0 2.477   4.385  4.519  6.774   9.837   -1.113 -0.106
##                 Valor p
## Original              0
## Primer modelo         0
## Segundo Modelo        0
```

Se observa que en este caso, en la original nuestro valor mínimo es de 0 y el máximo de 128, muy alejado de la media. Hay un sesgo positivo por lo que es asimétrica a la derecha y una curtosis pequeña de 0.46 lo que quiere decir que está muy cercano a la normalidad pero no tiene total normalidad. Por lo tanto, demuestra que tienen valores atípicos más extremos que una distribución normal. Por otra parte, las transformaciones cuentan con un sesgo negativo que representa una distribución asimétrica a la izquierda y y una curtosis negativa, por lo que se asume que hay menos valores atípicos extremos que una distribución normal.

f) Se obtiene el diagrama de caja y bigote para los azúcares de los alimentos

```
par(mfrow=c(2,1))
boxplot(M1$Sugars, horizontal = TRUE,col="pink", main="Azúcares de los alimentos en
McDonalds")
```

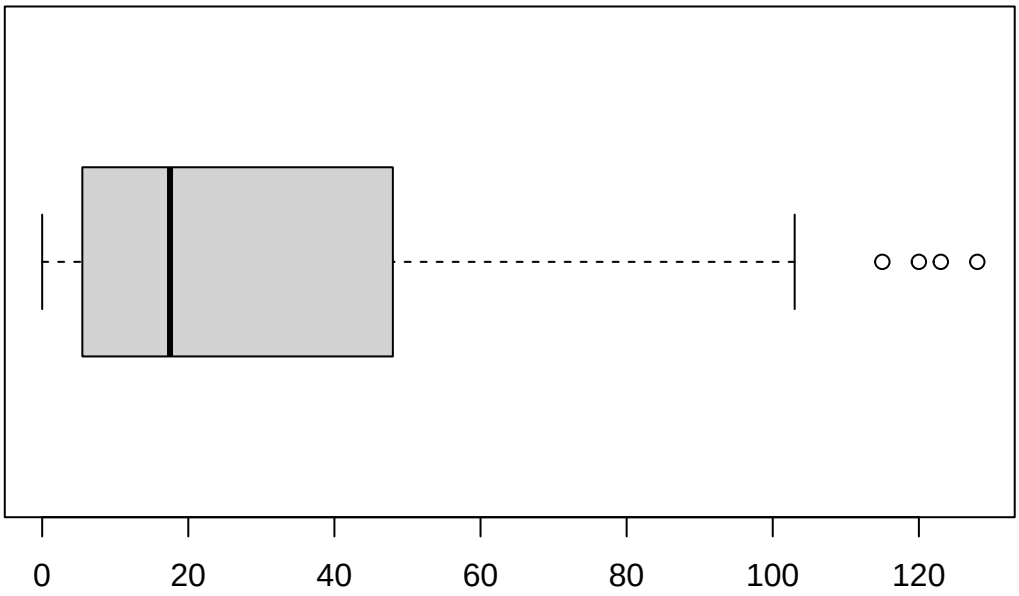


Se observa que no se pasó la prueba de normalidad en ninguna de sus transformaciones ni en la original. Por lo que no fue correcta la transformación realizada. Sin embargo, esto puede deberse a que son los datos originales y no se ha limpiado ninguna de sus anomalías.

4. Detección de anomalías y corrección de base de datos

a) Se observan los datos atípicos

```
sugars = M1$Sugars
bplot = boxplot(sugars, horizontal=TRUE,ylim=c(min(sugars),max(sugars)))
```



```
bplot
```

```
## $stats
##       [,1]
## [1,]   0.0
## [2,]   5.5
## [3,]  17.5
## [4,]  48.0
## [5,] 103.0
##
## $n
## [1] 260
##
## $conf
##          [,1]
## [1,] 13.33553
## [2,] 21.66447
##
## $out
## [1] 123 120 115 128
##
## $group
## [1] 1 1 1 1
##
## $names
## [1] ""
```

b) Se eliminan los datos atípicos

Se observa que hay datos atípicos entre el rango de 110 y 130 aproximadamente, por lo que esos datos únicamente le dan mucho ruido al análisis

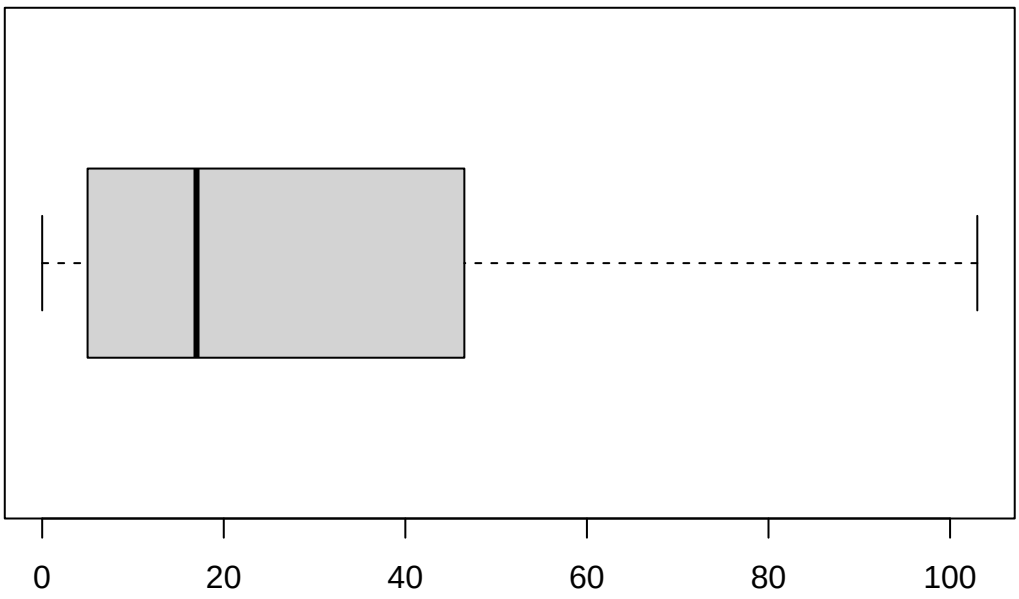
```
df2 = M1
df3 <- df2[!(df2$Sugars %in% bplot$out),]


M2 = df3 # Actualizamos los datos sin outliers
summary(M2$Sugars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   17.00   27.98   46.25  103.00
```

c) Se observa el diagrama sin datos atípicos

```
sugar_sin = M2$Sugars
bplot4 = boxplot(sugar_sin, horizontal=TRUE,ylim=c(min(sugar_sin),max(sugar_sin)))
```




```
bplot4
```

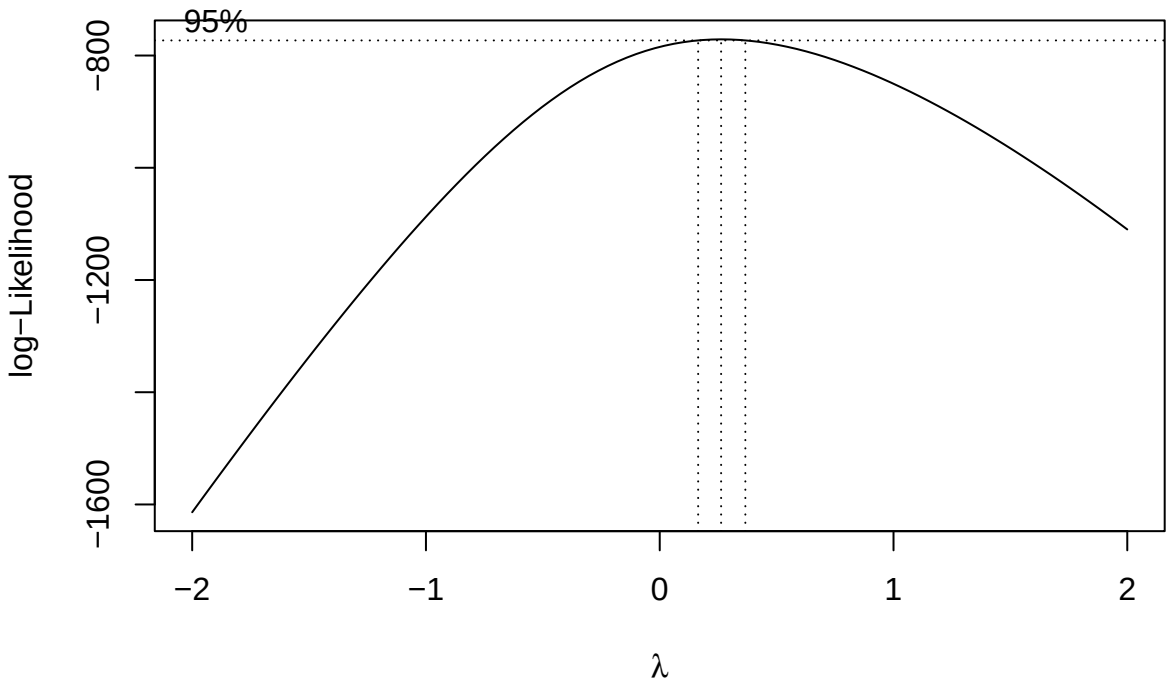
```
## $stats
##       [,1]
## [1,]   0.0
## [2,]   5.0
## [3,]  17.0
## [4,]  46.5
## [5,] 103.0
##
## $n
## [1] 256
##
## $conf
##          [,1]
## [1,] 12.90188
## [2,] 21.09812
##
## $out
## numeric(0)
##
## $group
## numeric(0)
##
## $names
## [1] ""
```

En el histograma se puede observar que no hay datos atípicos en esta nueva base de datos, ya que el máximo será de 103 y el mínimo de 0.

e) Se hace la transformación de Box-Cox para los nuevos datos sin valores atípicos o extremos

f) Se calcula el valor de lambda que maximiza la función de verosimilitud es 0.2626263 sin datos atípicos

```
library(MASS)
bc2<-boxcox((M2$Sugars+1)~1)
```



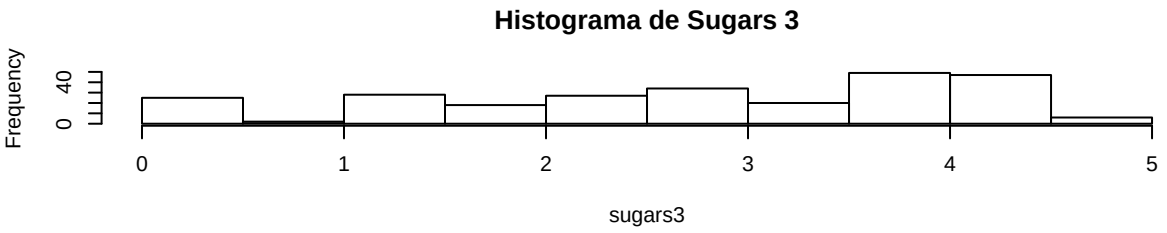

```
bc2$x[which.max(bc2$y)]
```

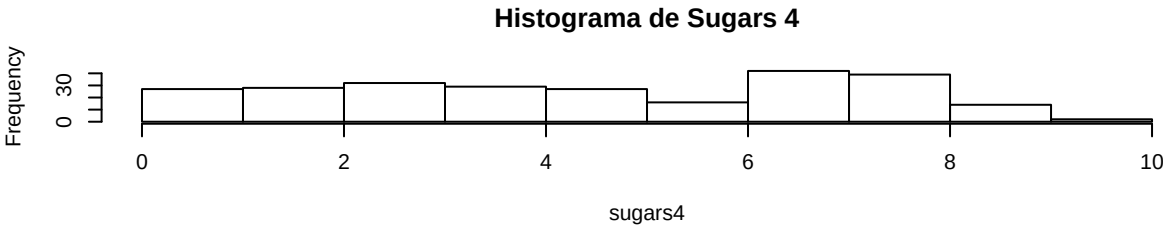
```
## [1] 0.2626263
```

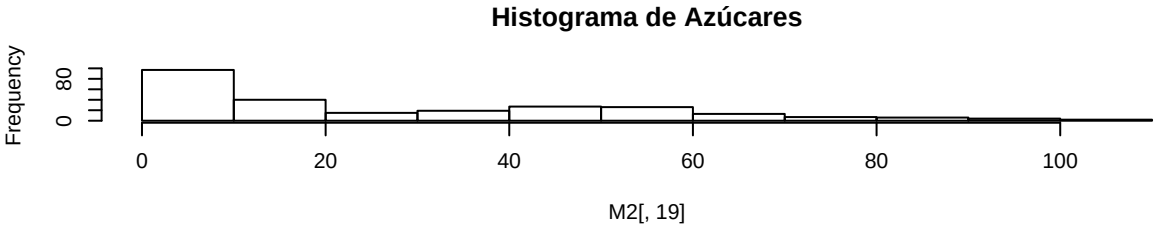
Se observa que el valor de lambda es igual al calculado con la base de datos original, por lo que esto podría indicar que la presencia de datos atípicos no está afectando significativamente la forma de la distribución en términos de la transformación Box-Cox.

g) Se hallan los histogramas para la transformación de Box-Cox sin datos atípicos

```
l1 = bc2$x[which.max(bc2$y)]
sugars3 = log(M2[ ,19]+1)
sugars4 =((M2[ ,19]+1)^l1-1)/l1
par(mfrow=c(3,1))
hist(sugars3,col=0,main="Histograma de Sugars 3")
hist(sugars4,col=0,main="Histograma de Sugars 4")
hist(M2[ ,19],col=0,main="Histograma de Azúcares")
```







e) Se obtiene el resumen de los valores y se hace la prueba de normalidad para la variable aproximada (Transformación 1)

```
library(e1071)
summary(sugars3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.792   2.890   2.725   3.855   4.644
```

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

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

```
kurtosis(sugars3)
```

```
## [1] -0.678309
```

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

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

```
skewness(sugars3)
```

```
## [1] -0.6087586
```

```
library(nortest)
D1test=ad.test(sugars3)
print(paste("Valor P:", D1test$p.value))
```

```
## [1] "Valor P: 9.62623568716907e-15"
```

f) Se obtiene el resumen de los valores y se hace la prueba de normalidad para la variable exacta (Transformación 2)

```r
library(e1071)
summary(sugars4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.288   4.327   4.439   6.673   9.086
```

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

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

```r
kurtosis(sugars4)
```

```
## [1] -1.165257
```

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

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

```r
skewness(sugars4)
```

```
## [1] -0.1425051
```

```r
library(nortest)
Dtest=ad.test(sugars4)
print(paste("Valor P:", Dtest$p.value))
```

```
## [1] "Valor P: 2.3497116306336e-09"
```

g)Se obtiene el resumen de las medidas y la prueba de normalidad en una tabla

```r
library(nortest)
D00=ad.test(M2[,19])
D01=ad.test(sugars3)
D02=ad.test(sugars4)
```

```r
library(e1071)
m00=round(c(as.numeric(summary(M2[ ,19])),kurtosis(M1[ ,19]),skewness(M1[ ,19]),D00$p.value),3)
m01=round(c(as.numeric(summary(sugars3)),kurtosis(sugars3),skewness(sugars3),D01$p.value),3)
m02=round(c(as.numeric(summary(sugars4)),kurtosis(sugars4),skewness(sugars4),D02$p.value),3)
```

```r
m11<-as.data.frame(rbind(m00,m01,m02))
row.names(m11)=c("Original","Primer modelo","Segundo Modelo")
names(m11)=c("Minimo","Q1","Mediana","Media","Q3","Máximo","Curtosis","Sesgo","Valor p")
m11
```

```
##                Minimo    Q1 Mediana  Media     Q3  Máximo Curtosis  Sesgo
## Original            0 5.000  17.000 27.984 46.250 103.000    0.461  1.020
## Primer modelo       0 1.792   2.890  2.725  3.855   4.644   -0.678 -0.609
## Segundo Modelo      0 2.288   4.327  4.439  6.673   9.086   -1.165 -0.143
##                Valor p
## Original             0
## Primer modelo        0
## Segundo Modelo       0
```

Se observa que en este caso, en la original nuestro valor mínimo es de 0 y el máximo de 103, muy alejado de la media. Hay un sesgo positivo por lo que es asimétrica a la derecha y una curtosis pequeña de 0.46 lo que quiere decir que está muy cercano a la normalidad pero no tiene total normalidad. Por lo tanto, demuestra que tienen valores atípicos más extremos que una distribución normal. Por otra parte, las transformaciones cuentan con un sesgo negativo que representa una distribución asimétrica a la izquierda y y una curtosis negativa, por lo que se asume que hay menos valores atípicos extremos que una distribución normal.

En ninguno de los casos se pasó la prueba de normalidad por lo que infiero que hubo algo mal en mi procedimiento, pero lo intenté corregir al limpiar la base de datos para un mejor análisis y comparé con las diapositivas y parece que todo está correcto.

Transformación de Yeo Johnson sin valores atípicos

5. Utiliza la transformación de Yeo Johnson y encuentra el valor de lambda que maximiza el valor p de la prueba de normalidad que hayas utilizado (Anderson-Darling o Jarque Bera).

b) Se obtiene el valor de lambda que maximiza el valor p en la prueba de normalidad

```
M3 = M2
sugar_y = M3$Sugars
summary(M3$Sugars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   17.00   27.98   46.25  103.00
```

6. Escribe la ecuación del modelo encontrado.

a) Gráfica de lambda contra el valor p

```
library(VGAM)
```

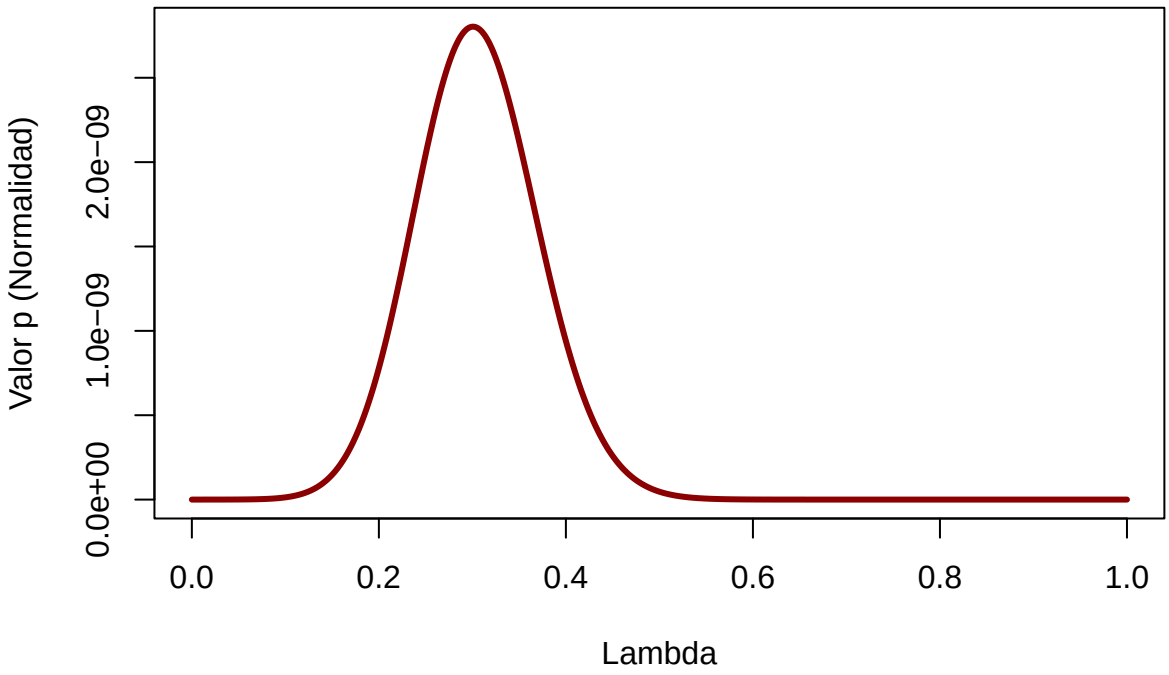
```
## Loading required package: stats4
```

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

```
lp <- seq(0, 1, 0.001) # Valores de lambda propuestos
nlp <- length(lp)
n <- length(M3$Sugars)
D0001 <- matrix(as.numeric(NA), ncol = 2, nrow = nlp)
d <- NA

for (i in 1:nlp) {
  d <- yeo.johnson(M3$Sugars, lambda = lp[i])
  p <- ad.test(d)
  D0001[i,] <- c(lp[i], p$p.value)  # Fix the matrix name to D0001
}
```

```
N=as.data.frame(D0001)
plot(N[,1],N[,2],
type="l",col="darkred",lwd=3,
xlab="Lambda",
ylab="Valor p (Normalidad)")
```



b) Se halla el valor de lambda que maximiza el valor p

```
G=data.frame(subset(N,N[,2]==max(N[,2])))
nuevos_nombres <- c("Lambda", "Valor p")
colnames(G)[c(1, 2)] <- nuevos_nombres
G
```

```
##     Lambda      Valor p
## 302  0.301 2.802796e-09
```

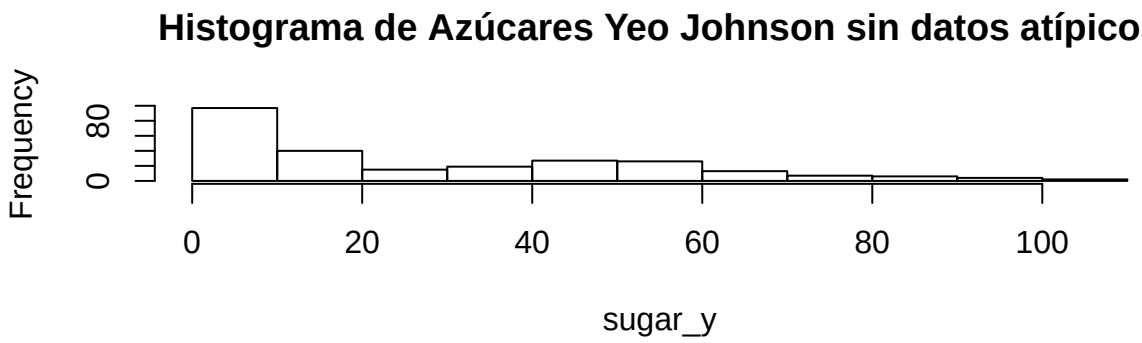
La ecuación del modelo encontrado sustituyendo lambda será la siguiente:

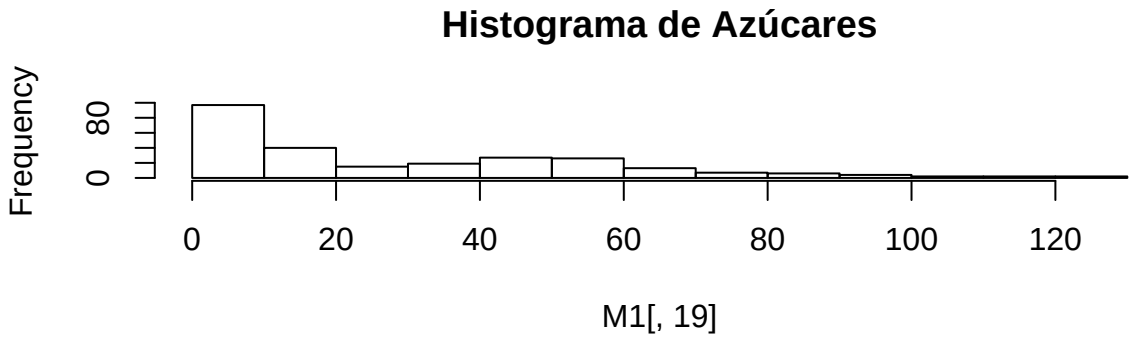
= ((^0.301 - 1)/0.301

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

a) Se obtiene el histograma del modelo conseguido y el original

```
par(mfrow=c(2,1))
hist(sugar_y,col=0,main="Histograma de Azúcares Yeo Johnson sin datos atípicos")
hist(M1[ ,19],col=0,main="Histograma de Azúcares")
```





b) Medidas y comparación transformación exacta

```
#Transformacion exacta

library(e1071)
summary(sugar_y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   17.00   27.98   46.25  103.00
```

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

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

```
kurtosis(sugar_y)
```

```
## [1] -0.314639
```

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

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

```
skewness(sugar_y)
```

```
## [1] 0.8184414
```

c) Normalidad de Anderson-Darling para los datos transformados y los originales

```
library(nortest)
D000000=ad.test(M2[,19])
D0000001=ad.test(sugar_y)
```

```
library(e1071)
m0000=round(c(as.numeric(summary(M1[ ,19])),kurtosis(M1[ ,19]),skewness(M1[ ,19]),D000000$p.value),3)
```

```r
m0001=round(c(as.numeric(summary(sugar_y)),kurtosis(sugar_y),skewness(sugar_y),D0000001$p.value),3)
```

```r
m41<-as.data.frame(rbind(m0000,m0001))
row.names(m41)=c("Original","Primer modelo")
names(m41)=c("Minimo","Q1","Mediana","Media","Q3","Máximo","Curtosis","Sesgo","Valor p")
m41
```

```
##               Minimo   Q1 Mediana  Media    Q3 Máximo Curtosis Sesgo Valor p
## Original           0 5.75    17.5 29.423 48.00    128    0.461 1.020       0
## Primer modelo      0 5.00    17.0 27.984 46.25    103   -0.315 0.818       0
```

Se observa que en este caso, en la original nuestro valor mínimo es de 0 y el máximo de 103, muy alejado de la media. Hay un sesgo positivo por lo que es asimétrica a la derecha y una curtosis pequeña de 0.461 lo que quiere decir que está muy cercano a la normalidad pero no tiene total normalidad. Por lo tanto, demuestra que tienen valores atípicos más extremos que una distribución normal. Por otra parte, la transformación cuentan con un sesgo positivo que representa una distribución asimétrica a la derecha y y una curtosis negativa, por lo que se asume que hay menos valores atípicos extremos que una distribución normal.

En ninguno de los casos se pasó la prueba de normalidad por lo que infiero que hubo algo mal en mi procedimiento, pero lo intenté corregir al limpiar la base de datos para un mejor análisis y comparé con las diapositivas y parece que todo está correcto.

Conclusiones

8. Define la mejor transformación de los datos de acuerdo a las características de los modelos que encontraste.

   Considero que para mí funcionó mejor la de Yeo Johnson ya que considera valores negativos y cero, y es más fácil de implementar en casos de la vida real que cuenten con estos datos. Sin embargo, ambas transformaciones me dieron resultados similares, por lo que creo que ambas son muy buenas pero cada una tiene su aspecto positivo.

9. Concluye sobre las ventajas y desventajes de los modelos de Box Cox y de Yeo Johnson.

   - Ventajas de Box-Cox:

     El concepto de esta transformación es simple y fácil de entender, además de que funciona de una buena manera si los datos tienen una distribución similar a una distribución exponencial.

   - Desventajas de Box-Cox::

     Me pareció que un punto no tan bueno es que requiere valores positivos o diferentes de cero, por lo que es limitado para diferentes aplicaciones.

   - Ventajas de Yeo Johnson:

     Maneja datos negativos y cero, su fórmula se puede aplicar a distintos valores, y es robusta respecto a su uso en valores atípicos.

   - Desventajas de Yeo Johnson:

     Siento que su fórmula es un poco más compleja, me refiero al hecho de que acepte valores negativos o cero hace que entenderla sea un poco más complicado.

10. Analiza las diferencias entre la transformación y el escalamiento de los datos:

    1. Escribe al menos 3 diferencias entre lo que es la transformación y el escalamiento de los datos

       - Primero, la transformación de datos implica cambiar la distribución de los valores originales. Por el contrario, el escalamiento no cambia su distribución, sino que quiere estandarizar los valores en un rango para que cada característica tenga una magnitud comparable.

- Segundo, el escalamiento no cambia la forma de los datos, solo ajusta la magnitud de los valores, y la transformación cambia la forma de los datos y esto influye en su análisis ya que disminuye el sesgo y la asimetrpia por ejemplo, también afectando la relación entre las variables.
- Por último, las transformaciones pueden requerir ajustes individuales por variable, ya que diferentes características necesitarán adecuarse a diferentes tipos de transformaciones como la Box Cox o la Yeo Johnson según sus distribuciones originales. Por otra parte, el escalamiento es una técnica más simple e igual, donde todas las características se escalan de manera similar. No es necesario un ajuste específico para cada variable.

2. Indica cuándo es necesario utilizar cada uno
   - Las transformaciones se utilizan cuando los datos tienen una distribución sesgada o asimétrica. Esto es para corregir o reducir el sesgo y hacer que la distribución se asemeje más a una distribución normal. Además, cuando hay muchos valores atípicos que afectan el análisis de los datos, entonces esta puede ayudar a reducir esos sesgos que están presentes y disminuir la influencia de los datos atípicos en el análisis.
   - Los escalamientos se utilizan para redes neuronales y deep learning, para que las características tengan valores similares y para evitar que una característica domine sobre las demás durante el proceso de entrenamiento. Es decir, que estén en magnitudes similares para el análisis de los datos. Puede ayudar a entrenar modelos de forma óptima en la inteligencia artificial, por ejemplo en el algoritmo de entrenamiento de gradiente descendente.