# Actividad 5

Oskar Arturo Gamboa Reyes

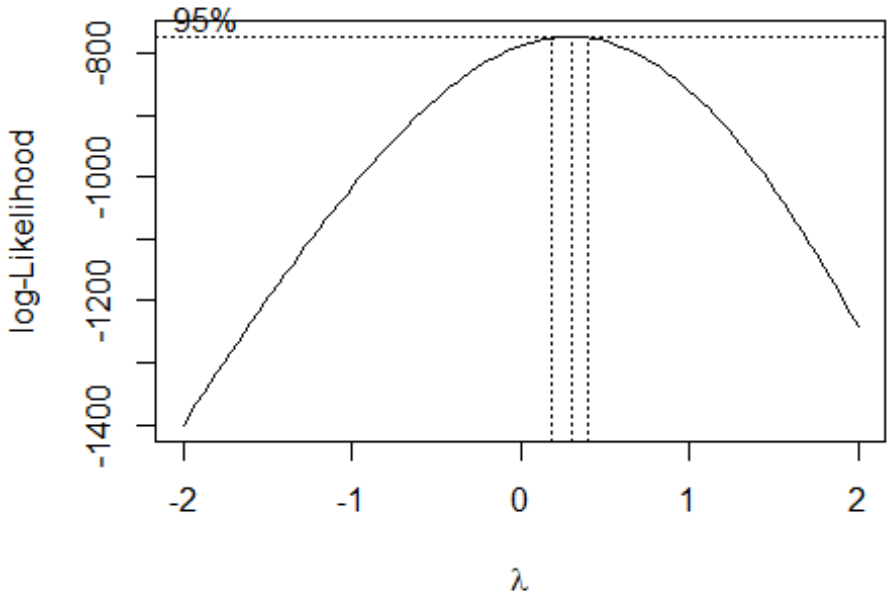
2024-08-14

1. Leer archivo con datos

```
M=read.csv("mc-donalds-menu.csv")

fat = M$Total.Fat
```

2. Transformación con Box-Cut

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




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

## [1] "lambda:  0.303030303030303"
```

Modelos sugeridos a partir de lambda

Aproximado = $\sqrt{x+1}$ Exacto = $\frac{(x+1)^{0.62}-1}{0.62}$

```r
fatM1 = sqrt(fat+1)
fatM2 = (((fat+1)^0.62)-1)/0.62
```

Comparación de variables

```r
library(e1071)

print("Original")

## [1] "Original"

summary(fat)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.375  11.000  14.165  22.250 118.000

print("Curtosis")

## [1] "Curtosis"

kurtosis(fat)

## [1] 10.35171

print("Sesgo")

## [1] "Sesgo"

skewness(fat)

## [1] 2.128023

print("Aproximación")

## [1] "Aproximación"

summary(fatM1)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.836   3.464   3.450   4.822  10.909

print("Curtosis")

## [1] "Curtosis"

kurtosis(fatM1)

## [1] -0.08053187

print("Sesgo")

## [1] "Sesgo"

skewness(fatM1)
```

```
## [1] 0.3078819
```

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

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

```
summary(fatM2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   1.814   5.915   6.195   9.731  29.608
```

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

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

```
kurtosis(fatM2)
```

```
## [1] 0.9486195
```

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

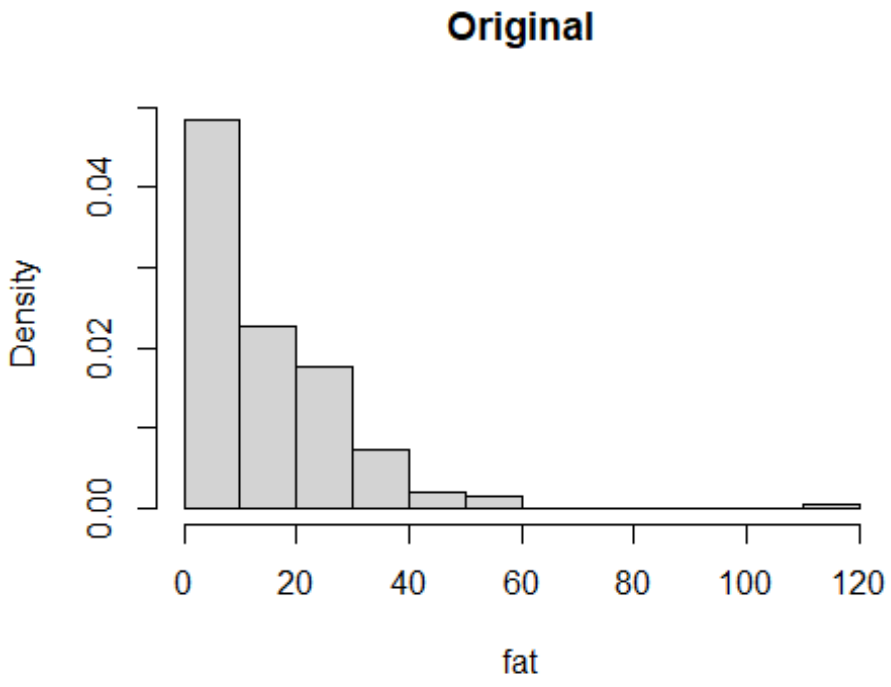
```
## [1] "Sesgo"
```

```
skewness(fatM2)
```

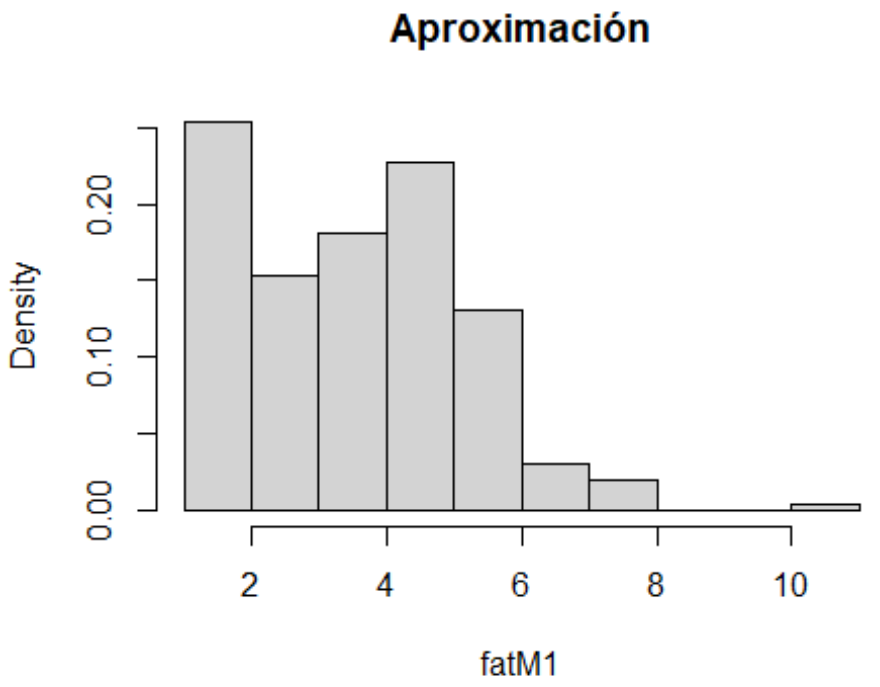
```
## [1] 0.6293498
```

Histogramas

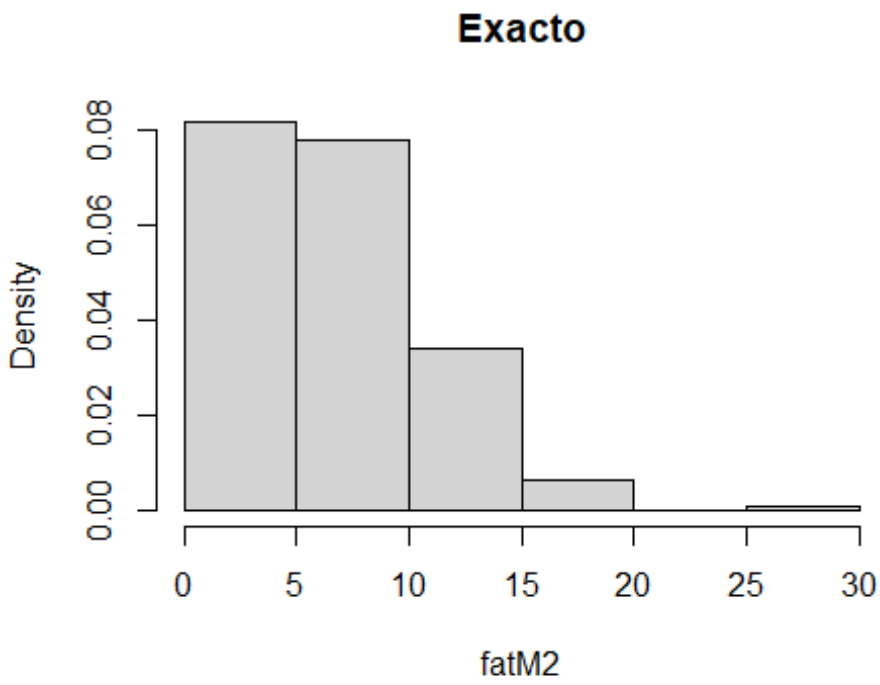
```
hist(fat,freq=FALSE, main="Original")
```



```r
hist(fatM1,freq=FALSE, main="Aproximación")
```




```r
hist(fatM2,freq=FALSE, main="Exacto")
```



Pruebas de Normalidad

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

```
##
##  Anderson-Darling normality test
##
## data:  fat
## A = 6.7424, p-value < 2.2e-16
```

```
ad.test(fatM1)
```

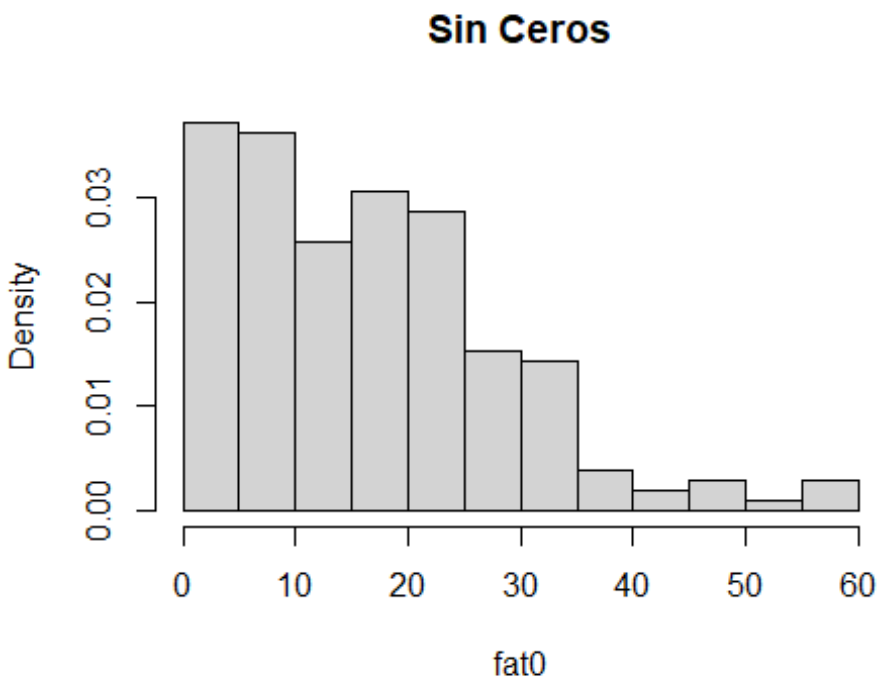
```
##
##  Anderson-Darling normality test
##
## data:  fatM1
## A = 3.9624, p-value = 6.861e-10
```

```
ad.test(fatM2)
```

```
##
##  Anderson-Darling normality test
##
## data:  fatM2
## A = 3.6333, p-value = 4.31e-09
```

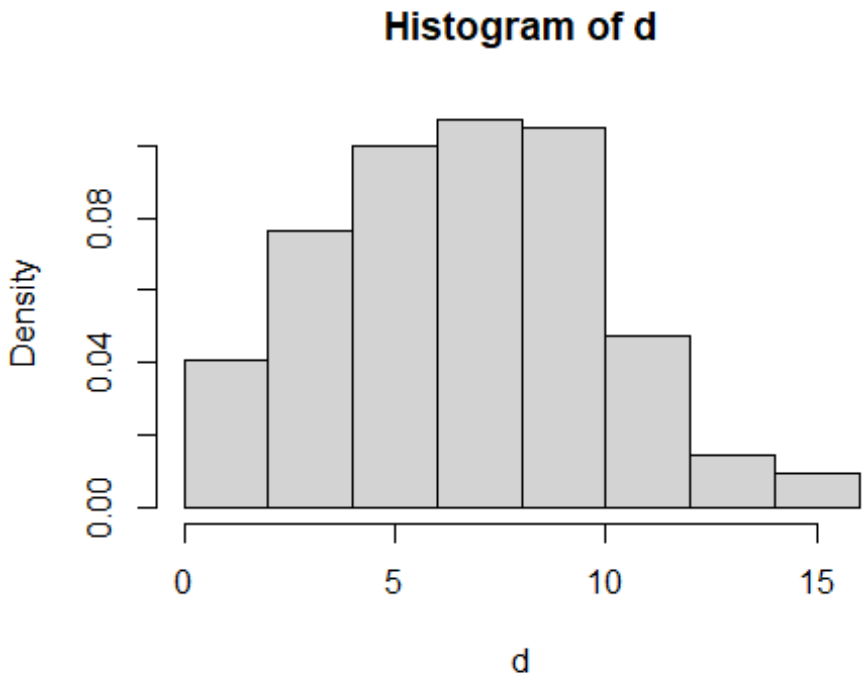
Quitando ceros de los datos originales y dato atípico.

```
fat0=subset(fat,fat>0 & fat<100)
hist(fat0,freq=FALSE, main="Sin Ceros")
```




Lambda máximo con modelo Yeo Johnson sin ceros

```
library(VGAM)

## Loading required package: stats4

## Loading required package: splines

library(e1071)
lp <- seq(0,1,0.001) # Valores de lambda propuestos
nlp <- length(lp)
n=length(fat0)
D <- matrix(as.numeric(NA),ncol=2,nrow=nlp)
d <-NA
for (i in 1:nlp){
  d= yeo.johnson(fat0, lambda = lp[i])
  p=ad.test(d)
  D[i,]=c(lp[i],p$p.value)
}
N=as.data.frame(D)
G=data.frame(subset(N,N$V2==max(N$V2)))
print(paste("Lambda con mayor p-value",G$V1))

## [1] "Lambda con mayor p-value 0.546"

d= yeo.johnson(fat0, lambda = G$V1)
hist(d,freq=FALSE)
```




```
ad.test(d)

##
##  Anderson-Darling normality test
##
## data:  d
## A = 0.63888, p-value = 0.09449
```

Modelo encontrado

Aproximado = $\sqrt{x+1}$ Exacto = $\frac{(x+1)^{0.54}-1}{0.54}$

```
fatM3 = sqrt(fat0+1)
fatM4 = (((fat0+1)^0.54)-1)/0.54

ad.test(fat0)

##
##  Anderson-Darling normality test
##
## data:  fat0
## A = 2.5838, p-value = 1.537e-06

ad.test(fatM3)

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

```
##
## data:  fatM3
## A = 0.66147, p-value = 0.08307
```

```
ad.test(fatM4)
```

```
##
##  Anderson-Darling normality test
##
## data:  fatM4
## A = 0.63926, p-value = 0.09429
```

```
library(e1071)
```

```
print("Original sin Ceros")
```

```
## [1] "Original sin Ceros"
```

```
summary(fat0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.50    8.00   16.00   16.98   23.00   60.00
```

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

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

```
kurtosis(fat0)
```

```
## [1] 1.118341
```

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

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

```
skewness(fat0)
```

```
## [1] 0.9648949
```

```
library(e1071)
```

```
print("Aproximación")
```

```
## [1] "Aproximación"
```

```
summary(fatM3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.225   3.000   4.123   3.987   4.899   7.810
```

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

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

```r
kurtosis(fatM3)
```

```
## [1] -0.3883092
```

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

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

```r
skewness(fatM3)
```

```
## [1] 0.0693232
```

```r
library(e1071)
```

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

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

```r
summary(fatM4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4533  4.2141  6.6998  6.4436  8.4501 15.1966
```

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

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

```r
kurtosis(fatM4)
```

```
## [1] -0.3463189
```

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

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

```r
skewness(fatM4)
```

```
## [1] 0.1417924
```

A partir de las pruebas de normalidad de Anderson-Darling podemos determinar que la mejor manera de normalizar una variable, es quitando los datos atípicos y ceros, haciendo el metodo de Yeo-Johnson exacto, ya que esto resulto en el mayor p-value. El mejor resultado es un p-value con valor de 0.09, al ser mayor a 0.05 nos indica que ya tiene una distribución normal.

La mayor diferencia entre el modelo Yeo-Johnson y Box-Cox es que el primero te permite trabajar con 0 y datos negativos, mientras que Box-Cox no tiene en consideración esta cualidad de los datos.

La transformación de datos modifica la distribución de los datos, esto nos permite hacer un analisis de datos con mayor facilidad ya que una curva normal tiene muchas propiedades que ya conocemos. Mientras que el escalamiento solo modifica la escala, por lo que nos

permite comparar dos variables que tienen similar distribución pero escalas parecidas, simplemente no cambia las proporciones basicas de la variable.