# Transformaciones

Arturo

2023-09-02

**Autor: Arturo Garza Campuzano**

**Matrícula: A00828096**

## Transformaciones

### Importar módulos

```
# Instalacion y carga de paquetes
if (!require(VGAM) || !require(e1071) || !require(nortest) || !require(MASS)) {
  install.packages("VGAM")
  install.packages("e1071")
  install.packages('nortest')
  install.packages("MASS")
}
```

```
## Loading required package: VGAM
```

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

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

```
## Loading required package: e1071
```

```
## Loading required package: nortest
```

```
## Loading required package: MASS
```

```
library(VGAM)
library(e1071)
library(nortest)
library(MASS)
```

### Cargar datos

La variable seleccionada para la transformación ha sido *Carbohydrates*.

```
# Lectura del archivo (cambiar routeo si es necesario)
M = read.csv("mc_donalds_menu_1.csv")

# Variable seleccionada para la transformación
carbohidratos <- M$Carbohydrates
```

## Transformación

**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.**

La **transformación Box-Cox** es una familia de funciones indexadas por el parámetro $\lambda$. Considerando que en este caso $\lambda \neq 0$, se tiene que:

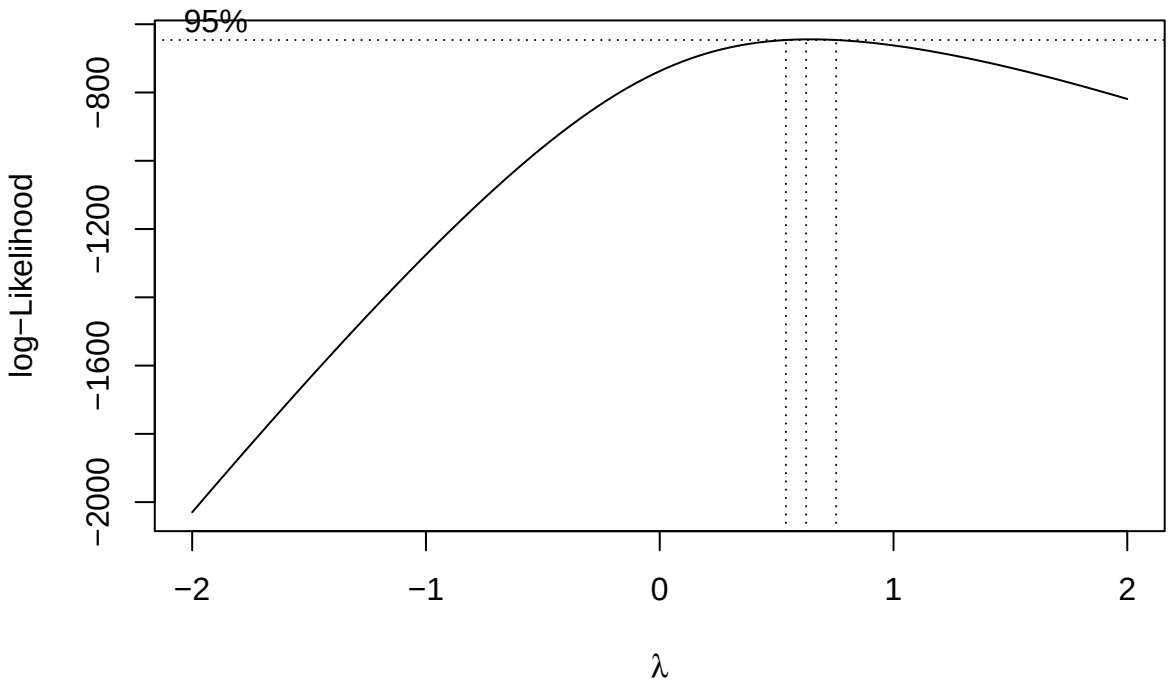
$$f(x, \lambda) = \frac{x^\lambda - 1}{\lambda}$$

A continuación se presenta la tabla de las transformaciones estimadas para una determinada $\lambda$:

| $\lambda$ | Transformación |
|---|---|
| -2 | $\frac{1}{x^2}$ |
| -1 | $\frac{1}{x}$ |
| -0.5 | $\frac{1}{\sqrt{x}}$ |
| 0 | $log(x)$ |
| 0.5 | $\sqrt{x}$ |
| 1 | $x$ |
| 2 | $x^2$ |

Para utilizar esta transformación es necesario trabajar con valores positivos, es por esto que se hace una traslación de la variable *Carbohydrates*. Para obtener el valor aproximado se utiliza la tabla anterior y el valor de lambda exacto.

```
# Valores de lambda por transformacion
data <- data.frame(
lambda = c(-2, -1, -0.5, 0, 0.5, 1, 2)
)

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



```r
lambda_exacto <- bc$x[which.max(bc$y)]

# Calcular las diferencias entre los valores de lambda y el valor exacto
diferencias <- abs(data$lambda - lambda_exacto)

# Encontrar el valor de lambda con la menor diferencia
lambda_aproximado <- data$lambda[which.min(diferencias)]

cat("Valor de lambda del modelo exacto:", lambda_exacto,"\n")
```

```
## Valor de lambda del modelo exacto: 0.6262626
```

```r
cat("Valor de lambda del modelo aproximado:", lambda_aproximado)
```

```
## Valor de lambda del modelo aproximado: 0.5
```

**2. Escribe las ecuaciones de los modelos encontrados.**

Considerando la traslación de la variable, se tienen las siguientes ecuaciones:

- Modelo exacto: $f(x) = \frac{(x+1)^{0.63}-1}{0.63}$
- Modelo aproximado: $f(x) = \sqrt{x+1}$

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

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

```r
carbohidratos1 <- sqrt(carbohidratos + 1)
carbohidratos2 <- ((carbohidratos + 1)^lambda_exacto - 1) / lambda_exacto

medidas <- function(variable, nombre_variable){
  cat(nombre_variable, "\n")
  print(summary(variable))
  cat("Curtosis:",kurtosis(variable), "\n")
  cat("Sesgo:",skewness(variable), "\n")
}

medidas(carbohidratos1, "Carbohidratos 1")
```

```
## Carbohidratos 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.568   6.708   6.583   7.810  11.916
## Curtosis: 0.90923
## Sesgo: -0.4939626
```

```r
medidas(carbohidratos2, "Carbohidratos 2")
```

```
## Carbohidratos 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.12   15.72   15.67   19.36   33.98
## Curtosis: 0.6381974
## Sesgo: -0.08250202
```

```r
medidas(carbohidratos, "Carbohidratos")
```

```
## Carbohidratos
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

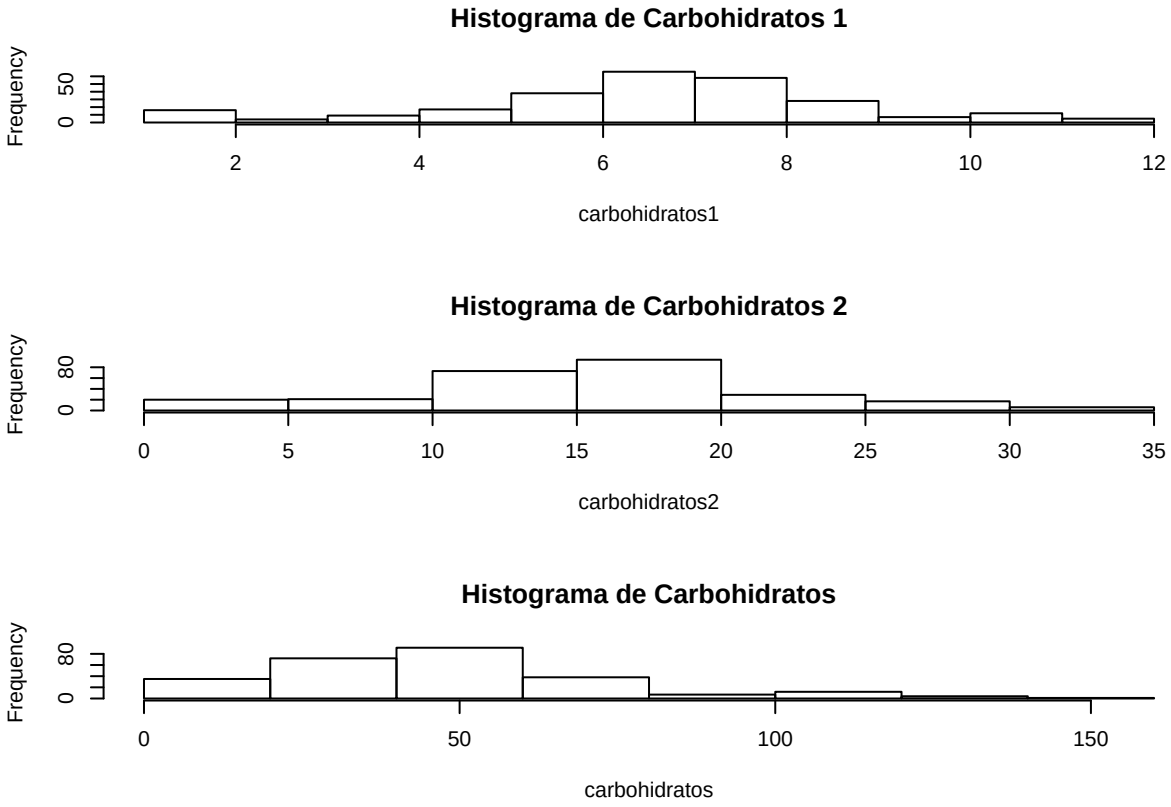
```
##    0.00   30.00   44.00   47.35   60.00  141.00
## Curtosis: 1.324083
## Sesgo: 0.9021952
```

Comparando las medidas se hacen las siguientes observaciones:

- La diferencia entre la mediana y la media es menor en los modelos propuestos que en la variable original.
- El sesgo y la curtosis en los modelos son menores que en la variable original.
- La distribución de los modelos parece ser más simétrica y platicúrtica que la distribución de la variable original.

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

```
par(mfrow=c(3,1))
hist(carbohidratos1,col=0,main="Histograma de Carbohidratos 1")
hist(carbohidratos2,col=0,main="Histograma de Carbohidratos 2")
hist(carbohidratos,col=0,main="Histograma de Carbohidratos")
```




Se puede contemplar que las distribuciones de los modelos propuestos parecen seguir un comportamiento más normal que la distribución de la variable original.

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

Considerando un alfa de 0.05 se realiza la prueba de normalidad de Anderson-Darling para cada modelo.

```
anderson_darling <- function(variable, nombre_variable){
  print(nombre_variable)
  D <- ad.test(variable)
  D$p.value
  cat("Estadística de Anderson-Darling:", D$statistic, "\n")
```

```
  cat("P-valor:", D$p.value, "\n")
  alpha <- 0.05
  if (D$p.value < alpha) {
    cat("La variable parece no seguir una distribución normal.\n")
  } else {
    cat("La variable parece seguir una distribución normal.\n")
  }
}

anderson_darling(carbohidratos1, "Carbohidratos 1")
```

```
## [1] "Carbohidratos 1"
## Estadística de Anderson-Darling: 4.452416
## P-valor: 4.481723e-11
## La variable parece no seguir una distribución normal.
```

```
anderson_darling(carbohidratos2, "Carbohidratos 2")
```

```
## [1] "Carbohidratos 2"
## Estadística de Anderson-Darling: 3.107604
## P-valor: 8.1823e-08
## La variable parece no seguir una distribución normal.
```

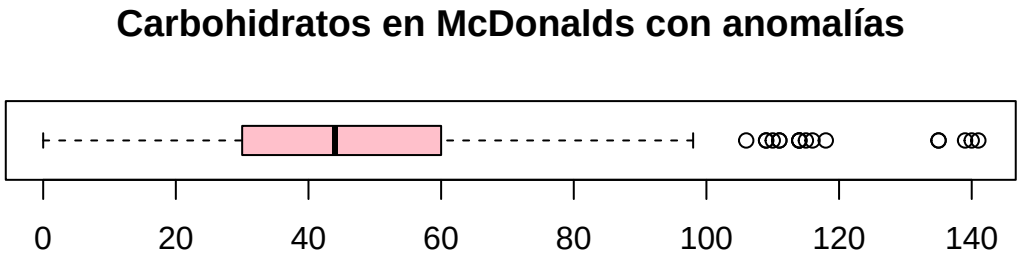
```
anderson_darling(carbohidratos, "Carbohidratos")
```

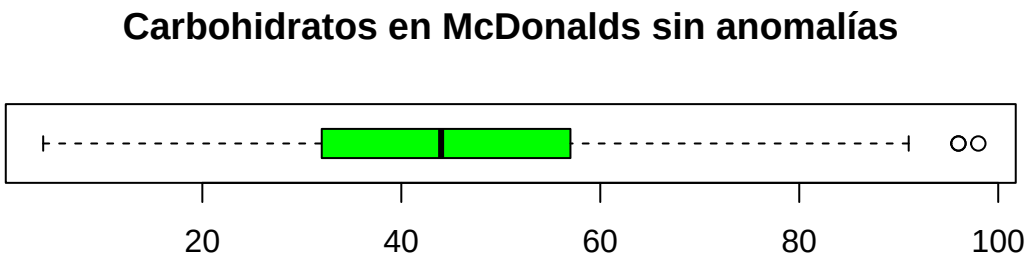
```
## [1] "Carbohidratos"
## Estadística de Anderson-Darling: 4.140224
## P-valor: 2.546548e-10
## La variable parece no seguir una distribución normal.
```

Al parecer, todos los modelos están muy lejos de seguir una distribución normal.

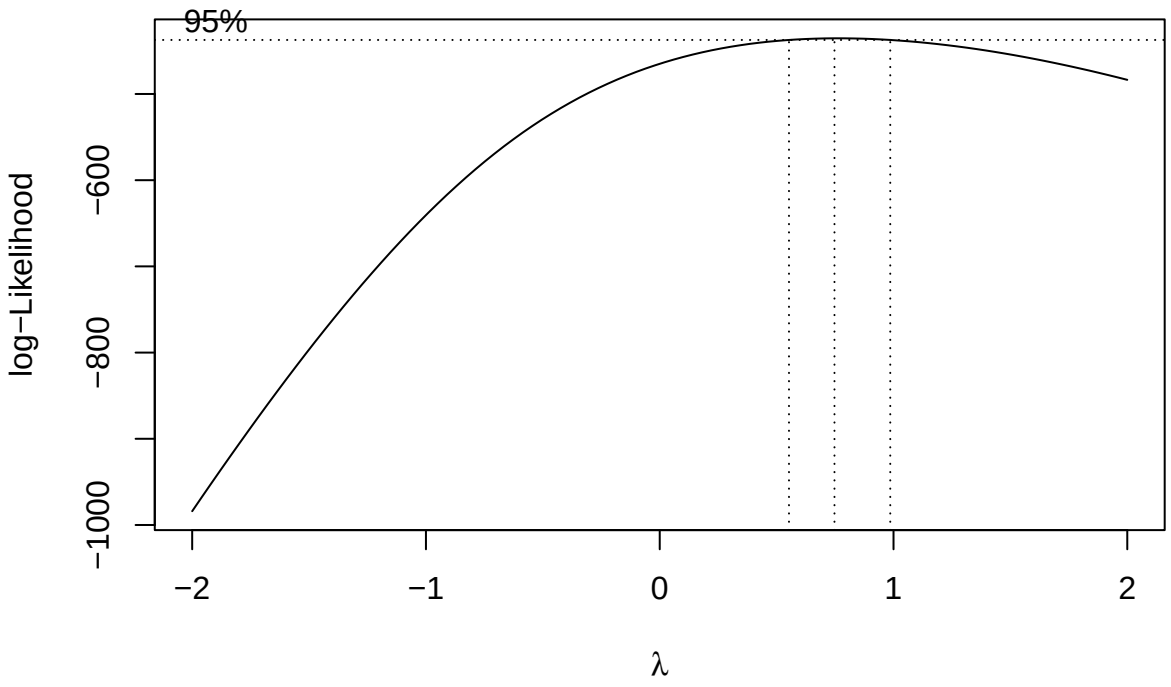
**4. Detecta anomalías y corrige tu base de datos (datos atípicos, ceros anómalos, etc).**

```
q1 <- quantile(carbohidratos, 0.25)
q3 <- quantile(carbohidratos, 0.75)
ri <- q3 - q1

# Filtrar los datos para eliminar valores cero
carbohidratos_filtrados <- carbohidratos[carbohidratos <= q3 + 1.5 * ri & carbohidratos > 0 & carbohidra

# Crear gráficos de boxplot para carbohidratos
par(mfrow=c(2,1))
boxplot(carbohidratos, horizontal = TRUE, col="pink", main="Carbohidratos en McDonalds con anomalías")
boxplot(carbohidratos_filtrados, horizontal = TRUE, col="green", main="Carbohidratos en McDonalds sin a
```




```
# Reconstrucción de modelos
par(mfrow=c(1,1))
nuevo_bc <- boxcox(carbohidratos_filtrados ~ 1)
```




```
nueva_lambda <- nuevo_bc$x[which.max(nuevo_bc$y)]
print(nueva_lambda)
```

```
## [1] 0.7474747
```

```
carbohidratos1_filtrados <- sqrt(carbohidratos_filtrados)
carbohidratos2_filtrados <- ((carbohidratos_filtrados)^nueva_lambda - 1) / nueva_lambda
```

```r
medidas(carbohidratos1_filtrados, "Carbohidratos 1 filtrados")
```

```
## Carbohidratos 1 filtrados
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   5.657   6.633   6.567   7.550   9.899
## Curtosis: 0.3648873
## Sesgo: -0.4208334
```

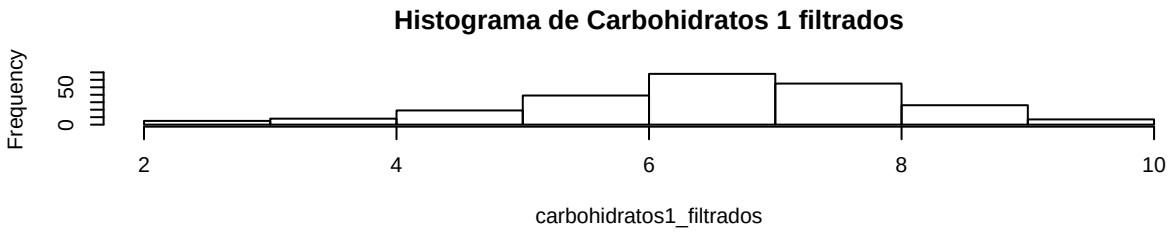
```r
medidas(carbohidratos2_filtrados, "Carbohidratos 2 filtrados")
```

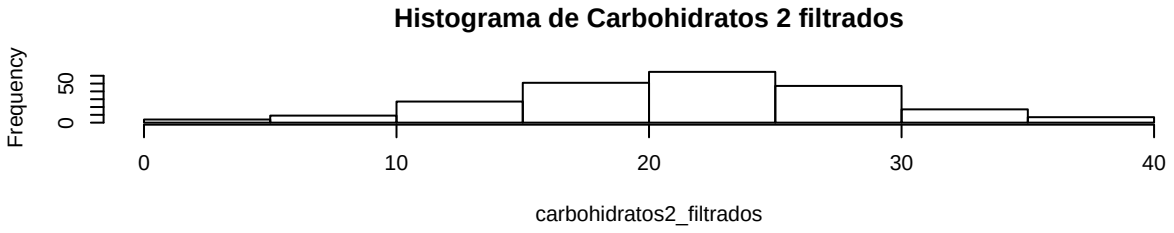
```
## Carbohidratos 2 filtrados
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.433  16.505  21.300  21.382  26.133  39.852
## Curtosis: -0.005745343
## Sesgo: -0.03778588
```

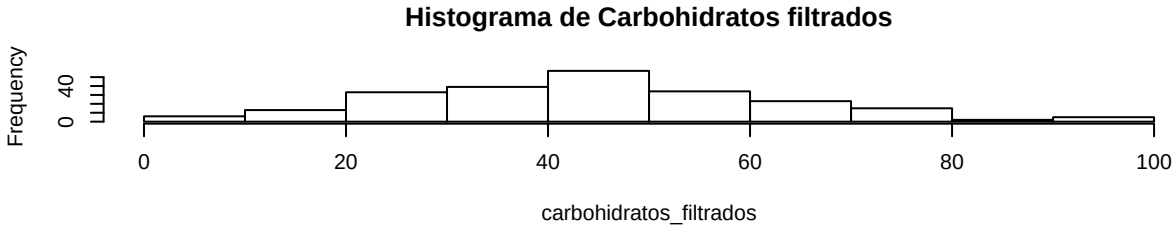
```r
medidas(carbohidratos_filtrados, "Carbohidratos filtrados")
```

```
## Carbohidratos filtrados
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   32.00   44.00   45.26   57.00   98.00
## Curtosis: 0.02586435
## Sesgo: 0.3014074
```

```r
par(mfrow=c(3,1))
hist(carbohidratos1_filtrados,col=0,main="Histograma de Carbohidratos 1 filtrados")
hist(carbohidratos2_filtrados,col=0,main="Histograma de Carbohidratos 2 filtrados")
hist(carbohidratos_filtrados,col=0,main="Histograma de Carbohidratos filtrados")
```



```r
anderson_darling(carbohidratos1_filtrados, "Carbohidratos 1 filtrados")
```

```
## [1] "Carbohidratos 1 filtrados"
## Estadística de Anderson-Darling: 0.751107
## P-valor: 0.04987871
## La variable parece no seguir una distribución normal.
```

```r
anderson_darling(carbohidratos2_filtrados, "Carbohidratos 2 filtrados")
```

```
## [1] "Carbohidratos 2 filtrados"
## Estadística de Anderson-Darling: 0.2261886
## P-valor: 0.8159778
## La variable parece seguir una distribución normal.
```

```r
anderson_darling(carbohidratos_filtrados, "Carbohidratos filtrados")
```

```
## [1] "Carbohidratos filtrados"
## Estadística de Anderson-Darling: 0.3963877
## P-valor: 0.3669956
## La variable parece seguir una distribución normal.
```

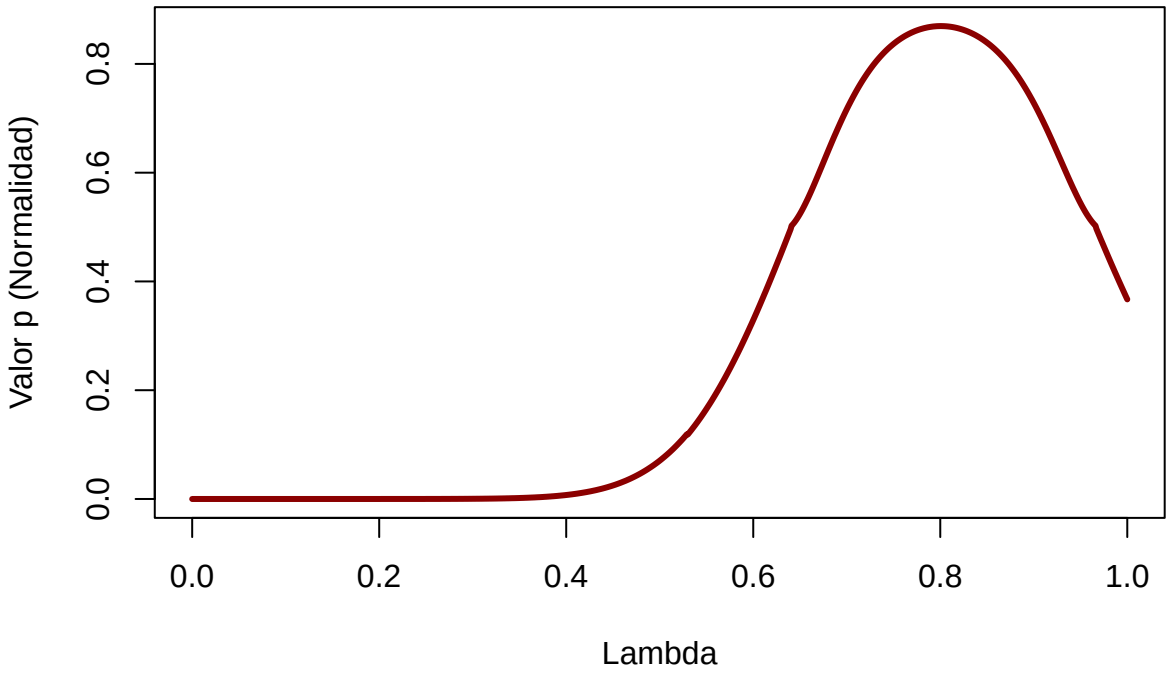
Se eliminaron los **datos atípicos** de la variable original, tomando en cuenta que un **dato atípico** es aquel que esté por debajo de $Q1 - 1.5 \times IRQ$, por encima de $Q3 + 1.5 \times IRQ$ o que su valor sea cero. Después se realizó la transformación Box-Cox con los datos sin anomalías y se hizo una prueba de normalidad sobre los modelos propuestos y la variable original, cuyos resultados fueron los siguientes:

- El modelo aproximado parece no seguir una distribución normal, esto se debe a que el valor de lambda exacto está cerca del límite intermedio entre dos valores (0.5 y 1).
- El modelo exacto parece seguir una distribución normal y tiene el valor de p más cercano a uno.
- La variable original parece seguir una distribución normal y cuenta con un valor p más grande que el modelo aproximado pero mucho menor que el modelo exacto.

**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).**

```r
carbohidratos3 <- yeo.johnson(carbohidratos_filtrados, nueva_lambda)
lp <- seq(0,1,0.001)
nlp <- length(lp)
n <- length(carbohidratos_filtrados)
D <- matrix(as.numeric(NA),ncol=2,nrow=nlp)
d <- NA

# Obtener el valor de lambda que maximiza el valor p en la prueba de normalidad
for (i in 1:nlp){
  d = yeo.johnson(carbohidratos_filtrados, lp[i])
  p = ad.test(d)
  D[i,] = c(lp[i],p$p.value)}

# Grafica de lambda contra valor p
N <- as.data.frame(D)
N <- N[complete.cases(N), ]
minimo_V2 <- min(N$V2)
maximo_V2 <- max(N$V2)
plot(N$V1,N$V2,
     type="l",col="darkred",lwd=3,
     xlab="Lambda",
     ylab="Valor p (Normalidad)",
```

```
    x_limit <- c(0, 1),
    y_limit <- c(minimo_V2, max(N$V2)))
```




```
G <- data.frame(subset(N,N$V2==max(N$V2)))
lambda_max <- G$V1
```

Yeo y Johnson indican que los valores de y deben acotarse. En este caso, como se tiene que $x \geq 0$ y $\lambda \neq 0$ se propone la siguiente transformación:

$$f(x, \lambda) = \frac{(x+1)^{\lambda} - 1}{\lambda}$$

En esta transformación se obtiene el valor de lambda que maximiza el valor p de la prueba de normalidad: 0.801.

**6. Escribe la ecuación del modelo encontrado.**

Haciendo uso de la tabla de transformaciones, se sabe que el modelo aproximado sería equivalente a la variable original.

El modelo exacto está definido por la siguiente ecuación:

$$f(x, \lambda) = \frac{(x+1)^{0.801} - 1}{0.801}$$

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

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

```
carbohidratos3 <- ((carbohidratos_filtrados + 1)^lambda_max - 1) / lambda_max
medidas(carbohidratos3, "Carbohidratos 3")
```

```
## Carbohidratos 3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.283  19.296  25.090  25.307  31.027  48.282
## Curtosis: -0.03363509
## Sesgo: 0.04631274
```

```
medidas(carbohidratos_filtrados, "Carbohidratos filtrados")
```

```
## Carbohidratos filtrados
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   32.00   44.00   45.26   57.00   98.00
## Curtosis: 0.02586435
## Sesgo: 0.3014074
```

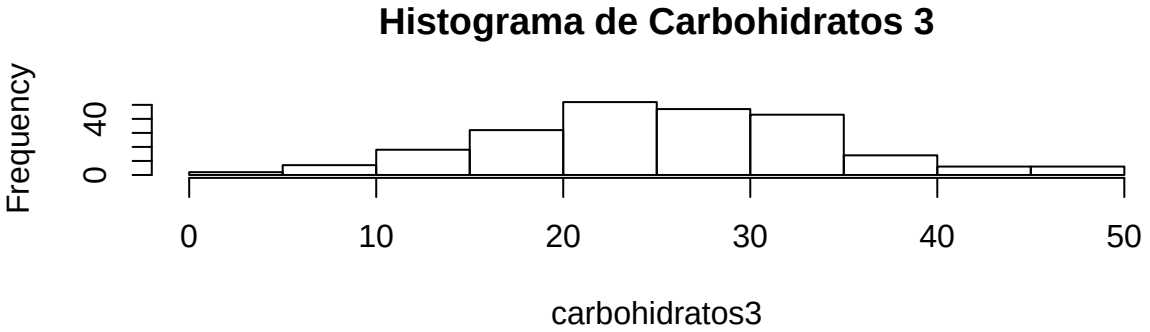
Comparando las medidas se hacen las siguientes observaciones:

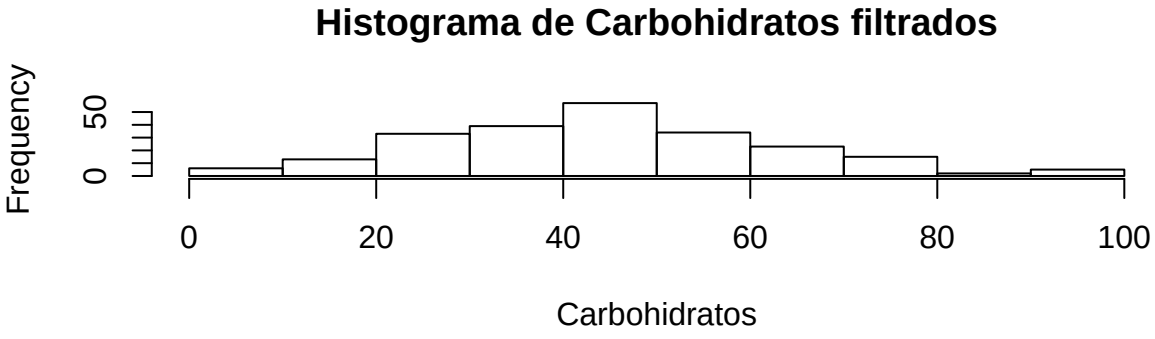
- La diferencia entre la mediana y la media es menor en el modelo propuesto que en la variable original.
- El sesgo y la curtosis en el modelo son menores que en la variable original.
- La distribución del modelo parece ser más simétrica y platicúrtica que la distribución de la variable original.

**7.2. Obtén el histograma de los 2 modelos obtenidos (exacto y aproximado) y los datos originales.**

En este caso sólo se tiene el modelo exacto, ya que el aproximado sería equivalente a la variable original.

```
par(mfrow=c(2,1))
hist(carbohidratos3,col=0,main="Histograma de Carbohidratos 3")
hist(carbohidratos_filtrados,col=0,main="Histograma de Carbohidratos filtrados",xlab="Carbohidratos")
```




Se puede contemplar que la distribución del modelo propuesto parece cargarse más hacia la derecha que la distribución de la variable original.

**7.3. Realiza la prueba de normalidad de Anderson-Darling para los datos transformados y los originales.**

```r
anderson_darling(carbohidratos3, "Carbohidratos 3")
```

```
## [1] "Carbohidratos 3"
## Estadística de Anderson-Darling: 0.2057914
## P-valor: 0.8695641
## La variable parece seguir una distribución normal.
```

```r
anderson_darling(carbohidratos_filtrados, "Carbohidratos filtrados")
```

```
## [1] "Carbohidratos filtrados"
## Estadística de Anderson-Darling: 0.3963877
## P-valor: 0.3669956
## La variable parece seguir una distribución normal.
```

Se puede observar que el valor p del modelo exacto supera el valor p de la variable original y también supera el valor máximo de p alcanzado por el modelo exacto propuesto de la transformación Box-Cox.

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

Si sólo se tomaran en cuenta los resultados de la prueba de normalidad Anderson-Darling, la mejor transformación de los datos sería la transformación Yeo-Johnson, debido a que el modelo obtenido por esta transformación alcanzó el valor de p más cercano al uno.

No obstante, si se considera el sesgo y la curtosis parece ser que la mejor transformación de los datos sería la transformación Box-Cox, ya que su modelo exacto cuenta con más simetría y una distribución más mesocúrtica. Además, la transformación Yeo-Jhonson depende de la transformación Box-Cox para obtener un buen resultado.

Por lo tanto, se puede considerar que la mejor transformación es la transformación Box-Cox debido a su independencia y sus resultados más normales en cuanto a sesgo y curtosis.

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

MODELO BOX COX

Ventajas:

1. Más simple de entender y aplicar. Solo requiere de un único parámetro de transformación lambda.
2. Solo es válida para datos positivos, lo cual puede ser útil cuando se trabaja con variables que siempre debe de ser positivas.
3. Puede tener una interpretación directa en términos de porcentaje de cambio cuando la lambda no es igual a cero.

Desventajas:

1. Solo puede ser aplicado a datos que sean estrictamente positivos. No se puede utilizar con datos que incluyan ceros o valores negativos.
2. La elección del valor de lambda puede ser un desafío, y elegir un valor incorrecto puede no lograr la normalidad o incluso empeorar la distribución.
3. Puede ser sensible a outliers extremos en los datos.

MODELO YEO JOHNSON

Ventajas:

1. Es más flexible que el modelo de Box-Cox ya que puede ser aplicado a datos positivos y negativos.
2. Permite un rango más amplio de valores de lambda, lo que puede ser útil para encontrar una transformación que se ajuste mejor a los datos.
3. Es menos sensible a outliers extremos en comparación con el modelo de Box-Cox.

Desventajas:

1. Es más complejo en comparación con el modelo de Box-Cox debido a la inclusión de términos adicionales para manejar datos negativos.
2. La interpretación directa del cambio porcentual no es tan sencilla.
3. Puede generar valores negativos en los datos transformados, lo que podría no ser adecuado para ciertos contextos.

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

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

Transformación

1. Implica cambiar la distribución o la forma de los datos originales.
2. Se utiliza para lograr una distribución más cercana a la normalidad, estabilizar la varianza, o ajustar relaciones no lineales.
3. Puede cambiar drásticamente la apariencia y la interpretación de los datos.

Escalamiento

1. Implica ajustar la escala de los datos sin cambiar su distribución.
2. Se utiliza para estandarizar las unidades de medida de diferentes características, de modo que todas las variables tengan un rango similar y no se vea afectada su importancia relativa debido a sus escalas originales.
3. No cambia la distribución ni la forma de los datos.

**10.2. Indica cuándo es necesario utilizar cada uno.**

La transformación de los datos se utiliza cuando los datos no cumplen con los supuestos de normalidad o cuando las relaciones entre variables no son lineales. Por otro lado, el escalamiento de datos se utiliza cuando las características tienen diferentes unidades de medida y se busca evitar que una variable con una escala mayor domine el análisis.