Explorando bases

Nallely Serna

2024-08-13

```
M=read.csv("mc-donalds-menu.csv") #leer la base de datos
#M$variable   #para llamar una variable, aunque también la puedes leer
con corchetes cuadrados M[renglón, columna]
```

2. Analiza 2 de las siguientes variables en cuanto a sus datos atípicos y normalidad:

Calorias Carbohidratos Proteinas Sodio Azucares (Sugars)

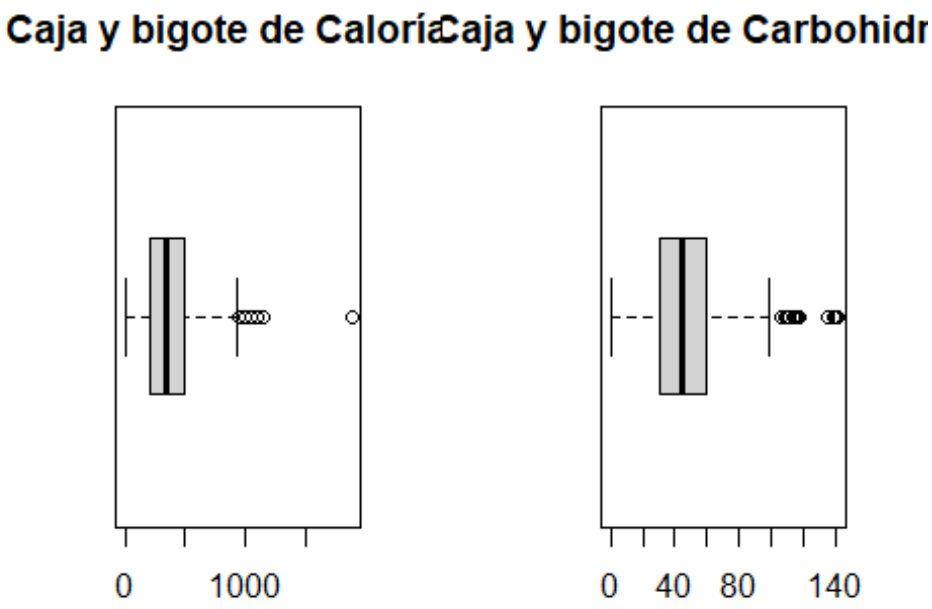
```
calorias <- M$Calories
carbohidratos <- M$Carbohydrates
```

#3. Para analizar datos atípicos se te sugiere:

1. Graficar el diagrama de caja y bigote
2. Calcula el rango intercuartílico y los cuartíles
3. Identifica la cota de 1.5 rangos intercuartílicos para datos atípicos, ¿hay datos atípicos de acuerdo con este criterio?
4. Identifica la cota de 3 desviaciones estándar alrededor de la media, ¿hay datos atípicos de acuerdo con este criterio?
5. Toma una decisión de si conviene o no quitar los datos atípicos (para ello interpreta la variable en el contexto del problema y determina si es necesario quitarlos o no quitarlos)

1. Graficar el diagrama de caja y bigote

```
par(mfrow = c(1, 2))  # Ventana gráfica de 1x2
boxplot(calorias, main = "Caja y bigote de Calorías", horizontal = TRUE)
boxplot(carbohidratos, main = "Caja y bigote de Carbohidratos",
horizontal = TRUE)
```



2. Calcula el rango intercuartílico y los cuartíles

```
# Para Calorías
q1_cal <- quantile(calorias, 0.25)
q3_cal <- quantile(calorias, 0.75)
iqr_cal <- IQR(calorias)

# Para Carbohidratos
q1_carb <- quantile(carbohidratos, 0.25)
q3_carb <- quantile(carbohidratos, 0.75)
iqr_carb <- IQR(carbohidratos)

# Mostrar resultados
q1_cal; q3_cal; iqr_cal
```

```
## 25%
## 210
```

```
## 75%
## 500
```

```
## [1] 290
```

```
q1_carb; q3_carb; iqr_carb
```

```
## 25%
##  30
```

```
## 75%
##  60

## [1] 30
```

#3. Identifica la cota de 1.5 rangos intercuartílicos para datos atípicos, ¿hay datos atípicos de acuerdo con este criterio?

```
# Para Calorías
limite_inferior_cal <- q1_cal - 1.5 * iqr_cal
limite_superior_cal <- q3_cal + 1.5 * iqr_cal

atipicos_cal <- calorias[calorias < limite_inferior_cal | calorias >
limite_superior_cal]

# Para Carbohidratos
limite_inferior_carb <- q1_carb - 1.5 * iqr_carb
limite_superior_carb <- q3_carb + 1.5 * iqr_carb

atipicos_carb <- carbohidratos[carbohidratos < limite_inferior_carb |
carbohidratos > limite_superior_carb]

# Mostrar los datos atípicos
atipicos_cal

## [1] 1090 1150  990 1050  940 1880

atipicos_carb

##  [1] 111 116 110 115 118 111 109 135 114 140 114 141 109 135 139 106
114
```

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

```
# Para Calorías
media_cal <- mean(calorias)
desviacion_estandar_cal <- sd(calorias)

limite_inferior_cal_sd <- media_cal - 3 * desviacion_estandar_cal
limite_superior_cal_sd <- media_cal + 3 * desviacion_estandar_cal

atipicos_cal_sd <- calorias[calorias < limite_inferior_cal_sd | calorias
> limite_superior_cal_sd]

# Para Carbohidratos
media_carb <- mean(carbohidratos)
desviacion_estandar_carb <- sd(carbohidratos)

limite_inferior_carb_sd <- media_carb - 3 * desviacion_estandar_carb
```

```r
limite_superior_carb_sd <- media_carb + 3 * desviacion_estandar_carb

atipicos_carb_sd <- carbohidratos[carbohidratos < limite_inferior_carb_sd
| carbohidratos > limite_superior_carb_sd]

# Mostrar los datos atípicos
atipicos_cal_sd
```

```
## [1] 1090 1150 1880
```

```r
atipicos_carb_sd
```

```
## [1] 135 140 141 135 139
```

4. Para analizar normalidad se te sugiere:

Realiza pruebas de normalidad univariada de las variables (selecciona entre los métodos vistos en clase) Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos) para cada variable Calcula el coeficiente de sesgo y el coeficiente de curtosis de cada variable. Compara las medidas de media, mediana y rango medio de cada variable. Realiza el histograma y su distribución teórica de probabilidad (sugerencia, adapta el código: hist(datos,freq=FALSE) lines(density(datos),col="red") curve(dnorm(x,mean=mean(datos,sd=sd(datos)), from=-6, to=6, add=TRUE, col="blue",lwd=2) Comenta los gráficos y los resultados obtenidos con vías a interpretar normalidad de los datos.

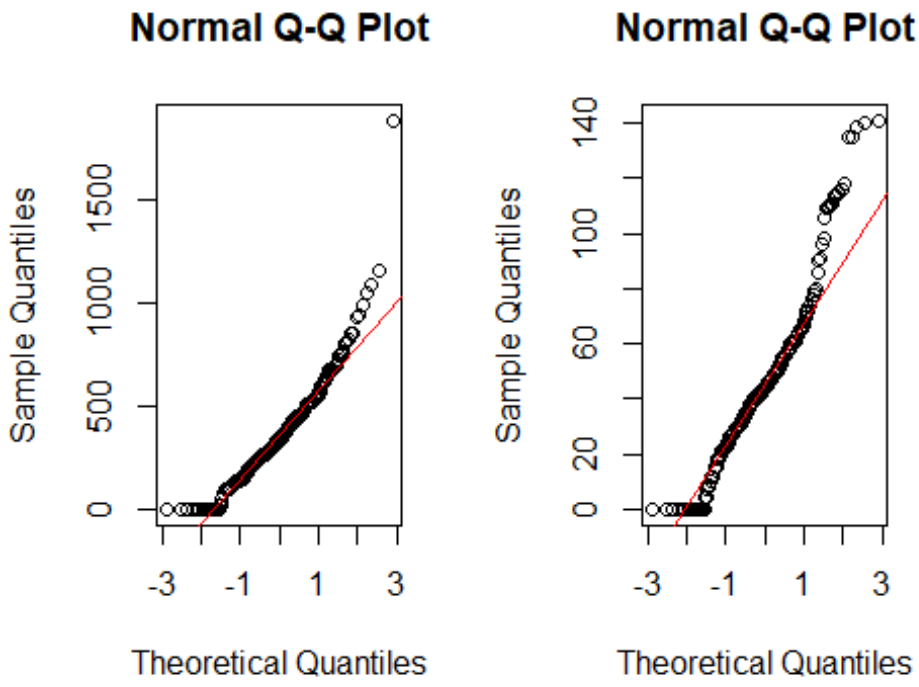
```r
shapiro.test(calorias)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  calorias
## W = 0.91902, p-value = 1.119e-10
```

```r
shapiro.test(carbohidratos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  carbohidratos
## W = 0.93666, p-value = 3.931e-09
```

```r
par(mfrow = c(1, 2))
qqnorm(calorias)
qqline(calorias, col = "red")
qqnorm(carbohidratos)
qqline(carbohidratos, col = "red")
```




```
install.packages("moments")
```

```
## Warning: package 'moments' is in use and will not be installed
```

```
library(moments)
```

```
# Calcular sesgo y curtosis
sesgo_calorias <- skewness(calorias)
curtosis_calorias <- kurtosis(calorias)

sesgo_carbohidratos <- skewness(carbohidratos)
curtosis_carbohidratos <- kurtosis(carbohidratos)

# Mostrar resultados
sesgo_calorias
```

```
## [1] 1.444105
```

```
curtosis_calorias
```

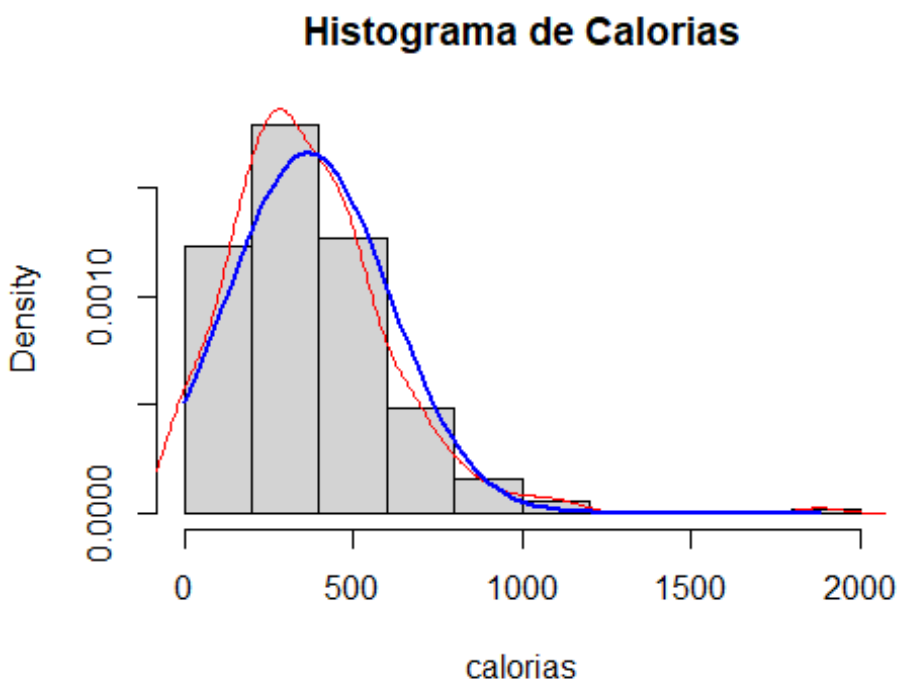
```
## [1] 8.645274
```

```
sesgo_carbohidratos
```

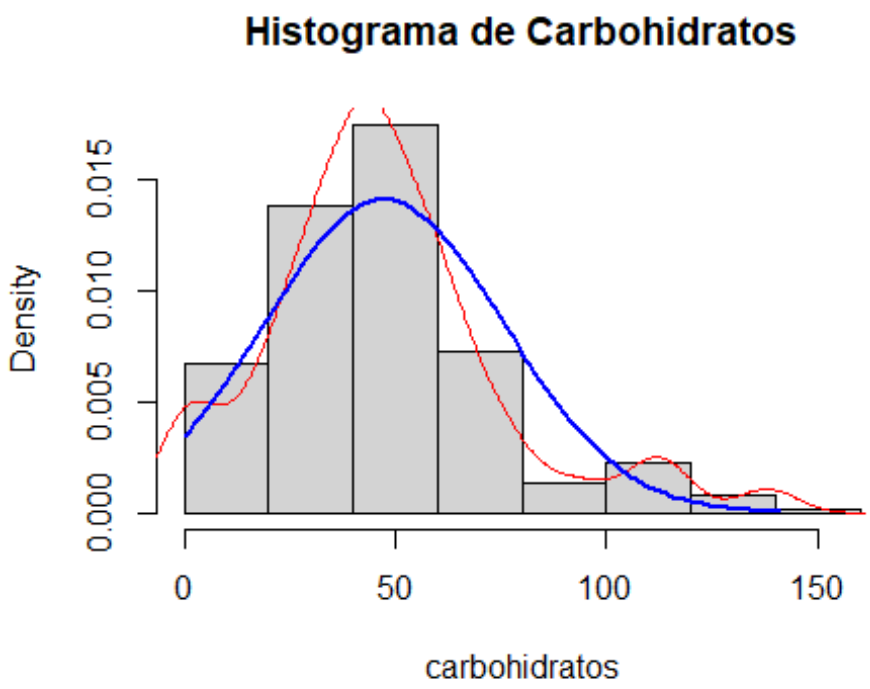
```
## [1] 0.9074253
```

```
curtosis_carbohidratos
```

```
## [1] 4.357538

# Calorias
mean_calorias <- mean(calorias)
median_calorias <- median(calorias)
rango_medio_calorias <- (max(calorias) + min(calorias)) / 2

# Carbohidratos
mean_carbohidratos <- mean(carbohidratos)
median_carbohidratos <- median(carbohidratos)
rango_medio_carbohidratos <- (max(carbohidratos) + min(carbohidratos)) /
2

# Mostrar resultados
mean_calorias

## [1] 368.2692

median_calorias

## [1] 340

rango_medio_calorias

## [1] 940

mean_carbohidratos

## [1] 47.34615

median_carbohidratos

## [1] 44

rango_medio_carbohidratos

## [1] 70.5

# Para Calorias
hist(calorias, freq=FALSE, main="Histograma de Calorias")
lines(density(calorias), col="red")
curve(dnorm(x, mean=mean(calorias), sd=sd(calorias)),
      from=min(calorias), to=max(calorias),
      add=TRUE, col="blue", lwd=2)
```

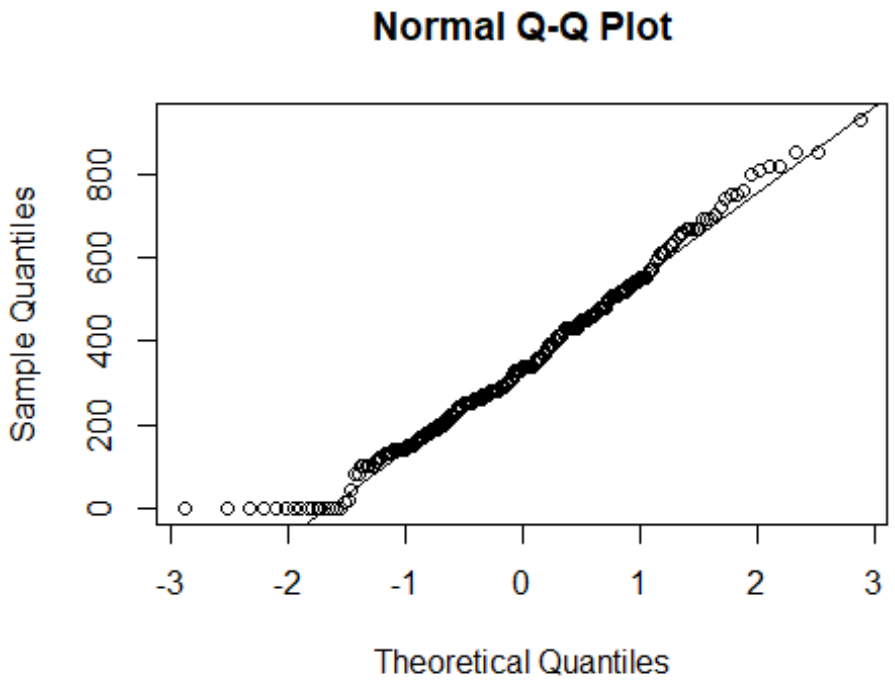



```r
# Para Carbohidratos
hist(carbohidratos, freq=FALSE, main="Histograma de Carbohidratos")
lines(density(carbohidratos), col="red")
curve(dnorm(x, mean=mean(carbohidratos), sd=sd(carbohidratos)),
      from=min(carbohidratos), to=max(carbohidratos),
      add=TRUE, col="blue", lwd=2)
```

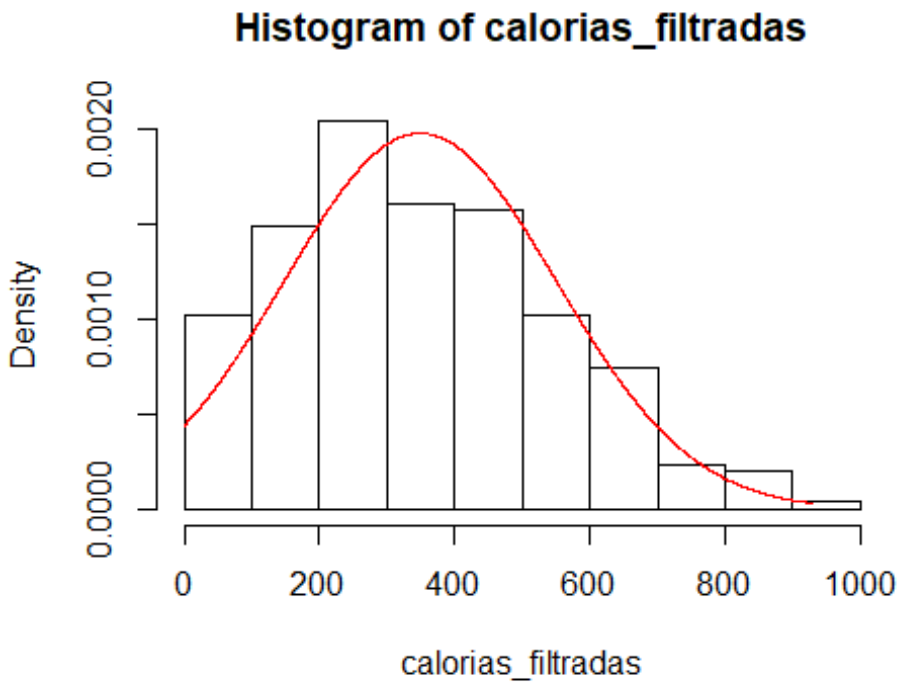



```
# Para Calorias
q1_cal <- quantile(calorias, 0.25)
q3_cal <- quantile(calorias, 0.75)
ri_cal <- q3_cal - q1_cal

# Remover valores atípicos
calorias_filtradas <- calorias[calorias < (q3_cal + 1.5 * ri_cal)]

# Para Carbohidratos
q1_carb <- quantile(carbohidratos, 0.25)
q3_carb <- quantile(carbohidratos, 0.75)
ri_carb <- q3_carb - q1_carb

# Remover valores atípicos
carbohidratos_filtrados <- carbohidratos[carbohidratos < (q3_carb + 1.5 *
ri_carb)]

# Resúmenes antes y después
summary(calorias)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   210.0   340.0   368.3   500.0  1880.0

summary(calorias_filtradas)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   202.5   335.0   349.0   480.0   930.0
```

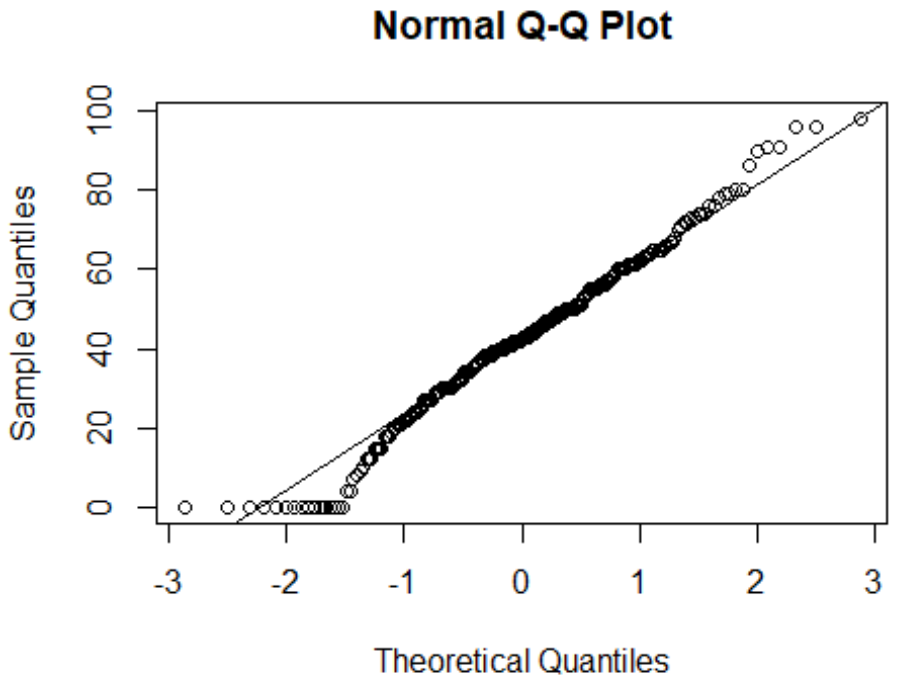
```
summary(carbohidratos)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   44.00   47.35   60.00  141.00

summary(carbohidratos_filtrados)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   43.00   42.28   56.00   98.00

# Para Calorias
qqnorm(calorias_filtradas)
qqline(calorias_filtradas)
```



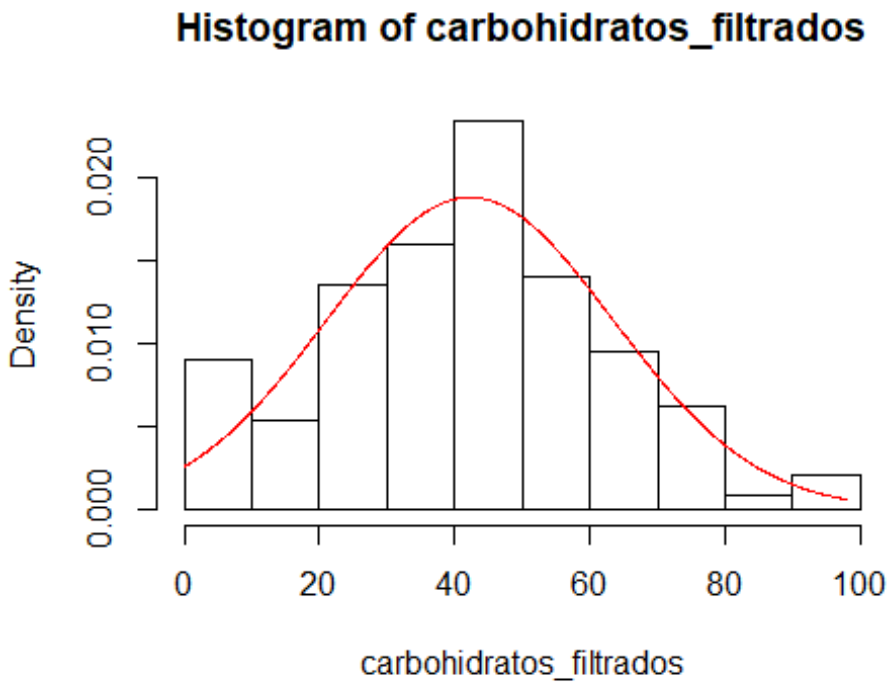

```
hist(calorias_filtradas, prob=TRUE, col=0)
x_cal <- seq(min(calorias_filtradas), max(calorias_filtradas), 0.1)
y_cal <- dnorm(x_cal, mean=mean(calorias_filtradas),
sd=sd(calorias_filtradas))
lines(x_cal, y_cal, col="red")
```




```r
# Para Carbohidratos
qqnorm(carbohidratos_filtrados)
qqline(carbohidratos_filtrados)
```



```
hist(carbohidratos_filtrados, prob=TRUE, col=0)
x_carb <- seq(min(carbohidratos_filtrados), max(carbohidratos_filtrados),
0.1)
y_carb <- dnorm(x_carb, mean=mean(carbohidratos_filtrados),
sd=sd(carbohidratos_filtrados))
lines(x_carb, y_carb, col="red")
```



En este caso, no recomendaría eliminar los datos atípicos. Dado que los valores extremos son representativos de productos reales en el menú de McDonald's, eliminarlos podría distorsionar el análisis. Además, McDonald's tiene una amplia variedad de productos, desde opciones muy bajas en calorías y carbohidratos hasta opciones muy altas. Esto es parte integral del análisis y, por lo tanto, eliminar estos datos podría omitir información crucial sobre el rango completo de productos ofrecidos.