Actividad 4. Explorando bases

José Carlos Sánchez Gómez

2024-08-14

```
data =
read.csv("C:\\Users\\jcsg6\\Documentos\\Uni\\SeptimoSemestre\\Estadistica
\\mc-donalds-menu.csv")
```

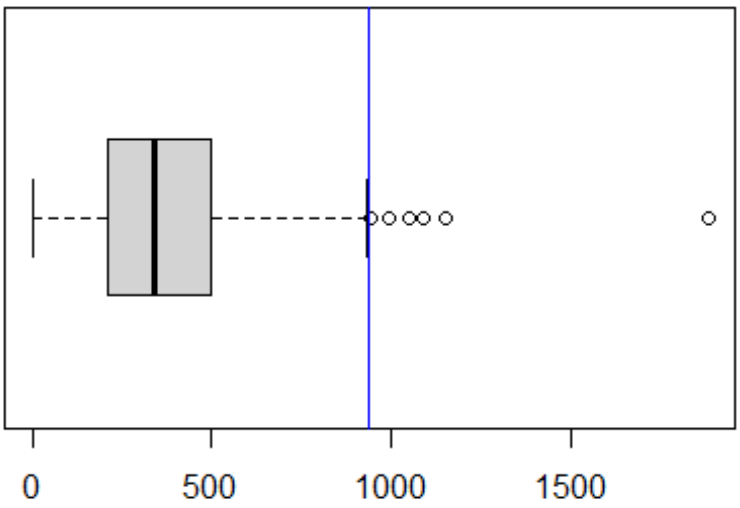
Análisis de las Calorías

Explorando los datos, analizando los datos atípicos y su normalidad

Análisis de datos atípicos

```
calorias = data$Calories
q1_calorias = quantile(calorias, 0.25)
q3_calorias = quantile(calorias, 0.75)
ri_calorias = q3_calorias - q1_calorias
# ri2 = IQR(calorias) comprobar que si da lo mismo

boxplot(calorias, horizontal = TRUE)
abline(v = q3_calorias + 1.5 * ri_calorias, col="blue")  #linea vertical
en el límite de los datos atípicos o extremos
```



```r
calorias_filtradas = data[data$Calories < q3_calorias +1.5 * ri_calorias,
c("Calories")]  #En la matriz M, quitar datos más allá de 3 rangos
intercuartílicos arriba de q3 de la variable X
summary(calorias)
```

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

```r
summary(calorias_filtradas)
```

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

Total de datos atípicos con respecto a la cota de rangos intercuartilicos

```r
cota_inferior_calorias = q1_calorias - 1.5 * ri_calorias
cota_superior_calorias = q3_calorias + 1.5 * ri_calorias
calorias_fuera_cotas = calorias[calorias < cota_inferior_calorias |
calorias > cota_superior_calorias]
calorias_fuera_cotas
```

```
## [1] 1090 1150  990 1050  940 1880
```

```r
data$Item[calorias < cota_inferior_calorias | calorias >
cota_superior_calorias]
```

```
## [1] "Big Breakfast with Hotcakes (Regular Biscuit)"
## [2] "Big Breakfast with Hotcakes (Large Biscuit)"
## [3] "Big Breakfast with Hotcakes and Egg Whites (Regular Biscuit)"
## [4] "Big Breakfast with Hotcakes and Egg Whites (Large Biscuit)"
## [5] "Chicken McNuggets (20 piece)"
## [6] "Chicken McNuggets (40 piece)"
```

Total de datos atípicos con respecto a la cota de 3 desviaciones estandar alrededor de la media

```r
media_calorias = mean(calorias)
calorias_sd = sd(calorias)
cota_inferior_calorias_sd = media_calorias - 3 * calorias_sd
cota_superior_calorias_sd = media_calorias + 3 * calorias_sd
calorias_fuera_cota_sd = calorias[calorias < cota_inferior_calorias_sd |
calorias > cota_superior_calorias_sd]
calorias_fuera_cota_sd
```

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

```r
data$Item[calorias < cota_inferior_calorias_sd | calorias >
cota_superior_calorias_sd]
```

```
## [1] "Big Breakfast with Hotcakes (Regular Biscuit)"
## [2] "Big Breakfast with Hotcakes (Large Biscuit)"
## [3] "Chicken McNuggets (40 piece)"
```

Tras analizar los datos atípicos según diferentes criterios y evaluar a qué corresponden, concluyo que no deberían eliminarse. Estos valores atípicos representan opciones del menú con porciones significativamente más grandes que el promedio, lo que explica la disparidad observada en comparación con otros datos. Dado que estos valores reflejan variaciones reales en el tamaño de las porciones y no errores en los datos, es importante conservarlos para mantener la integridad y representatividad del análisis.

Análisis de normalidad

Comparación de pruebas mediante Anderson-Darling y Kolmorogov-Smirnov; junto con qqplot de calorías

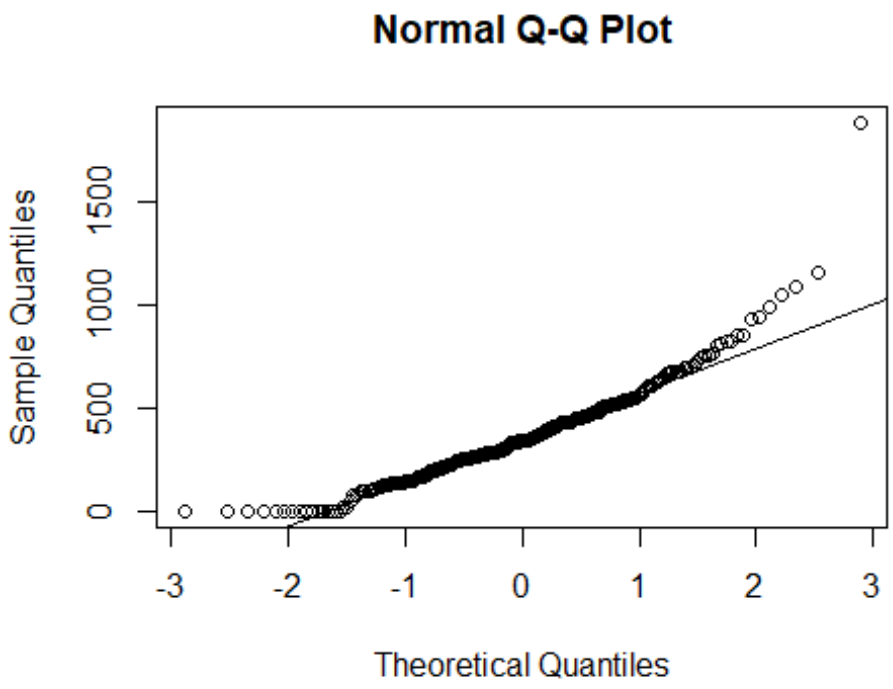
```
library(nortest)
lillie.test(calorias)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  calorias
## D = 0.073753, p-value = 0.001611
```

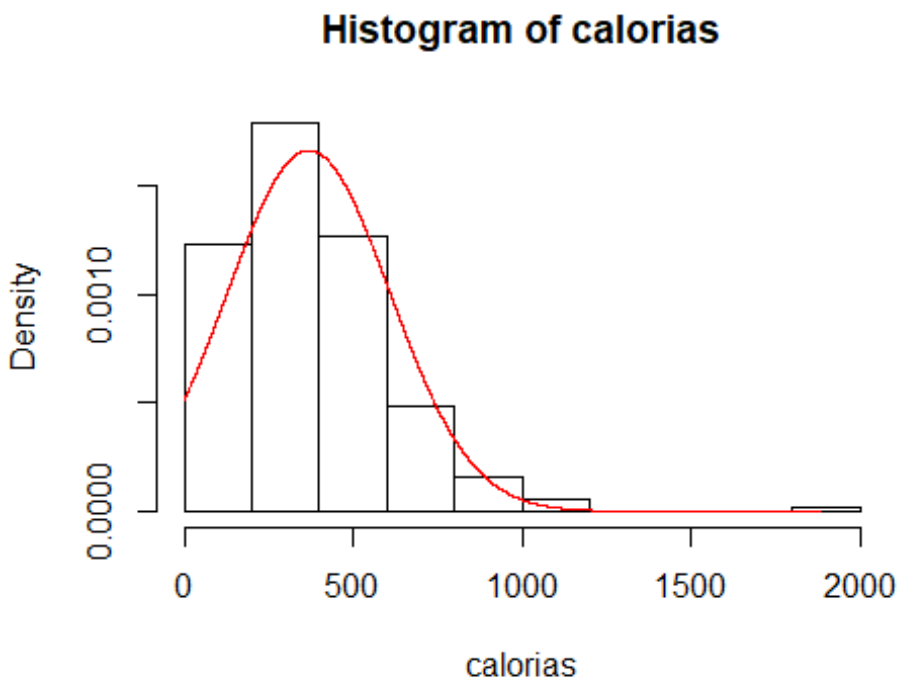
```
ad.test(calorias)
```

```
##
##  Anderson-Darling normality test
##
## data:  calorias
## A = 2.5088, p-value = 2.369e-06
```

```
qqnorm(calorias)
qqline(calorias)
```




```r
hist(calorias,prob=TRUE,col=0)
x=seq(min(calorias),max(calorias),0.1)
y=dnorm(x,mean(calorias),sd(calorias))
lines(x,y,col="red")
```



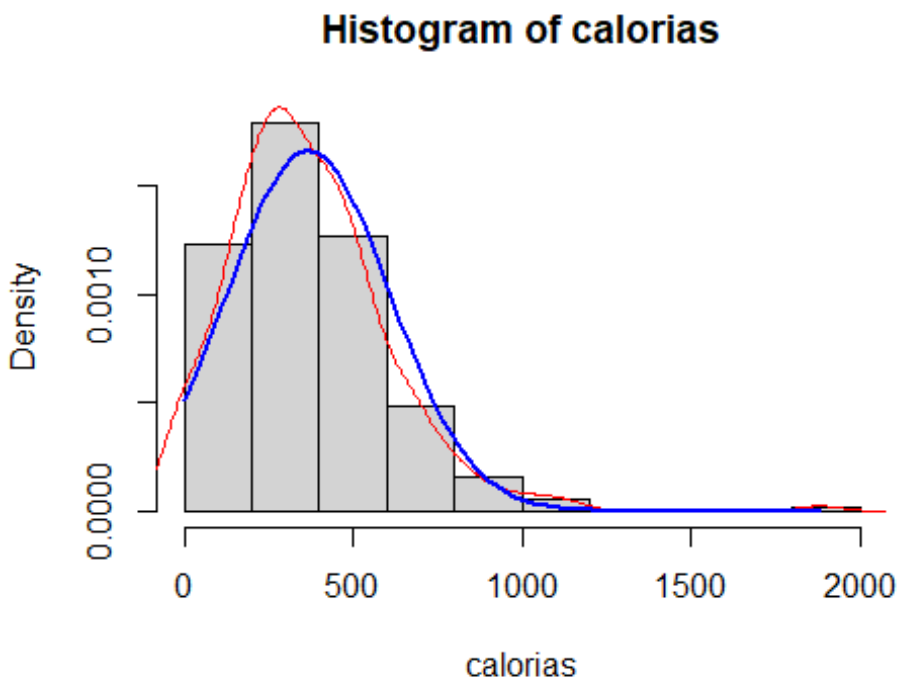
Obtención de sesgo, coeficiente curtosis, media, mediana, y rango medio

```
library(e1071)
data.frame(
  Sesgo = skewness(calorias),
  Curtosis = kurtosis(calorias),
  Media = mean(calorias),
  Mediana = median(calorias),
  Rango_Medio = (min(calorias) + max(calorias)) / 2
)
```

```
##      Sesgo Curtosis    Media Mediana Rango_Medio
## 1 1.435782   5.5789 368.2692     340         940
```

Histograma y distribucion teórica de probabilidad

```
hist(calorias, freq = FALSE)
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)
```



Tras analizar ambas gráficas pude darme cuenta de que no siguen una distribución normal. El QQPlot muestra una desviación significativa de la línea de referencia en los extremos, mientras que el histograma tiene un sesgo positivo con una cola larga a la derecha. Por esto mismo, deberíamos de tener en cuenta los valores atípicos que crean estos comportamientos en futuros análisis.

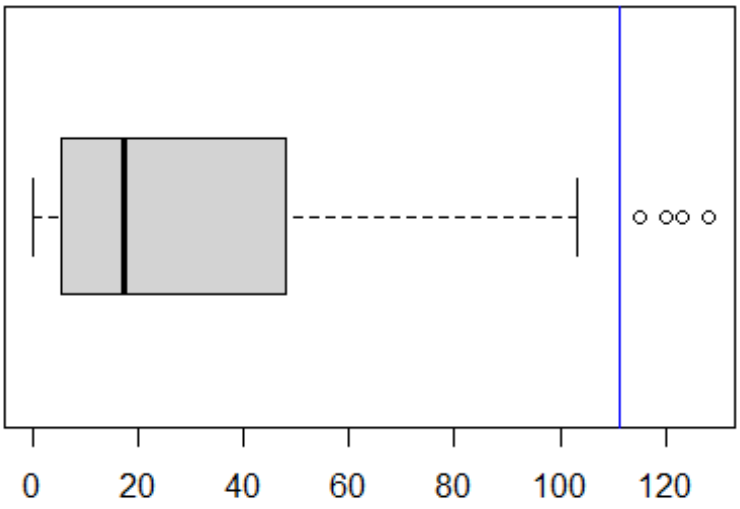
Análisis de los Azucares

Explorando los datos, analizando los datos atípicos y su normalidad

Análisis de datos atípicos

```
azucares = data$Sugars
q1_azucares = quantile(azucares, 0.25)
q3_azucares = quantile(azucares, 0.75)
ri_azucares = q3_azucares - q1_azucares
# ri2 = IQR(calorias) comprobar que si da lo mismo

boxplot(azucares, horizontal = TRUE)
abline(v = q3_azucares + 1.5 * ri_azucares, col="blue")  #linea vertical
en el límite de los datos atípicos o extremos
```




```
azucares_filtradas = data[data$Sugars < q3_azucares +1.5 * ri_azucares,
c("Sugars")]  #En la matriz M, quitar datos más allá de 3 rangos
intercuartílicos arriba de q3 de la variable X
summary(azucares)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.75   17.50   29.42   48.00  128.00

summary(azucares_filtradas)

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

Total de datos atípicos con respecto a la cota de rangos intercuartilicos

```
cota_inferior_azucares = q1_azucares - 1.5 * ri_azucares
cota_superior_azucares = q3_azucares + 1.5 * ri_azucares
azucares_fuera_cotas = azucares[azucares < cota_inferior_azucares |
azucares > cota_superior_azucares]
azucares_fuera_cotas

## [1] 123 120 115 128

data$Item[azucares < cota_inferior_azucares | azucares >
cota_superior_azucares]

## [1] "Strawberry Shake (Large)"
## [2] "Chocolate Shake (Large)"
```

```
## [3] "Shamrock Shake (Large)"
## [4] "McFlurry with M&M's Candies (Medium)"
```

Total de datos atípicos con respecto a la cota de 3 desviaciones estandar alrededor de la media

```
media_azucares = mean(azucares)
azucares_sd = sd(azucares)
cota_inferior_azucares_sd = media_azucares - 3 * azucares_sd
cota_superior_azucares_sd = media_azucares + 3 * azucares_sd
azucares_fuera_cota_sd = azucares[azucares < cota_inferior_azucares_sd |
azucares > cota_superior_azucares_sd]
azucares_fuera_cota_sd
```

```
## [1] 123 120 128
```

```
data$Item[azucares < cota_inferior_azucares_sd | azucares >
cota_superior_azucares_sd]
```

```
## [1] "Strawberry Shake (Large)"
## [2] "Chocolate Shake (Large)"
## [3] "McFlurry with M&M's Candies (Medium)"
```

Similar al análisis de la variable anterior, los valores atípicos que se muestran son debido a la diferencia de porciones entre una opción del menú y otra. No por un error en la información. Por lo que considero de igual manera que es importante no eliminar estos valores, para mantener una integridad en los datos.

Análisis de normalidad

Comparación de pruebas mediante Anderson-Darling y Kolmorogov-Smirnov; junto con qqplot de calorias

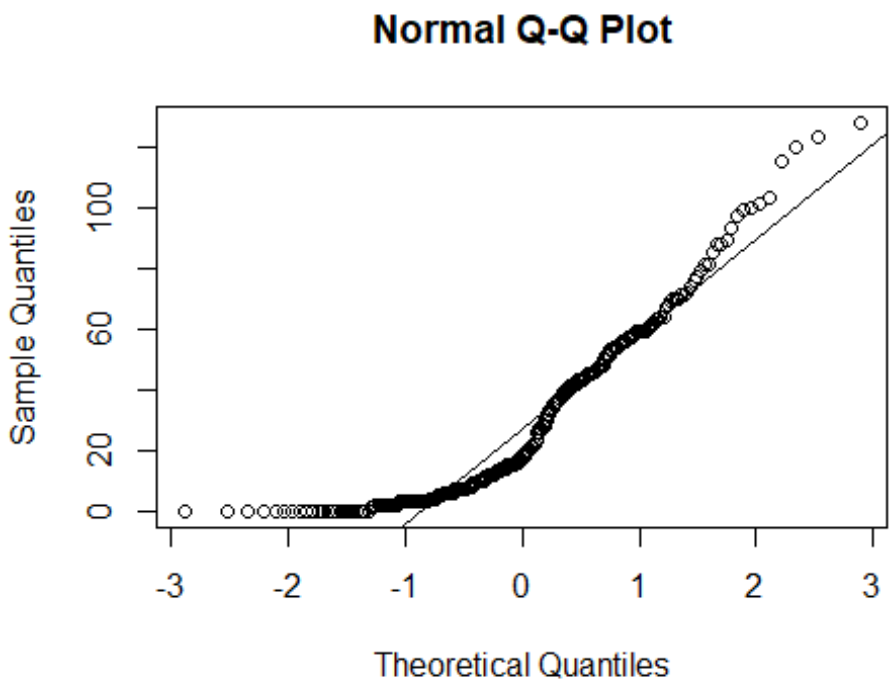
```
library(nortest)
lillie.test(azucares)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  azucares
## D = 0.16858, p-value < 2.2e-16
```

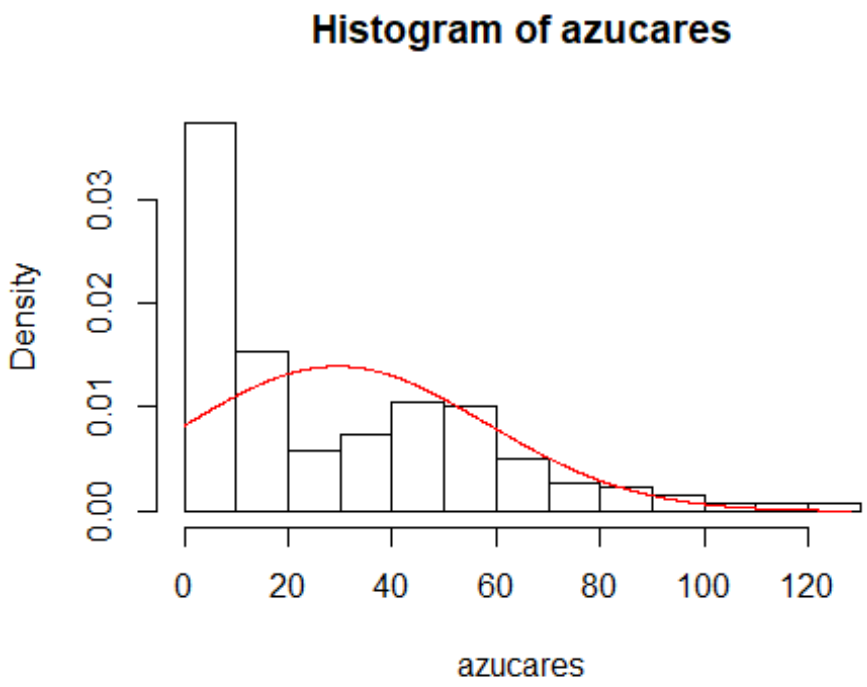
```
ad.test(azucares)
```

```
##
##  Anderson-Darling normality test
##
## data:  azucares
## A = 9.9899, p-value < 2.2e-16
```

```
qqnorm(azucares)
qqline(azucares)
```




```r
hist(azucares,prob=TRUE,col=0)
x=seq(min(azucares),max(azucares),0.1)
y=dnorm(x,mean(azucares),sd(azucares))
lines(x,y,col="red")
```



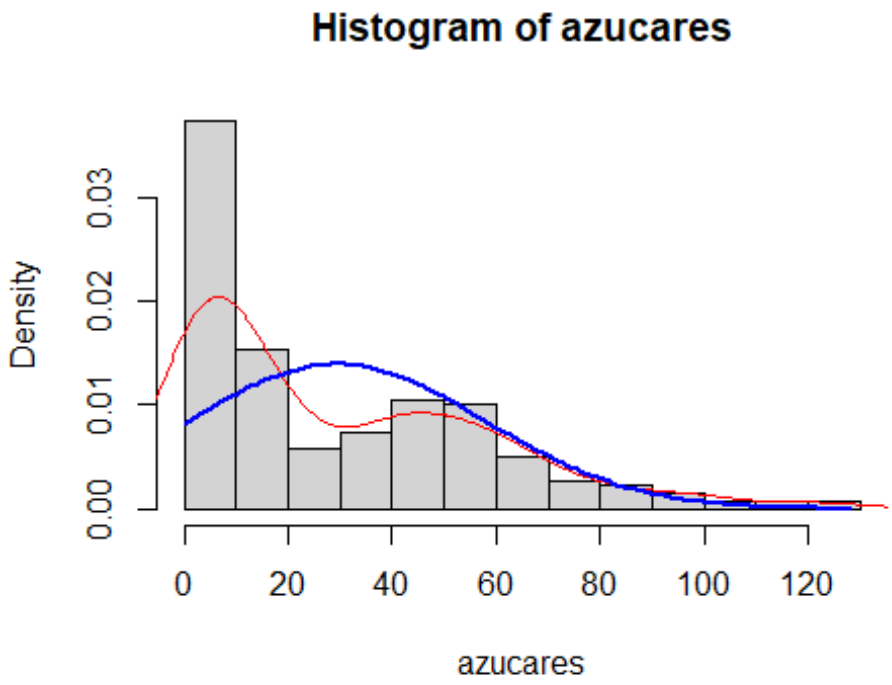
Obtención de sesgo, coeficiente curtosis, media, mediana, y rango medio

```
library(e1071)
data.frame(
  Sesgo = skewness(azucares),
  Curtosis = kurtosis(azucares),
  Media = mean(azucares),
  Mediana = median(azucares),
  Rango_Medio = (min(azucares) + max(azucares)) / 2
)
```

```
##      Sesgo Curtosis    Media Mediana Rango_Medio
## 1 1.020064 0.460967 29.42308    17.5          64
```

Histograma y distribucion teórica de probabilidad

```
hist(azucares, freq = FALSE)
lines(density(azucares), col = "red")
curve(dnorm(x, mean = mean(azucares), sd = sd(azucares)),
      from = min(azucares), to = max(azucares),
      add = TRUE, col = "blue", lwd = 2)
```



Similar a la variable anterior, las gráficas muestran que no hay una distribución normal dentro de los datos. El QQPlot muestra una desviación significativa de la línea de referencia en sus extremos y mitad; esto nos dice que los datos no siguen una distribución normal. Mientas que el histograma muestra una distribución bastante asimétrica. La curva de densidad nos informa que existe una gran concentración de valores bajos, con una cola creciente hacia la derecha que muestra la presencia de valores atípicos mayores. Debido a que estos datos presentan una anormalidad mayor al de la variable anterior, considero que se deberían de transformar estos datos para obtener un análisis más adecuado.