Act2_ExplorandoBases

Ana Lucía Cárdenas Pérez

2023-08-18

```
# Cargamos archivo
data <- read.csv("mc-donalds-menu-1.csv")
```

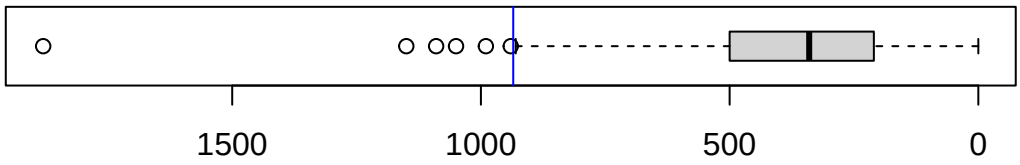
```
# Seleccionamos 2 variables y las guardamos
calories <- data$Calories
carbohydrates <- data$Carbohydrates
```

```
q1 = quantile(calories,0.25) # q1 variable calories
q3 = quantile(calories,0.75) #q3 calories
ri = IQR(calories) # rango intercuartílico de calories
par(mfrow = c(2,1))
boxplot(calories,horizontal=TRUE,ylim=c(max(calories),min(calories)))
abline(v=q3+1.5*ri,col="blue")
X1 = data[data$Calories<q3+1.5*ri,c("Calories")]
summary(X1)
```

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

```
summary(calories)
```

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

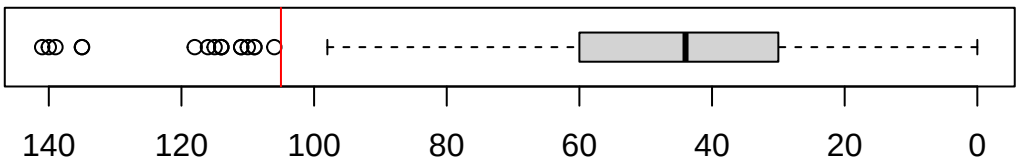



```
Carbq1 = quantile(carbohydrates,0.25) # q1 variable carbohydrates
Carbq3 = quantile(carbohydrates,0.75) #q3 carbohydrates
Carbri = IQR(carbohydrates) # rango intercuartílico de carbohydrates
par(mfrow = c(2,1))
boxplot(carbohydrates,horizontal=TRUE,ylim=c(max(carbohydrates),min(carbohydrates)))
abline(v=Carbq3+1.5*Carbri,col="red")
CarbX1 = data[data$Carbohydrates<Carbq3+1.5*Carbri,c("carbohydrates")]
summary(CarbX1)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

```
summary(carbohydrates)
```

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



#los valores en el boxplot de calorias muestran outliers del lado izquierdo, en comparación con el boxplot de carbohidratos, podemos ver que tenemos que revisar esos datos que se muestran fuera del boxplot para ver que información están registrando y ver si son datos necesarios, si se pueden corregir o si se pueden eliminar o modificar.

```
# 1. Pruebas de normalidad con el Test de Kolmogorov-Smirnov para ambas variables
KS_calories <- ks.test(calories, "pnorm")
```

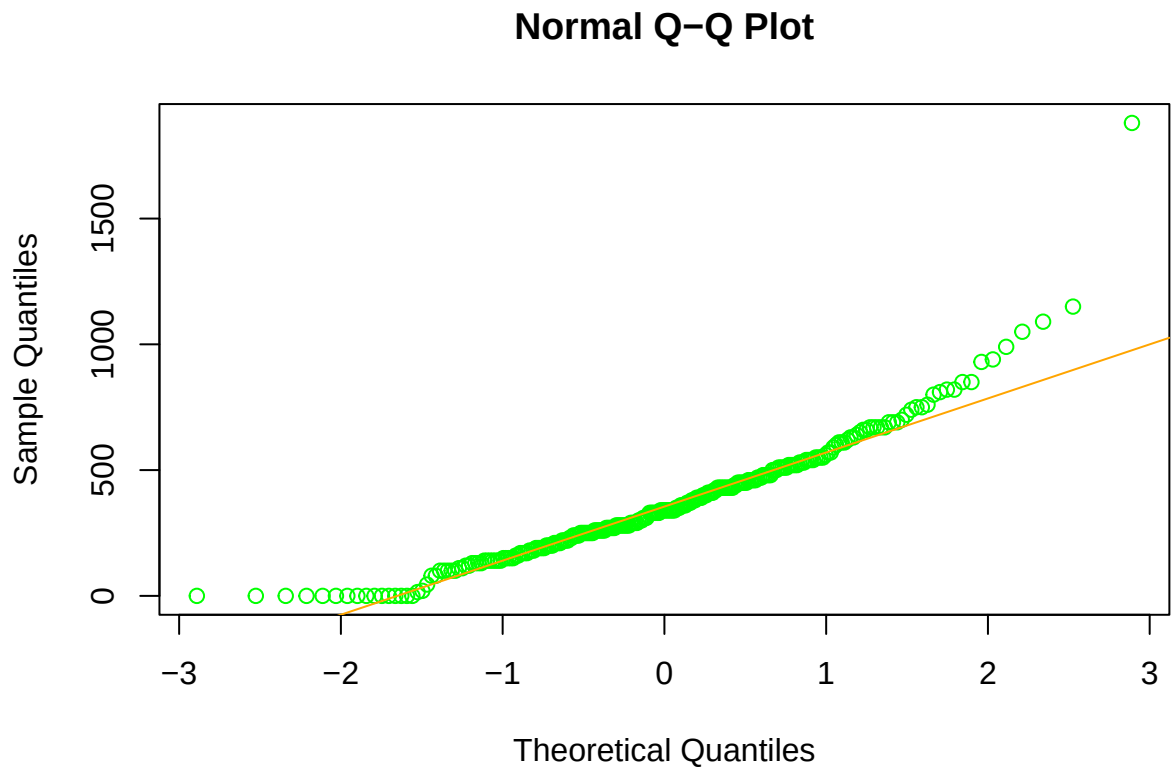
```
## Warning in ks.test.default(calories, "pnorm"): ties should not be present for
## the Kolmogorov-Smirnov test
```

```
KS_carbohydrates <- ks.test(carbohydrates, "pnorm")
```

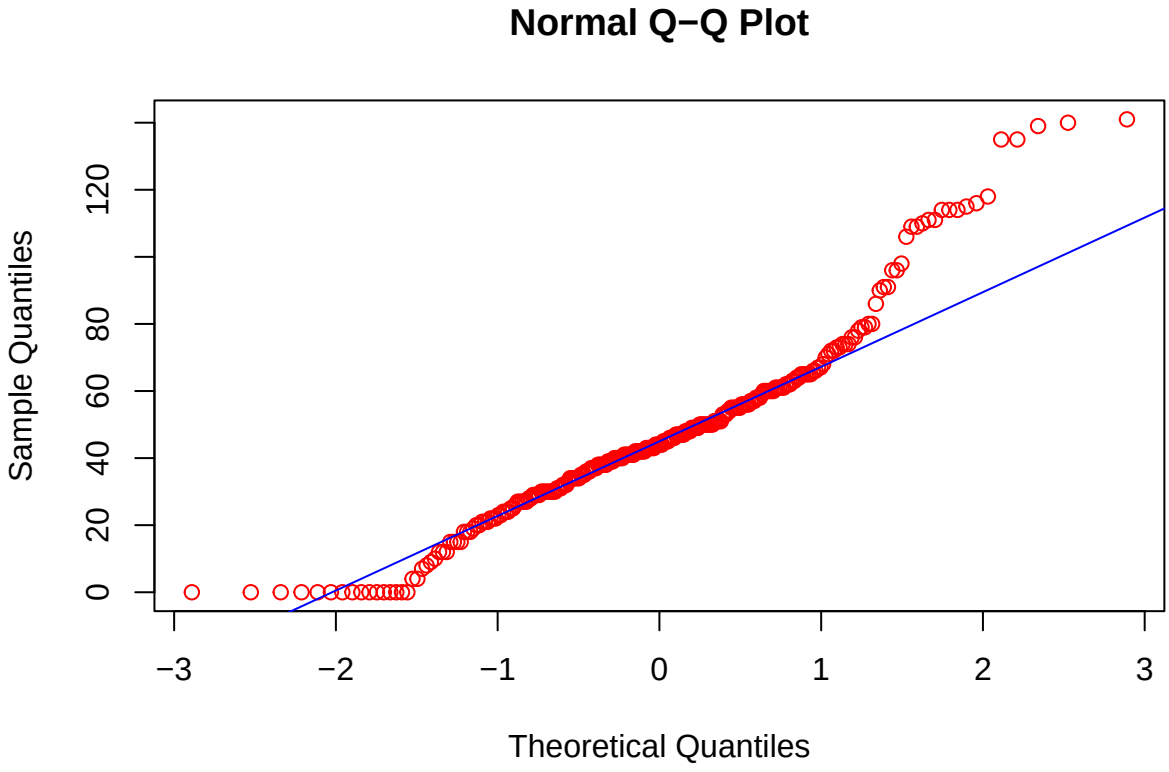
```
## Warning in ks.test.default(carbohydrates, "pnorm"): ties should not be present
## for the Kolmogorov-Smirnov test
```

```
# Graficar los datos con sus QQPlot - qqnorm y qqline para cada variable

qqnorm(calories,col="green")
qqline(calories,col="orange")
```




```
qqnorm(carbohydrates,col="red")
qqline(carbohydrates, col="blue")
```



Lo que las gráficas QQPlot nos muestran son qnorm y qline donde qnorm nos muestra los datos de cuantiles teóricos en una distribución normal. Qline nos ayuda a representar con una línea de referencia el como se espera que se vean los datos ya que entre más cerca de esa línea de referencia, más normal es la distribución.

Ahora analizando ambas gráficas podemos ver que en la gráfica de carbohidratos, del lado derecho podeemos ver una curva que sube exponencialmente, baja, y vuelve a subir, por lo que en comparación, la distibución de las calorias en más "normal".

```
# Cargamos Paquetes y Librerias
install.packages("moments")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(moments)
```

```
# Calcular el coeficiente de sesgo y curtosis de calories
sesgo_cal <- skewness(calories)
print("Sesgo Calorias: ")
```

```
## [1] "Sesgo Calorias: "
```

```
sesgo_cal
```

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

```
curtosis_cal <- kurtosis(calories)
print("Curtosis Calorias: ")
```

```
## [1] "Curtosis Calorias: "
```

```
curtosis_cal
```

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

```
# Calcular el coeficiente de sesgo y curtosis de carbohydrates
sesgo_carb <- skewness(carbohydrates)
print("Sesgo Carbohidratos: ")
```

```
## [1] "Sesgo Carbohidratos: "
```

```
sesgo_carb
```

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

```
curtosis_carb <- kurtosis(carbohydrates)
print("Curtosis Carbohidratos: ")
```

```
## [1] "Curtosis Carbohidratos: "
```

```
curtosis_carb
```

```
## [1] 4.357538
```

```
# Comparar medidas de media, mediana y rango medio de calories
mediaCalories <- mean(calories)
print("Media Calorias: ")
```

```
## [1] "Media Calorias: "
```

```
mediaCalories
```

```
## [1] 368.2692
```

```
medianaCalories <- median(calories)
print("Mediana Calorias: ")
```

```
## [1] "Mediana Calorias: "
```

```
medianaCalories
```

```
## [1] 340
```

```
rangoMedioCalories <- (max(calories) + min(calories)) / 2
print("Rango Medio Calorias: ")
```

```
## [1] "Rango Medio Calorias: "
```

```
rangoMedioCalories
```

```
## [1] 940
```

```
# Comparar medidas de media, mediana y rango medio de calories
mediaCarbs <- mean(carbohydrates)
print("Media Carbohidratos: ")
```

```
## [1] "Media Carbohidratos: "
```

```
mediaCarbs
```

```
## [1] 47.34615
```

```
medianaCarbs <- median(carbohydrates)
print("Mediana Cargohidratos: ")
```

```
## [1] "Mediana Cargohidratos: "
```

```
medianaCarbs
```

```
## [1] 44
```

```
rangoMedioCarbs <- (max(carbohydrates) + min(carbohydrates)) / 2
print("Rango Medio Carbohidratos: ")
```

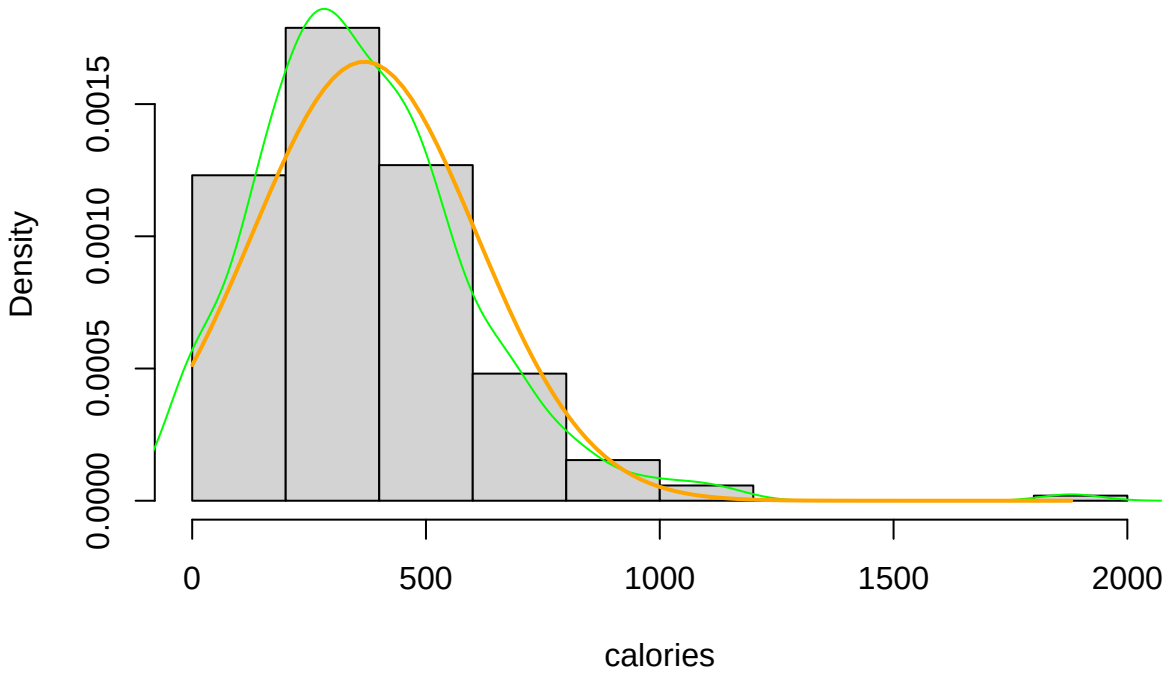
```
## [1] "Rango Medio Carbohidratos: "
```

```
rangoMedioCarbs
```

```
## [1] 70.5
```

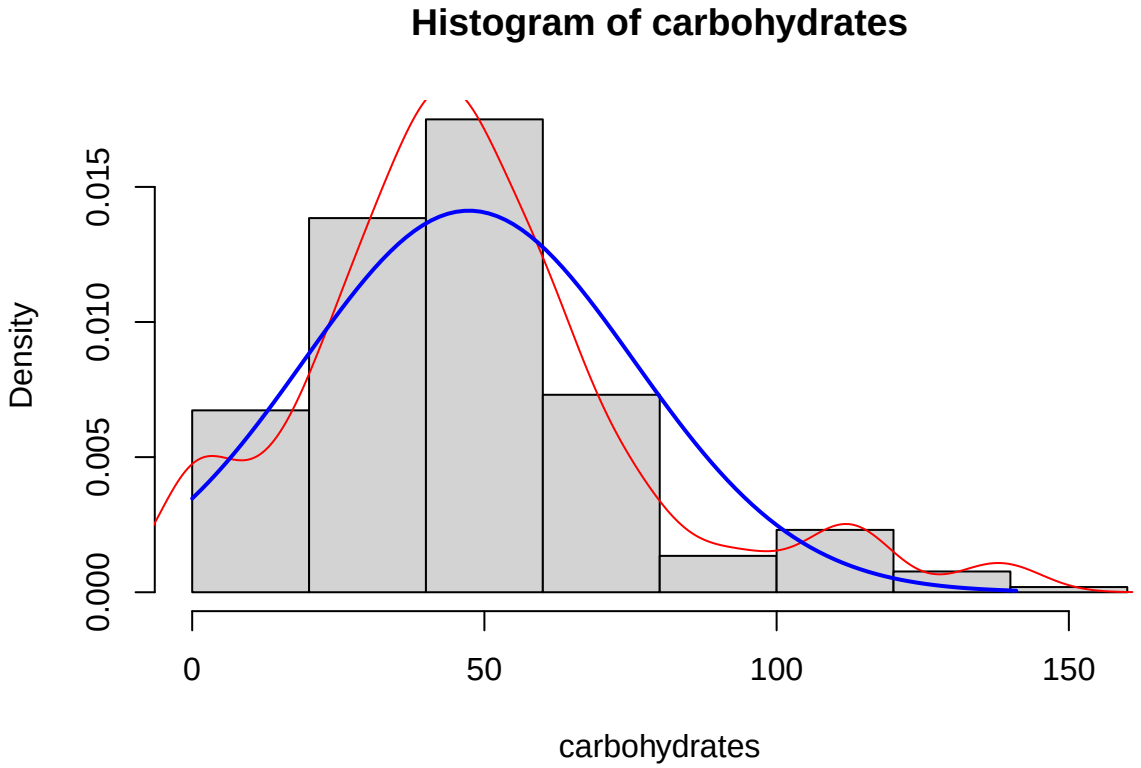
```
# Realizar histograma y distribución teórica de probabilidad - Calories

hist(calories,freq=FALSE)
lines(density(calories),col="green")
curve(dnorm(x, mean = mean(calories), sd = sd(calories)), from = min(calories), to = max(calories), add
```




```
# Realizar histograma y distribución teórica de probabilidad - Carboydrates

hist(carbohydrates,freq=FALSE)
lines(density(carbohydrates),col="red")
curve(dnorm(x, mean = mean(carbohydrates), sd = sd(carbohydrates)), from = min(carbohydrates), to = max
```




```
# un histograma nos va a decir si la distribución es normal o no dependiendo de la curva de las líneas.
```

en el caso del histograma de calorías tenemos 2 líneas, una naranja y una verde donde la verde representa la estimación de la densisdad de la probabilidad de una manera visual para ver como es la disribución de los datos. La curva naranja es la curva teórica por lo que de manera similar a QQPlot, entre mejor se ajusten las curvas al histograma, más normal es la distribución de los datos. En este caso podemos ver como la línea verde es similar en forma a la naranja.

En el caso del histograma de carbohidratos donde la línea roja son la estimación de la densidad de probabilidad, y la línea naranja es la distribución normal teórica, podemos observar que la normalidad de los datos no es cómo la de calorias. Si vemos bien, la línea roja del histograma presenta varias curvas a lo largo de la gráfica, mientras que la de calorías solo mostró una curva. Ya que la línea azul no presenta más de una curva, los datos de estimación no presentan una distribución normal.