Explorando bases

Catherine Rojas

2024-08-13

Carga de Datos

```
# Cargar el archivo de datos
data <- read.csv("mc-donalds-menu.csv")
```

Selección de variables

```
# Seleccionar las variables
variables <- data[, c("Protein", "Carbohydrates")]

# Seleccionar la variable Protein
protein_data <- data$Protein

# Número de datos en la variable Protein
num_datos_protein <- length(protein_data)
num_datos_protein
```

```
## [1] 260
```

```
# Seleccionar la variable Carbohydrates
carbohydrates_data <- data$Carbohydrates

# Número de datos en la variable Carbohydrates
num_datos_carbohydrates <- length(carbohydrates_data)
num_datos_carbohydrates
```

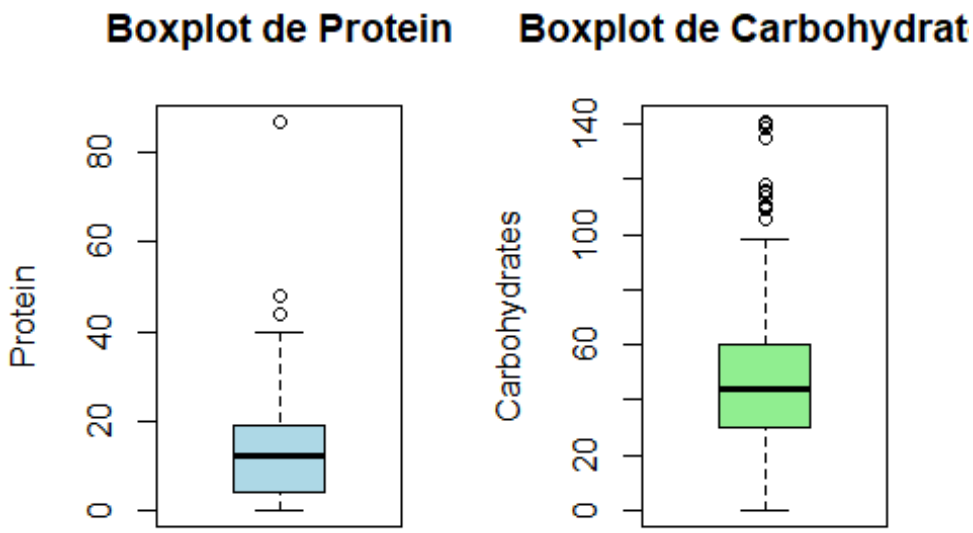
```
## [1] 260
```

1. Analizar datos atípicos

Diagrama de caja y bigote

```
# Graficar el diagrama de caja y bigote para Protein y Carbohydrates

par(mfrow=c(1, 2)) # Configurar la pantalla para mostrar 1x2 gráficos

# Boxplot para Protein
boxplot(protein_data, main="Boxplot de Protein", ylab="Protein",
col="lightblue")

# Boxplot para Carbohydrates
```

```
boxplot(carbohydrates_data, main="Boxplot de Carbohydrates",
ylab="Carbohydrates", col="lightgreen")
```



Interpretación de gráficas

Protein:

La mediana pareciera estar cercana al valor de 15g, esto significa que la mitad de los elementos del menú tienen un contenido de proteínas menor o igual a 15g y la otra mitad tiene un contenido mayor.

Aproximadamente el 50% de los datos se encuentran entre el primer (Q1) y tercer cuartil (Q3) que es aproximadamente entre el valor de 10 y 25 gramos respectivamente.

Se puede apreciar que existe una baja variabilidad en los datos debido al tamaño de la caja, sin embargo existen valores atípicos por encima del bigote superior, lo que indica que hay alimentos en el menú con un contenido de proteínas significativamente mayor que la mayoría. Estos valores se encuentran cerca del valor de 45g y 50g y otro más en el de 90g.

Carbohydrates:

La mediana se encuentra alrededor de 45g. Esto significa que la mitad de los elementos del menú tienen un contenido de carbohidratos menor o igual a 45g, y la otra mitad tiene un contenido mayor.

Aproximadamente el 50% de los datos se encuentran entre el primer (Q1) y tercer cuartil (Q3) que es aproximadamente entre el valor de 30 y 60 gramos respectivamente.

Hay varios puntos por encima del bigote superior, indicando alimentos con un contenido de carbohidratos significativamente mayor que la mayoría. Estos valores atípicos están alrededor de 120 gramos.

Rango intercuartílico y los cuartíles

```
# Calcular los cuartiles y el rango intercuartílico (IQR) para Protein
quartiles_protein <- quantile(protein_data, probs=c(0.25, 0.5, 0.75))

IQR_protein <- IQR(protein_data)

# Calcular los cuartiles y el rango intercuartílico (IQR) para
Carbohydrates
quartiles_carbohydrates <- quantile(carbohydrates_data, probs=c(0.25,
0.5, 0.75))

IQR_carbohydrates <- IQR(carbohydrates_data)


# Tabla con resultados
resultados <- data.frame(
  Variable = c("Protein", "Carbohydrates"),
  Q1 = c(quartiles_protein[1], quartiles_carbohydrates[1]),
  Q2 = c(quartiles_protein[2], quartiles_carbohydrates[2]),
  Q3 = c(quartiles_protein[3], quartiles_carbohydrates[3]),
  IQR = c(IQR_protein, IQR_carbohydrates)
)

resultados
```

```
##        Variable Q1 Q2 Q3 IQR
## 1       Protein  4 12 19  15
## 2 Carbohydrates 30 44 60  30
```

Interpretación de resultados

Protein:

1er cuartil (Q1): El 25% de los elementos del menú tienen un contenido de proteínas menor o igual a 4 gramos.

Mediana (Q2): La mediana del contenido de proteínas es de 12 gramos, lo que significa que el 50% de los elementos tienen un contenido de proteínas menor o igual a este valor.

3er Cuartil (Q3): El 75% de los elementos del menú tienen un contenido de proteínas menor o igual a 19 gramos.

Rango Intercuartílico (IQR): La diferencia entre Q3 y Q1 (19 - 4), indica la dispersión del contenido de proteínas en el 50% central de los datos.

Carbohydrates:

Primer Cuartil (Q1): El 25% de los elementos del menú tienen un contenido de carbohidratos menor o igual a 30 gramos.

Mediana (Q2): La mediana del contenido de carbohidratos es de 44 gramos, indicando que el 50% de los elementos tienen un contenido de carbohidratos menor o igual a este valor.

Tercer Cuartil (Q3): El 75% de los elementos del menú tienen un contenido de carbohidratos menor o igual a 60 gramos.

Rango Intercuartílico (IQR): La diferencia entre Q3 y Q1 (60 - 30), indica la dispersión del contenido de carbohidratos en el 50% central de los datos.

Comparando ambas variables:

El contenido de proteínas tiene una menor dispersión (IQR = 15 gramos) en comparación con los carbohidratos. Esto indica que los valores de proteínas están más concentrados alrededor de la mediana y que los carbohidratos tienen una mayor variabilidad en los valores.

La mediana del contenido de carbohidratos (44 gramos) es significativamente mayor que la mediana del contenido de proteínas (12 gramos). Esto indica que los alimentos en el menú tienden a tener más carbohidratos que proteínas.

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

```
# Calcular los límites inferior y superior para detectar datos atípicos en Protein
limite_inferior_protein <- quartiles_protein[1] - 1.5 * IQR_protein
limite_superior_protein <- quartiles_protein[3] + 1.5 * IQR_protein

# Calcular los cuartiles y el rango intercuartílico (IQR) para Carbohydrates
quartiles_carbohydrates <- quantile(carbohydrates_data, probs=c(0.25, 0.5, 0.75))
IQR_carbohydrates <- IQR(carbohydrates_data)

# Calcular los límites inferior y superior para detectar datos atípicos en Carbohydrates
limite_inferior_carbohydrates <- quartiles_carbohydrates[1] - 1.5 * IQR_carbohydrates
limite_superior_carbohydrates <- quartiles_carbohydrates[3] + 1.5 * IQR_carbohydrates

# Tabla con resultados
```

```
resultados_limites <- data.frame(
  Variable = c("Protein", "Carbohydrates"),
  Limite_Inferior = c(limite_inferior_protein,
limite_inferior_carbohydrates),
  Limite_Superior = c(limite_superior_protein,
limite_superior_carbohydrates)
)

resultados_limites
```

```
##        Variable Limite_Inferior Limite_Superior
## 1       Protein           -18.5            41.5
## 2 Carbohydrates           -15.0           105.0
```

Interpretación de resultados

De acuerdo con el criterio de 1.5 veces el rango intercuartílico, cualquier valor de proteínas por debajo de -18.5 gramos o por encima de 41.5 gramos se consideraría un dato atípico. Debido a que el contenido de proteínas no puede ser negativo, solo los valores por encima de 41.5 gramos se considerarían atípicos.

Por otro lado, cualquier valor de carbohidratos por debajo de -15.0 gramos o por encima de 105.0 gramos se consideraría un dato atípico. Similarmente, el contenido de carbohidratos no puede ser negativo, así que solo los valores por encima de 105.0 gramos se considerarían atípicos.

Ante estos resultados y observando las gráficas anteriores, si hay datos atípicos de acuerdo con este criterio

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

```
# Calcular la media y la desviación estándar para Protein
media_protein <- mean(protein_data, na.rm = TRUE)
sd_protein <- sd(protein_data, na.rm = TRUE)


# Calcular los límites inferior y superior para detectar datos atípicos
en Protein
limite_inferior_protein <- media_protein - 3 * sd_protein
limite_superior_protein <- media_protein + 3 * sd_protein

# Calcular la media y la desviación estándar para Carbohydrates
media_carbohydrates <- mean(carbohydrates_data, na.rm = TRUE)
sd_carbohydrates <- sd(carbohydrates_data, na.rm = TRUE)


# Calcular los límites inferior y superior para detectar datos atípicos
en Carbohydrates
limite_inferior_carbohydrates <- media_carbohydrates - 3 *
```

```r
sd_carbohydrates
limite_superior_carbohydrates <- media_carbohydrates + 3 *
sd_carbohydrates

# Tabla con resultados
resultados_limites <- data.frame(
  Variable = c("Protein", "Carbohydrates"),
  Limite_Inferior = c(limite_inferior_protein,
limite_inferior_carbohydrates),
  Limite_Superior = c(limite_superior_protein,
limite_superior_carbohydrates)
)

resultados_limites
```

```
##        Variable Limite_Inferior Limite_Superior
## 1       Protein       -20.93998         47.6169
## 2 Carbohydrates       -37.41054        132.1028
```

Interpretación de resultados

Cualquier valor de proteínas por debajo de -20.93998 gramos o por encima de 47.6169 gramos se consideraría un dato atípico. Al igual que en el criterio anterior, solo se consideran los valores positivos, es decir, los valores por encima de 47.6169 gramos se considerarían atípicos.

Similarmente, en el caso de los carbohidratos el contenido no puede ser negativo, así que solo los valores por encima de 132.1028 gramos se considerarían atípicos

Considerando este critero, si hay datos atípicos presentes en ambas variables.

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)

Argumentación

En este caso, se esta analizando el contenido de proteínas y carbohidratos en los alimentos del menú de McDonald's. Los datos atípicos pueden representar alimentos con contenidos significativamente diferentes en comparación con la mayoría del menú.

Al ser un análisis general de las variables, se decide mantener los datos atípicos, ya que estos reflejan la realidad completa del menú. Además, los alimentos que tienen un contenido significativamente diferente al resto pueden ser productos muy específicos como hamburguesas de carne grandes para el caso de la proteina o batidos y postres en el caso del alto contenido de carbohidratos.

2. Realiza pruebas de normalidad univariada

Para todas las pruebas de normalidad, la hipótesis nula $[H_0]$ es que los datos siguen una distribución normal.

Un valor p bajo (generalmente < 0.05) indica que podemos rechazar la hipótesis nula, lo que sugiere que los datos no siguen una distribución normal.

```
# Instalar paquete
library(nortest)

# Prueba de normalidad de Anderson-Darling para la variable Protein
ad_test_protein <- ad.test(protein_data)
ad_test_protein

##
##  Anderson-Darling normality test
##
## data:  protein_data
## A = 4.7515, p-value = 8.515e-12

# Prueba de normalidad de Anderson-Darling para la variable Carbohydrates
ad_test_carbohydrates <- ad.test(carbohydrates_data)
ad_test_carbohydrates

##
##  Anderson-Darling normality test
##
## data:  carbohydrates_data
## A = 4.1402, p-value = 2.547e-10

# Prueba de normalidad de Kolmogorov-Smirnov para la variable Protein
lillie_test_protein <- lillie.test(protein_data)
lillie_test_protein

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  protein_data
## D = 0.12153, p-value = 5.806e-10

# # Prueba de Lilliefors (Kolmogorov-Smirnov modificada) para la variable
Carbohydrates
lillie_test_carbhohydrates <- lillie.test(carbohydrates_data)
lillie_test_carbhohydrates

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  carbohydrates_data
## D = 0.098548, p-value = 2.081e-06
```

```r
# Instalar paquete
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 4.3.3
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
# Prueba de normalidad de Jarque-Bera para la variable Protein
jb_test_protein <- jarque.bera.test(protein_data)
jb_test_protein
```

```
##
##  Jarque Bera Test
##
## data:  protein_data
## X-squared = 479.38, df = 2, p-value < 2.2e-16
```

```r
# Prueba de normalidad de Jarque-Bera para la variable Carbohydrates
jb_test_carbohydrates <- jarque.bera.test(carbohydrates_data)
jb_test_carbohydrates
```

```
##
##  Jarque Bera Test
##
## data:  carbohydrates_data
## X-squared = 55.646, df = 2, p-value = 8.251e-13
```

```r
# Prueba de Cramer-Von Mises para la variable Protein
cvm_test_protein <- cvm.test(protein_data)
cvm_test_protein
```

```
##
##  Cramer-von Mises normality test
##
## data:  protein_data
## W = 0.67776, p-value = 8.993e-08
```

```r
# Prueba de Cramer-Von Mises para la variable Carbohydrates
cvm_test_carbohydrates <- cvm.test(carbohydrates_data)
cvm_test_carbohydrates
```

```
##
##  Cramer-von Mises normality test
##
## data:  carbohydrates_data
## W = 0.63589, p-value = 1.862e-07
```

Interpretación de resultados

Prueba de Anderson-Darling:

Los valores p para ambas variables (proteínas y carbohidratos) son extremadamente bajos ($< 10^{-10}$), lo que indica que podemos rechazar la hipótesis nula de normalidad para ambas variables. Esto sugiere que los datos de proteínas y carbohidratos no siguen una distribución normal.

Prueba de Lilliefors: Los valores p para ambas variables son también extremadamente bajos ($< 10^{-6}$), lo que confirma nuevamente que los datos no siguen una distribución normal.

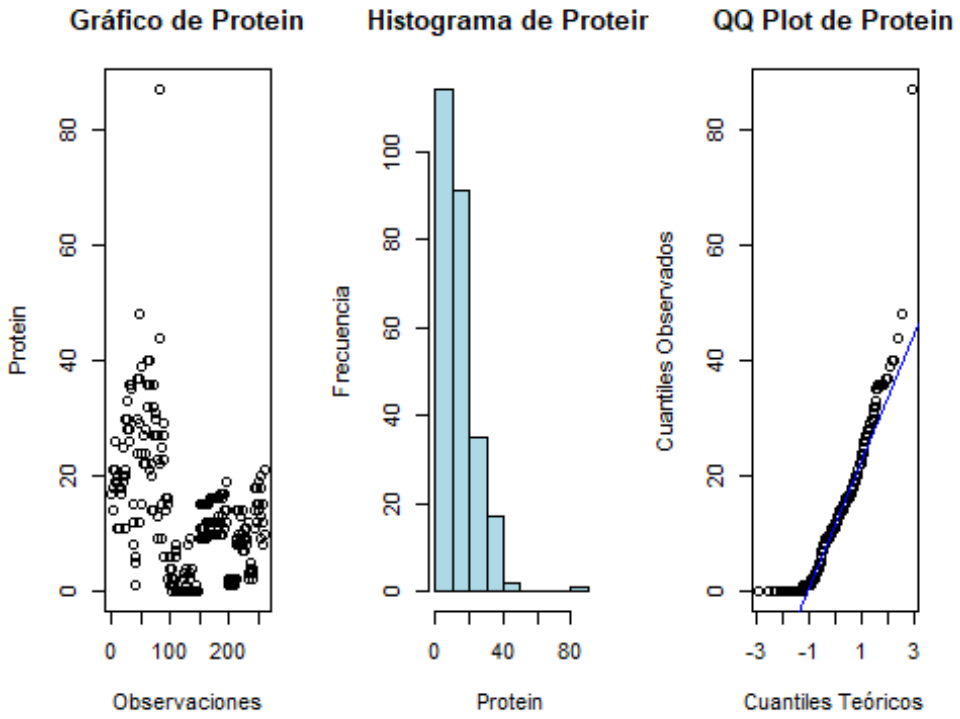
Prueba de Jarque-Bera: Los valores p son muy bajos ($< 10^{-13}$), indicando que los datos de proteínas y carbohidratos no siguen una distribución normal. La prueba de Jarque-Bera es especialmente sensible a la asimetría y la curtosis, y estos resultados sugieren que hay desviaciones significativas en estos aspectos.

Prueba de Cramer-von Mises: Los valores p para ambas variables son bajos ($< 10^{-7}$), lo que indica que los datos no siguen una distribución normal.

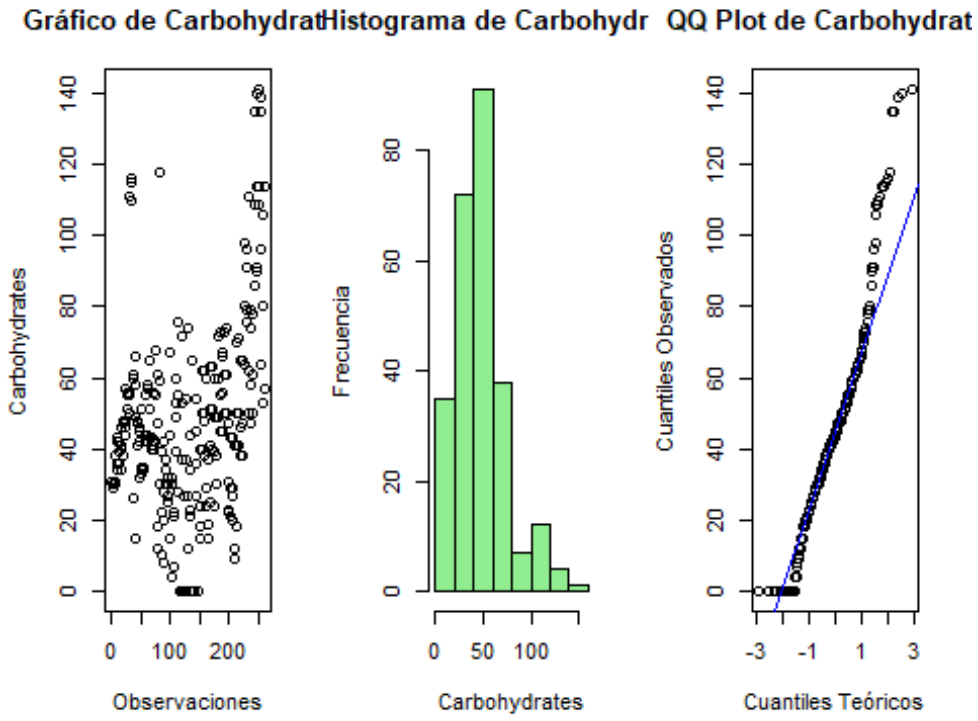
Todas las pruebas de normalidad realizadas (Anderson-Darling, Lilliefors, Jarque-Bera, y Cramer-von Mises) indican consistentemente que los datos de proteínas y carbohidratos del menú de McDonald's no siguen una distribución normal. Este resultado es importante para la elección de métodos estadísticos apropiados para un análisis posterior. En lugar de técnicas que asumen normalidad, se podría necesitar transformar los datos para acercarlos a una distribución normal.

3. Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos) para cada variable

```
# Graficar los datos de Protein (histograma) y QQ plot
par(mfrow=c(1,3))  # Dividir el área de gráficos en 1 fila y 2 columnas

# Gráficar los datos de la variable "Protein"
plot(data$Protein, main = "Gráfico de Protein", xlab = "Observaciones",
ylab = "Protein")

# Histograma para la variable "Protein"
hist(data$Protein, main = "Histograma de Protein", xlab = "Protein", ylab
= "Frecuencia",
     col = "lightblue", border = "black")

# QQ plot para la variable "Protein"
qqnorm(data$Protein, main = "QQ Plot de Protein", xlab = "Cuantiles
Teóricos",
       ylab = "Cuantiles Observados")
qqline(data$Protein, col = "blue")
```




```r
# Gráficar los datos de la variable "Carbohydrates"
plot(data$Carbohydrates, main = "Gráfico de Carbohydrates", xlab =
"Observaciones", ylab = "Carbohydrates")

# Histograma para la variable "Carbohydrates"
hist(data$Carbohydrates, main = "Histograma de Carbohydrates", xlab =
"Carbohydrates", ylab = "Frecuencia",
     col = "lightgreen", border = "black")

# QQ plot para la variable "Carbohydrates"
qqnorm(data$Carbohydrates, main = "QQ Plot de Carbohydrates", xlab =
"Cuantiles Teóricos",
       ylab = "Cuantiles Observados")
qqline(data$Carbohydrates, col = "blue")
```




```
# Restablecer la configuración de múltiples gráficos
par(mfrow=c(1,1))
```

Interpretación de gráficas

Protein: El gráfico de dispersión muestra la distribución de los valores de proteínas a lo largo de las observaciones. Hay algunos valores atípicos que alcanzan hasta más de 80 gramos.

El histograma muestra la frecuencia de los valores de proteínas en intervalos específicos. La mayoría de los alimentos tienen un contenido de proteínas entre 0 y 30 gramos, con una mayor concentración en el rango de 0 a 20 gramos. Posteriormente, existe una caída brusca en la frecuencia después de 30 gramos, con muy pocos alimentos superando este valor.

El gráfico Q-Q compara los cuantiles observados de la variable “Protein” con los cuantiles esperados de una distribución normal. Si los datos siguieran una distribución normal, los puntos se alinearían a lo largo de la línea diagonal, sin embargo, los puntos se desvían significativamente, especialmente en los extremos. Esto indica que los datos no siguen una distribución normal.

Carbohydrates: El gráfico de dispersión muestra que hay algunos valores atípicos que alcanzan hasta 130 gramos, lo cual indica la presencia de alimentos con contenido de carbohidratos significativamente más alto.

El histograma muestra que la mayoría de los alimentos tienen un contenido de carbohidratos entre 0 y 60 gramos, con una mayor concentración en el rango de 20 a

50 gramos. Posteriormente, hay una caída en la frecuencia después de 60 gramos, con muy pocos alimentos superando este valor.

El gráfico Q-Q muestra que los puntos se desvían significativamente de la línea diagonal, especialmente en los extremos. Esto indica que los datos no siguen una distribución normal.

4. Calcula el coeficiente de sesgo y el coeficiente de curtosis de cada variable.

```
# Instalar paquete
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.3.3
```

```
# Calcular el coeficiente de sesgo y el coeficiente de curtosis para
Protein
sesgo_protein <- skewness(protein_data, na.rm = TRUE)
curtosis_protein <- kurtosis(protein_data, na.rm = TRUE)

# Calcular el coeficiente de sesgo y el coeficiente de curtosis para
Carbohydrates
sesgo_carbohydrates <- skewness(carbohydrates_data, na.rm = TRUE)
curtosis_carbohydrates <- kurtosis(carbohydrates_data, na.rm = TRUE)

# Tabla con resultados
resultados_sesgo_curtosis <- data.frame(
  Variable = c("Protein", "Carbohydrates"),
  Sesgo = c(sesgo_protein, sesgo_carbohydrates),
  Curtosis = c(curtosis_protein, curtosis_carbohydrates)
)



resultados_sesgo_curtosis
```

```
##        Variable     Sesgo Curtosis
## 1       Protein 1.5617406 5.795500
## 2 Carbohydrates 0.9021952 1.324083
```

Interpretación de resultados

Sesgo:

Protein: El sesgo es positivo, indicando que la distribución de la variable Protein está inclinada hacia la derecha, es decir, hay una cola larga en el extremo derecho de la distribución, por la presencia de valores atípicos altos.

Carbohydrates: El sesgo es positivo, sin embargo, es menor en comparación con la variable Protein, lo que sugiere que la distribución de "Carbohydrates" es más simétrica.

Curtosis:

Protein: La curtosis indica que la distribución es leptocúrtica (más alta y con colas más pesadas que una distribución normal). Esto sugiere una alta concentración de valores cerca de la media, pero con valores atípicos extremos.

Carbohydrates: La curtosis sugiere una distribución ligeramente más plana que la distribución normal (mesocúrtica), pero aún con presencia de valores extremos.

5. Compara las medidas de media, mediana y rango medio de cada variable.

```
# Calcular las estadísticas para Protein
media_protein <- mean(protein_data, na.rm = TRUE)
mediana_protein <- median(protein_data, na.rm = TRUE)
rango_medio_protein <- (min(protein_data, na.rm = TRUE) +
max(protein_data, na.rm = TRUE)) / 2

# Calcular las estadísticas para Carbohydrates
media_carbohydrates <- mean(carbohydrates_data, na.rm = TRUE)
mediana_carbohydrates <- median(carbohydrates_data, na.rm = TRUE)
rango_medio_carbohydrates <- (min(carbohydrates_data, na.rm = TRUE) +
max(carbohydrates_data, na.rm = TRUE)) / 2

# Tabla con resultados
resultados_medidas <- data.frame(
  Variable = c("Protein", "Carbohydrates"),
  Media = c(media_protein, media_carbohydrates),
  Mediana = c(mediana_protein, mediana_carbohydrates),
  Rango_Medio = c(rango_medio_protein, rango_medio_carbohydrates)
)

resultados_medidas
```

```
##        Variable    Media Mediana Rango_Medio
## 1       Protein 13.33846      12        43.5
## 2 Carbohydrates 47.34615      44        70.5
```

Interpretación de resultados

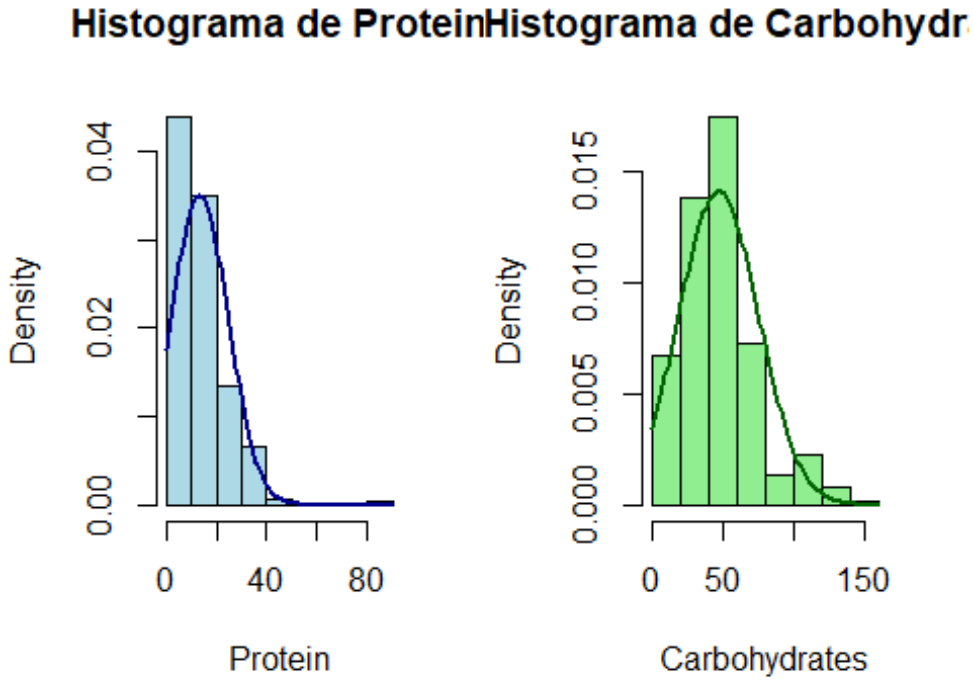
La media y la mediana de carbohidratos son significativamente mayores que las de proteínas, esto sugiere que los alimentos en el menú tienen en promedio un contenido de carbohidratos más alto que el de proteínas.

La diferencia entre la media y la mediana en ambas variables indica la presencia de outliers. Para proteínas, la media es ligeramente mayor que la mediana, lo que indica un sesgo positivo. Para carbohidratos, esta diferencia es menos pronunciada.

El rango medio de carbohidratos es mayor que el de proteínas, lo que sugiere una mayor dispersión y variabilidad en los valores de carbohidratos.

5. Realiza el histograma y su distribución teórica de probabilidad

```
# Crear gráficos
par(mfrow=c(1, 2)) # Configurar la pantalla para mostrar 1x2 gráficos

# Histograma y distribución teórica para Protein
hist(protein_data, prob=TRUE, main="Histograma de Protein",
xlab="Protein", col="lightblue", border="black")
curve(dnorm(x, mean=mean(protein_data, na.rm=TRUE), sd=sd(protein_data,
na.rm=TRUE)),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")

# Histograma y distribución teórica para Carbohydrates
hist(carbohydrates_data, prob=TRUE, main="Histograma de Carbohydrates",
xlab="Carbohydrates", col="lightgreen", border="black")
curve(dnorm(x, mean=mean(carbohydrates_data, na.rm=TRUE),
sd=sd(carbohydrates_data, na.rm=TRUE)),
      col="darkgreen", lwd=2, add=TRUE, yaxt="n")
```



###

Interpretación de gráficas

Protein: La distribución está fuertemente sesgada hacia la derecha con una concentración alta de valores bajos y algunos valores atípicos altos. La curva de densidad confirma esta asimetría.

Carbohydrates: La distribución también está sesgada hacia la derecha, pero de manera menos pronunciada.

6. Identifica cómo influyen los datos atípicos en la normalidad de los datos

Los datos atípicos pueden tener un impacto significativo en la distribución y normalidad de los datos.

Introducen Sesgo: Los outliers aumentan el sesgo positivo, desviando la distribución hacia la derecha.

Aumentan la Curtosis: Los outliers elevan la curtosis, resultando en colas más pesadas y picos más altos en la distribución.

Desvían Pruebas de Normalidad: Todas las pruebas de normalidad confirman que los datos no son normales, principalmente debido a los valores atípicos.

Impactan Medidas de Tendencia Central: La media es particularmente sensible a los outliers, lo que resulta en una mayor diferencia con la mediana.