

Estadística Descriptiva

Esteban Vizhñay

2024-07-02

```
library(moments)

library(readr)

## Warning: package 'readr' was built under R version 4.4.1
alimentos <- read_csv("Combo.csv")

## Rows: 500 Columns: 4
## -- Column specification -----
## Delimiter: ","
## chr (2): Papas, Refresco
## dbl (2): Carne, Salsa
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
head(alimentos,5)

## # A tibble: 5 x 4
##   Carne Salsa Papas Refresco
##   <dbl> <dbl> <chr>   <chr>
## 1  91.9  5.84 Medianas Mediano
## 2  89.2  5.59 Grandes Chico
## 3  88.2  5.48 Grandes Grande
## 4  86.8  5.26 Grandes Chico
## 5  88.8  5.55 Grandes Chico
```

1. Describa el comportamiento de la variable carne teniendo en cuenta la tendencia, variabilidad, distribución, valores atípicos entre otros.

```
carnes = alimentos$Carne
```

Visualizamos los primeros 30 valores

```
head(carnes, 30)

## [1] 91.8712 89.1723 88.2496 86.8345 88.8241 88.1539 87.1731 87.5638 91.2636
## [10] 93.4143 89.3216 90.2916 90.6802 91.1281 93.1330 90.5397 91.3261 88.7181
## [19] 90.3377 91.5828 90.4269 91.8140 89.2641 88.1940 91.3689 91.3685 90.8211
## [28] 88.1409 88.5290 89.9937
```

Medidas de tendencia central

Media

```
mean(carnes)
```

```
## [1] 90.02181
```

Mediana

```
median(carnes)
```

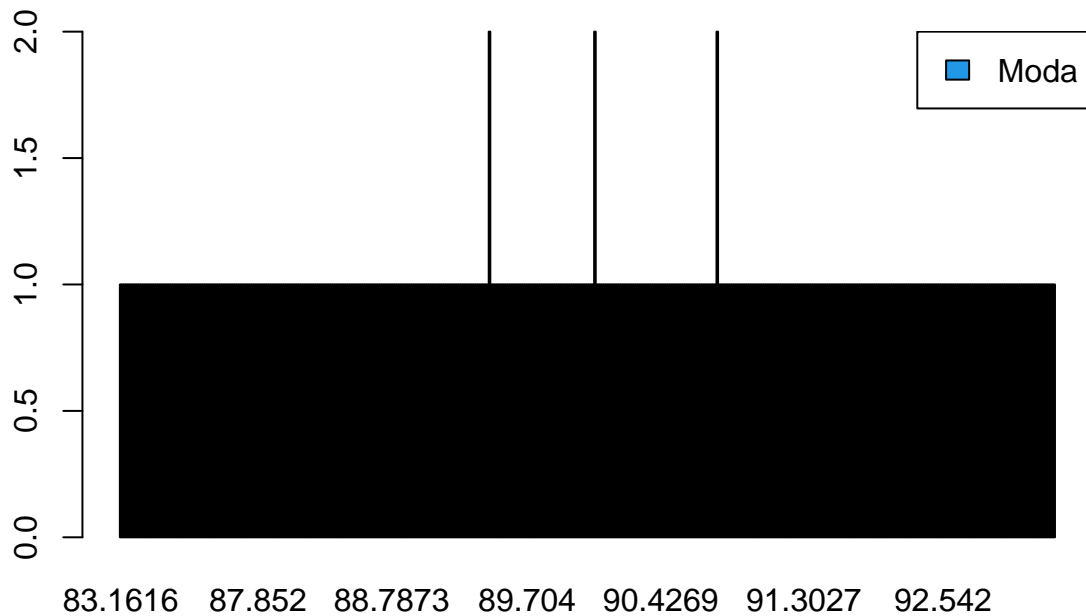
```
## [1] 90.02615
```

Moda

No es posible obtener una moda para este conjunto de datos debido a la naturaleza de los mismos ya que son variables continuas.

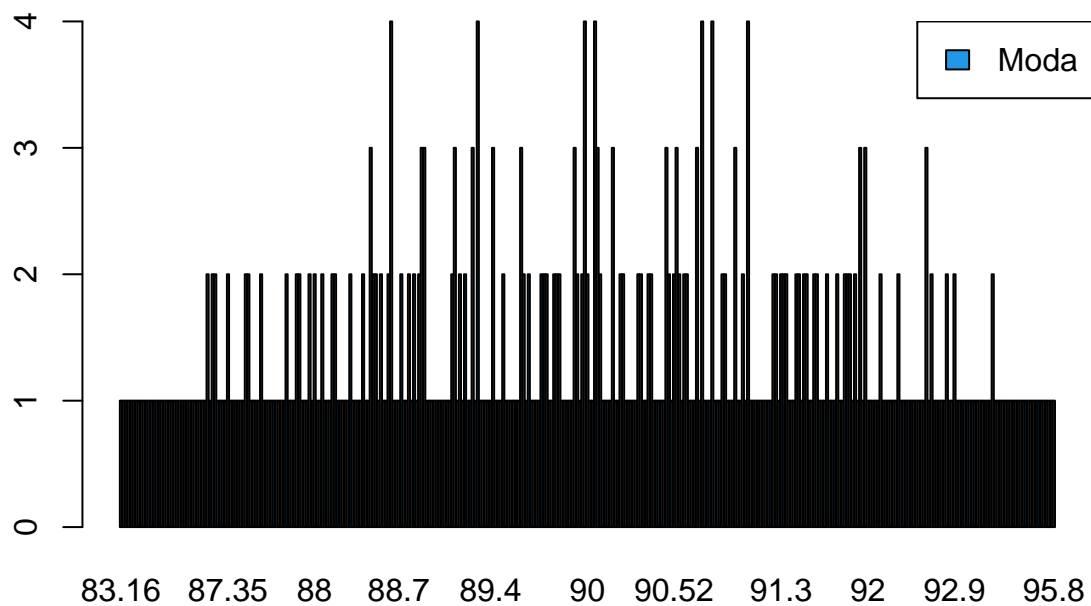
Una forma de evidenciar esto es graficar la tabla de frecuencias de “carne”

```
barplot(table(carnes), col = c(4, rep("gray", 4)))  
legend("topright", "Moda", fill = 4)
```



Se puede redondear los valores en 2 y generar la gráfica.

```
barplot(table(round(carnes,2)), col = c(4, rep("gray", 4)))  
legend("topright", "Moda", fill = 4)
```



Distribuciones

Asimetría y Curtosis

```
skewness(carnes)
```

```
## [1] 0.008819527
```

```
kurtosis(carnes)
```

```
## [1] 2.911783
```

Ahora consideramos los siguientes casos para los valores de asimetría y curtosis:

El valor de la asimetría es mayor o igual que -0.5 y menor o igual que 0.5 entonces podemos afirmar que la distribución es aproximadamente simétrica.

El valor de la curtosis es mayor a 0.5 podemos afirmar que los valores de la distribución tienen una gran concentración en torno a la media del conjunto.

Medidas de dispersión

Rango

```
rango = max(carnes) - min(carnes)
rango
```

```
## [1] 12.637
```

El valor anterior indica que los valores del conjunto de datos se encuentran en un intervalo pequeño.

Desviación estándar

```
sd(carnes)
```

```
## [1] 2.042254
```

El valor anterior indica que los valores del conjunto de datos están cerca de la media.

Más precisamente dentro del siguiente rango

```
rango_inferior = mean(carnes) - sd(carnes)
```

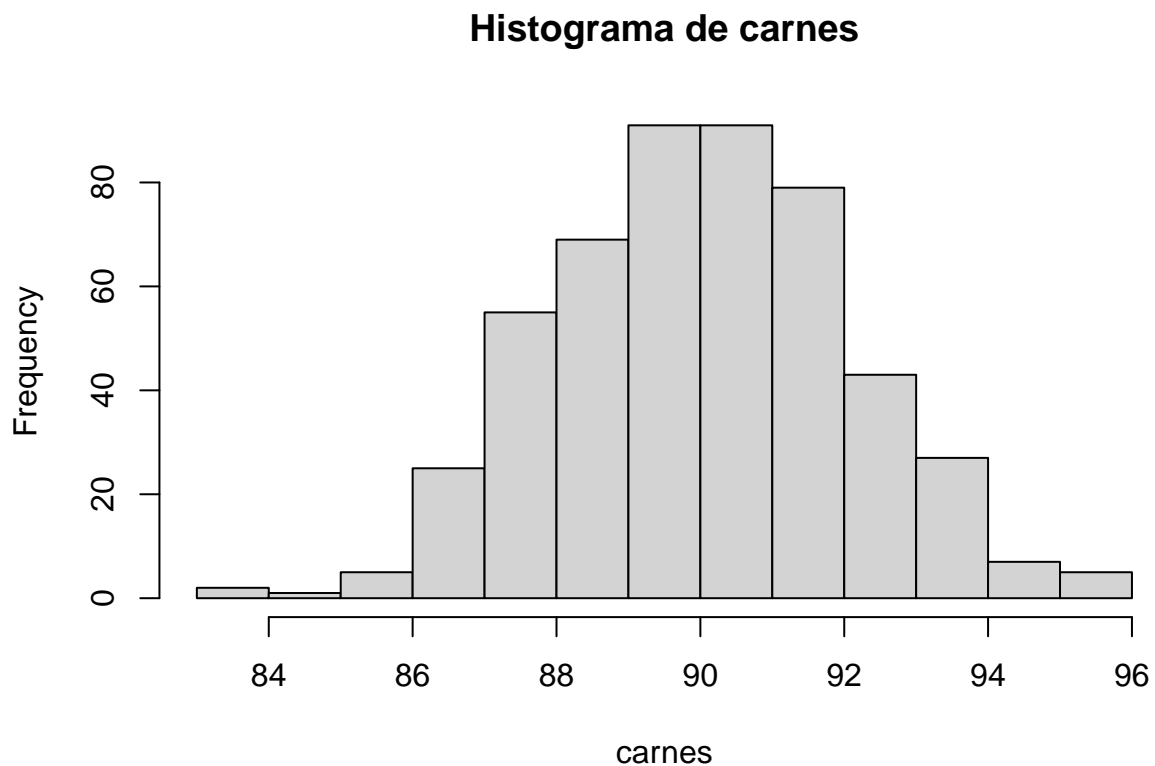
```
rango_superior = mean(carnes) + sd(carnes)
```

```
print(paste("[", rango_inferior, ",", rango_superior, "]"))
```

```
## [1] "[ 87.9795534760017 , 92.0640605239983 ]"
```

Una forma de poder evidenciar las conclusiones anteriores es crear un histograma del conjunto de datos

```
hist(carnes, main="Histograma de carnes")
```



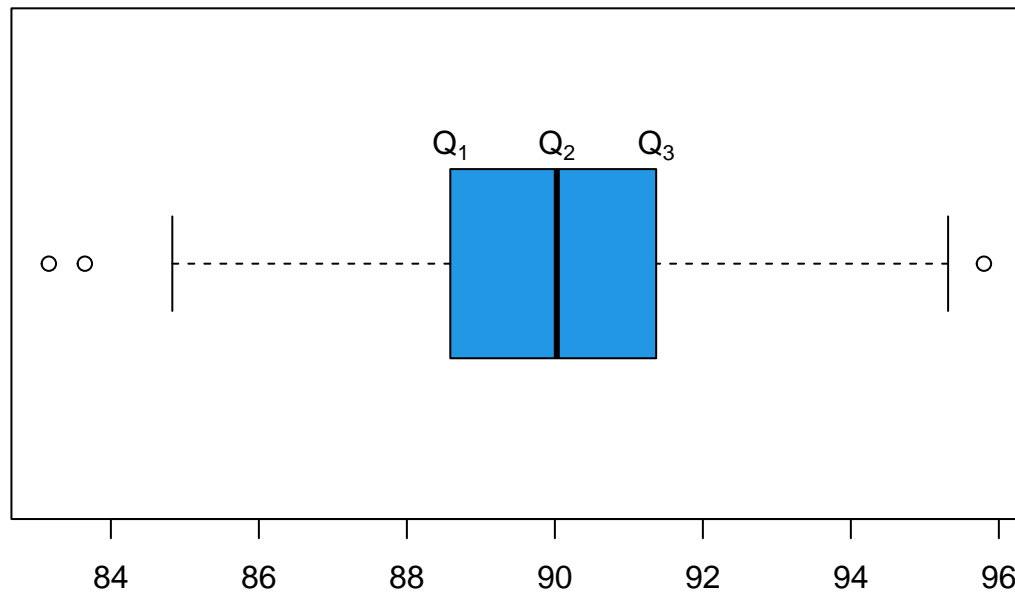
Medidas de posición

Cuartiles

```
cuantiles_carnes = quantile(carnes)
cuantiles_carnes
```

```
##          0%          25%          50%          75%         100%
```

```
## 83.16160 88.59600 90.02615 91.36860 95.79860
boxplot(carnes, col = 4, horizontal = TRUE)
text(cuantiles_carnes[2], 1.25, expression(Q[1]))
text(cuantiles_carnes[3], 1.25, expression(Q[2]))
text(cuantiles_carnes[4], 1.25, expression(Q[3]))
```



Vamos a revisar el rango intercuartílico

```
cuantiles_carnes[4] - cuantiles_carnes[2]
```

```
##      75%
## 2.7726
```

Con el valor obtenido podemos afirmar que los datos se encuentran agrupados con respecto a la media.

Valores atípicos

Una de las formas más utilizadas para calcular valores atípicos es

El promedio $\pm n$ * desviación estándar

- Para este caso n será igual a 3

https://help.highbond.com/helpdocs/analytics/141/user-guide/es/Content/analyzing_data/identifying_outliers.htm

```
limite_superior_carnes_at = mean(carnes) + (3 * sd(carnes))
limite_inferior_carnes_at = mean(carnes) - (3 * sd(carnes))
print(paste("[", limite_inferior_carnes_at, ",", limite_superior_carnes_at, "]"))
```

```
## [1] "[ 83.895046428005 , 96.148567571995 ]"
```

A partir de los límites obtenidos se pueden apreciar los valores atípicos

```
print( carnes[carnes < limite_inferior_carnes_at])
```

```
## [1] 83.6483 83.1616
```

```
print( carnes[carnes > limite_superior_carnes_at])
```

```
## numeric(0)
```

2. Describa el comportamiento de la variable salsa teniendo en cuenta la tendencia, variabilidad, distribución, valores atípicos entre otros.

Extraer carne del conjunto original

```
salsas = alimentos$Salsa
```

Visualizamos los primeros 30 valores

```
head(salsas, 30)
```

```
## [1] 5.83534 5.59279 5.47612 5.25916 5.55034 5.35418 5.20734 5.29052 5.84538
## [10] 5.76569 5.64502 5.51890 5.85529 5.65606 6.08032 5.65248 5.61313 5.46856
## [19] 5.46777 5.81372 5.46641 5.60135 5.53410 5.58889 5.69727 5.58146 5.74311
## [28] 5.22141 5.39952 5.54560
```

Medidas de tendencia central

Media

```
mean(salsas)
```

```
## [1] 5.60811
```

Mediana

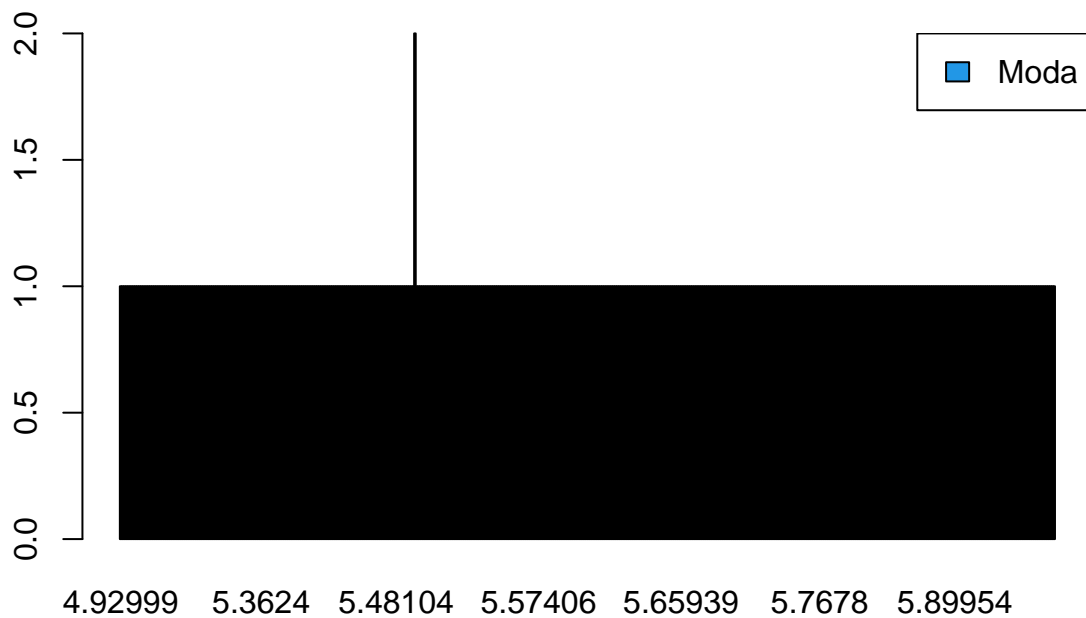
```
median(salsas)
```

```
## [1] 5.60025
```

Moda

Como primer paso, se gráfica la tabla de frecuencias de “salsa”

```
barplot(table(salsas), col = c(4, rep("gray", 4)))
legend("topright", "Moda", fill = 4)
```



De igual manera que el conjunto de datos de “carnes” no es posible obtener una moda.

Distribuciones

Asimetría y Curtosis

```
skewness(salsas)
```

```
## [1] 0.09954532
```

```
kurtosis(salsas)
```

```
## [1] 2.972769
```

Ahora consideramos los siguientes casos para los valores de asimetría y curtosis:

El valor de la asimetría es mayor o igual que -0.5 y menor o igual que 0.5 entonces podemos afirmar que la distribución es aproximadamente simétrica.

El valor de la curtosis es mayor a 0.5 podemos afirmar que los valores de la distribución tienen una gran concentración en torno a la media del conjunto.

Medidas de dispersión

Rango

```
max(salsas) - min(salsas)
```

```
## [1] 1.43587
```

El valor anterior indica que los valores del conjunto de datos se encuentran en un intervalo pequeño.

Desviación estándar

```
sd(salsas)
```

```
## [1] 0.2342666
```

El valor anterior indica que los valores del conjunto de datos están cerca de la media.

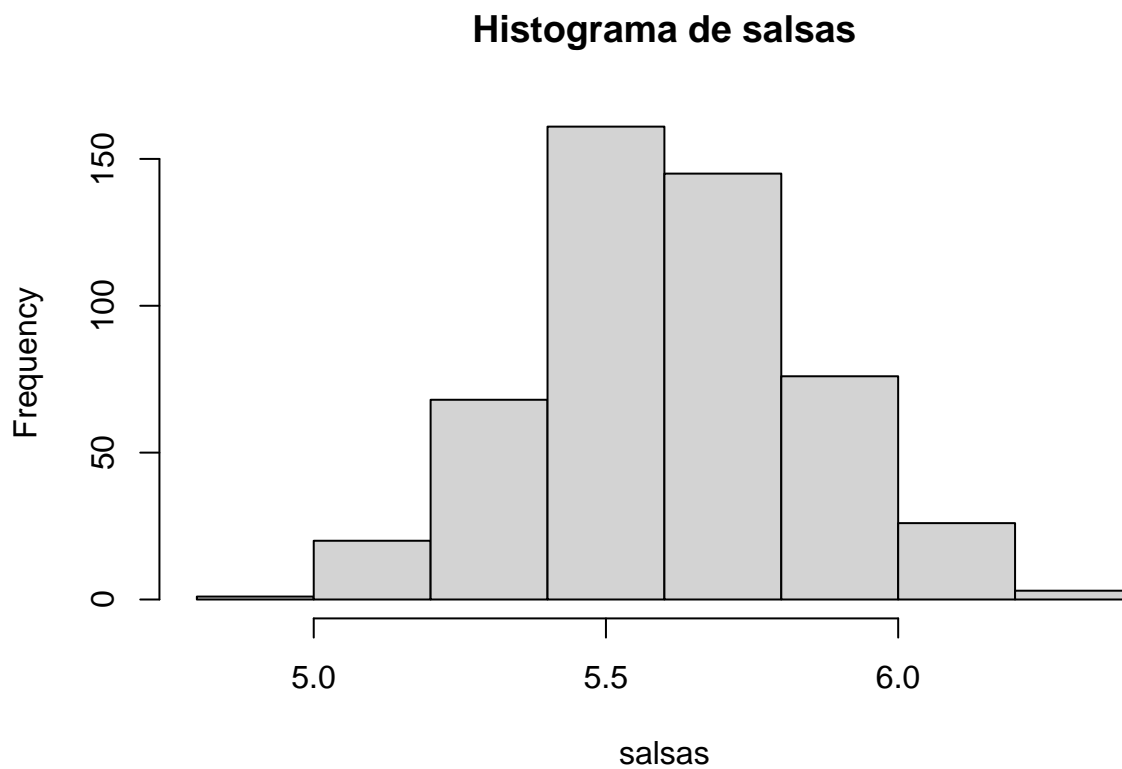
Más precisamente dentro del siguiente rango

```
rango_inferior = mean(salsas) - sd(salsas)
rango_superior = mean(salsas) + sd(salsas)
print(paste("[", rango_inferior, ",", rango_superior, "]"))
```

```
## [1] "[ 5.37384339219852 , 5.84237668780148 ]"
```

Una forma de poder evidenciar las conclusiones anteriores es crear un histograma del conjunto de datos

```
hist(salsas, main="Histograma de salsas")
```



Medidas de posición

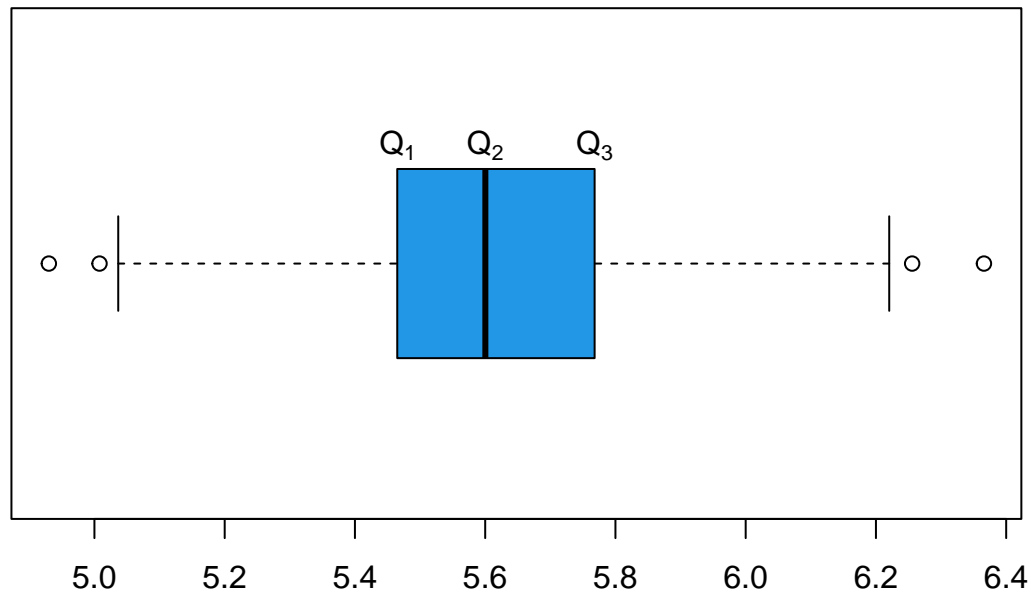
Cuartiles

```
cuantiles_salsas = quantile(salsas)
cuantiles_salsas
```



```
##      0%      25%      50%      75%     100%
## 4.92999 5.46569 5.60025 5.76792 6.36586
```

```
boxplot(salsas, col = 4, horizontal = TRUE)
text(cuantiles_salsas[2], 1.25, expression(Q[1]))
text(cuantiles_salsas[3], 1.25, expression(Q[2]))
text(cuantiles_salsas[4], 1.25, expression(Q[3]))
```



Vamos a revisar el rango intercuartílico

```
cuantiles_salsas[4] - cuantiles_salsas[2]
```

```
##      75%
## 0.30223
```

Con el valor obtenido podemos afirmar que los datos se encuentran agrupados con respecto a la media.

Valores atípicos

Una de las formas más utilizadas para calcular valores atípicos es

El promedio $\pm n$ * desviación estándar

- Para este caso n sera igual a 3

https://help.highbond.com/helpdocs/analytics/141/user-guide/es/Content/analyzing_data/identifying_outliers.htm

```
limite_superior_salsas_at = mean(salsas) + (3 * sd(salsas))
limite_inferior_salsas_at = mean(salsas) - (3 * sd(salsas))
```

```
print(paste("[", limite_inferior_salsas_at, ",", limite_superior_salsas_at, "]"))

## [1] "[ 4.90531009659556 , 6.31090998340444 ]"

A partir de los limites obtenidos se pueden apreciar los valores atípicos

print( salsas[salsas < limite_inferior_salsas_at])

## numeric(0)

print( salsas[salsas > limite_superior_salsas_at])

## [1] 6.36586
```

3. Describa el comportamiento de la variable papa

Extraer carne del conjunto original

```
papas = alimentos$Papas
```

Visualizamos los primeros 30 valores

```
head(papas, 30)

## [1] "Medianas" "Grandes" "Grandes" "Grandes" "Grandes" "Grandes"
## [7] "Grandes" "Medianas" "Grandes" "Medianas" "Medianas" "Chicas"
## [13] "Chicas" "Medianas" "Medianas" "Medianas" "Medianas" "Medianas"
## [19] "Grandes" "Medianas" "Medianas" "Medianas" "Grandes" "Grandes"
## [25] "Grandes" "Medianas" "Grandes" "Grandes" "Medianas" "Chicas"
```

Medidas de tendencia central

Moda

A pesar que el conjunto de datos no es numérico es posible obtener la moda ya que se busca el valor mas repetido de todo el conjunto.

```
table(papas)
```

```
## papas
## Chicas Grandes Medianas
##      87      166      247
```

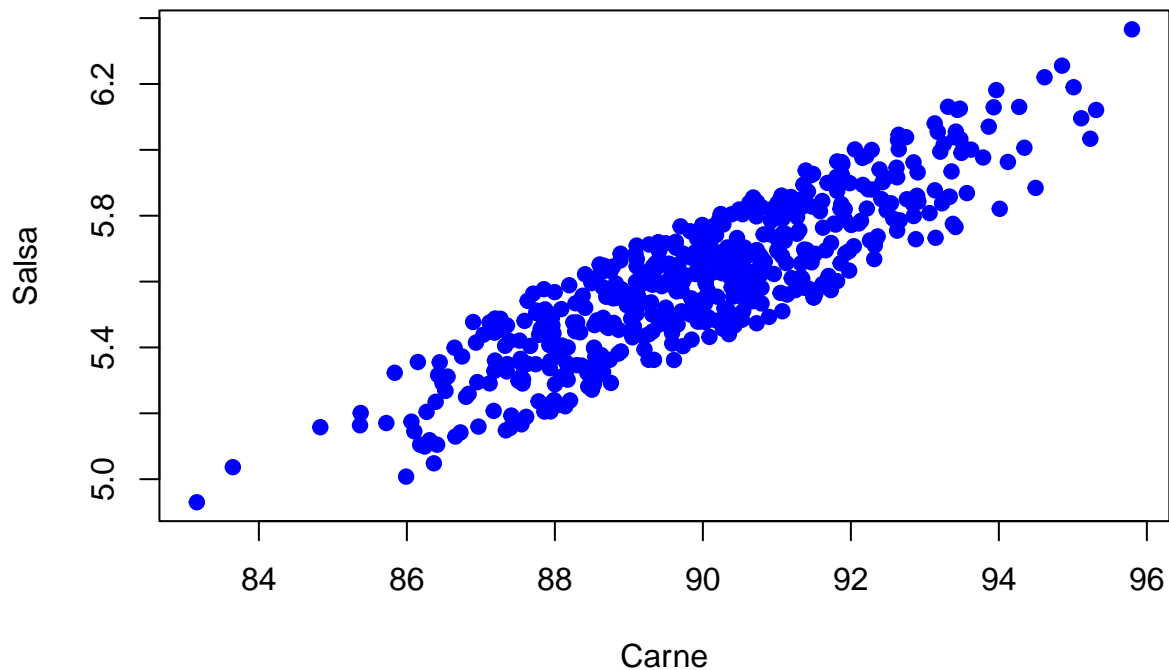
Para este caso el valor mas repetido es “Medianas”.

4. Describa el comportamiento bivariado de las variables carne y salsa, y calcule e interprete el coeficiente de correlación.

Obtenemos un gráfico de dispersión

```
# Crear un gráfico de dispersión
plot(carnes, salsas, main="Gráfico de Dispersión de Carne vs Salsa", xlab="Carne", ylab="Salsa", pch=19)
```

Gráfico de Dispersión de Carne vs Salsa



Coefficiente de correlación

```
cor(carnes, salsas)
```

```
## [1] 0.8816783
```

El valor anterior indica que existe una correlación fuerte entre carnes y salsas. A medida que aumente los valores de las carnes provoca que los valores de las salsas también aumenten.

5. ¿Qué combinación de papas y refresco es la más frecuente?

Obtenemos el conjunto de valores de refresco

```
refrescos = alimentos$Refresco
```

Obtenemos la tabla de frecuencias para los dos conjuntos de datos

```
papas_x_refrescos = table(papas, refrescos)
papas_x_refrescos
```

```
##           refrescos
## papas      Chico Grande Mediano
## Chicas      35      16      36
## Grandes     54      32      80
## Medianas    78      35     134
```

Obtenemos el valor máximo

```
max(papas_x_refrescos)
```

```
## [1] 134
```

Con el valor obtenido podemos observar que la combinación papas medianas y refresco mediano es la mas frecuente.

6. ¿Qué combinación de papas y refresco es la menos frecuente?

Obtenemos el valor mínimo

```
min(papas_x_refrescos)
```

```
## [1] 16
```

Con la tabla anterior podemos observar que la combinación papas chicas y refresco grande es la menos frecuente.

7. Calcula la probabilidad que hay de que un cliente seleccionado al azar haya pedido:

Papas medianas

$A = \text{Papas medianas}$

$P(A)$

La probabilidad A es

```
papas_medianas = alimentos [ which( alimentos$Papas=="Medianas"),]  
prob_a = (dim(papas_medianas)/length(papas))[1]  
prob_a
```

```
## [1] 0.494
```

Papas medianas o refresco chico

$A = \text{Papas medianas}$

$B = \text{Refresco chico}$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probabilidad de A

```
papas_medianas = alimentos [ which( alimentos$Papas=="Medianas"),]  
prob_a = (dim(papas_medianas)/length(papas))[1]
```

```
prob_a
```

```
## [1] 0.494
```

Probabilidad de B

```
refresco_chico = alimentos [ which(alimentos$Refresco=="Chico"),]  
prob_b = (dim(refresco_chico)/length(refrescos))[1]
```

```
prob_b
```

```
## [1] 0.334
```

Probabilidad de A intersección B

```
papas_medianas_refresco_chico = alimentos [ which(alimentos$Papas == "Medianas" & alimentos$Refresco == "Chico") ]
prob_a_b = (dim(papas_medianas_refresco_chico)/length(refrescos))[1]
```

```
prob_a_b
```

```
## [1] 0.156
```

Probabilidad de A unión B es

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

```
prob_a + prob_b - prob_a_b
```

```
## [1] 0.672
```

Papas grandes y refresco chico

La probabilidad de papas grandes y refresco chico es:

```
papas_grandes_refresco_chico = alimentos [ which(alimentos$Refresco == "Chico" & alimentos$Papas == "Grandes") ]
prob_a_b = (dim(papas_grandes_refresco_chico)/length(refrescos))[1]
prob_a_b
```

```
## [1] 0.108
```

Refresco chico si ya pidió papas grandes

A = Papas grandes

B = Refresco chico

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

La probabilidad de papas grandes y refresco chico es:

```
papas_grandes_refresco_chico = alimentos [ which(alimentos$Refresco == "Chico" & alimentos$Papas == "Grandes") ]
prob_a_b = (dim(papas_grandes_refresco_chico)/length(refrescos))[1]
prob_a_b
```

```
## [1] 0.108
```

La probabilidad de papas grandes

```
papas_grandes = alimentos [ which(alimentos$Papas == "Grandes"), ]
prob_a = (dim(papas_grandes)/length(papas))[1]
prob_a
```

```
## [1] 0.332
```

La probabilidad de refresco chico dado que pidió papas grandes es:

```
prob_a_b / prob_a
```

```
## [1] 0.3253012
```

9. ¿Los eventos papas grandes y refresco grande son independientes? Sí, No y Por qué.

$A = \text{Papas grandes}$

$B = \text{Refresco grande}$

Si son independientes cumplen la siguiente condición

$$P(A \cap B) = P(A) * P(B)$$

La probabilidad de papas grandes es:

```
papas_grandes = alimentos [ which(alimentos$Papas=="Grandes"),]  
prob_a = (dim(papas_grandes)/length(papas))[1]  
prob_a
```

```
## [1] 0.332
```

La probabilidad de refresco chico es:

```
refresco_chico = alimentos [ which(alimentos$Refresco=="Grande"),]  
prob_b = (dim(refresco_chico)/length(refrescos))[1]  
prob_b
```

```
## [1] 0.166
```

La probabilidad de papas grandes y refresco grande es:

```
papas_grandes_refresco_chico = alimentos [ which(alimentos$Refresco=="Grande" & alimentos$Papas=="Grandes"),]  
prob_a_b = (dim(papas_grandes_refresco_chico)/length(refrescos))[1]  
prob_a_b
```

```
## [1] 0.064
```

Validamos la condición

```
prob_a_b == prob_a * prob_b
```

```
## [1] FALSE
```

Como no cumplen la igualdad, podemos decir que los dos eventos no son independientes