# Taller 1 Multivariado

**Juan Jose Rodriguez Cubillos | Paula Sofia Torres Rodriguez**
**Mauricio Rodriguez Cordoba**
**Mariana Díaz Puentes**

**2024-08-10**

## Introducción

El objetivo del informe es realizar un análisis detallado para explorar la normalidad multivariada en un conjunto de datos. A lo largo del informe, se abordarán diferentes etapas del análisis, la obtención de estadísticas descriptivas, la visualización de correlaciones, y la aplicación de pruebas informales de normalidad multivariada. Este análisis permitirá determinar si los datos cumplen con las suposiciones de normalidad multivariada.

# Punto 1

En el archivo anexo1.csv contiene los datos de una muestra de vectores aleatorios de 3 componentes, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{40}$.

## a. Evalue la normalidad multivariada de la muestra dada.

### Verificacion de datos faltantes

Examinar los datos faltantes, con el fin de lograr asegurar que los resultados no estén sesgados ni incorrectos.

| | Total Datos Nulos |
|---|---|
| X1 | 0 |
| X2 | 0 |
| X3 | 0 |

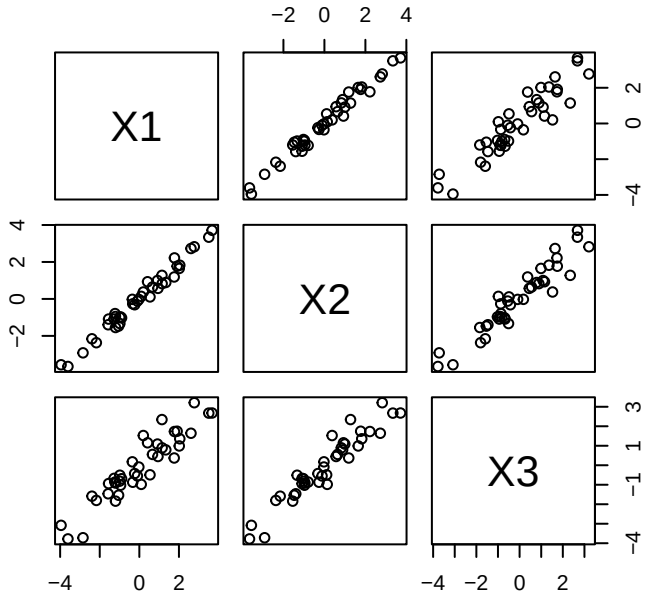
### Estadísticas descriptivas

Para determinar si este conjunto de datos se considera normal multivariado, se realizará una investigación inicial de forma descriptiva de cada variable. Se observará si presentan algún comportamiento normal bivariado y se analizarán las posibles correlaciones entre ellas.

### Grafico de dispersion de los datos

El gráfico de dispersión muestra la relación entre dos variables cuantitativas. Observando la disposición de los puntos, podemos identificar si existe una tendencia lineal o curvilínea entre las variables, o si no hay relación aparente.

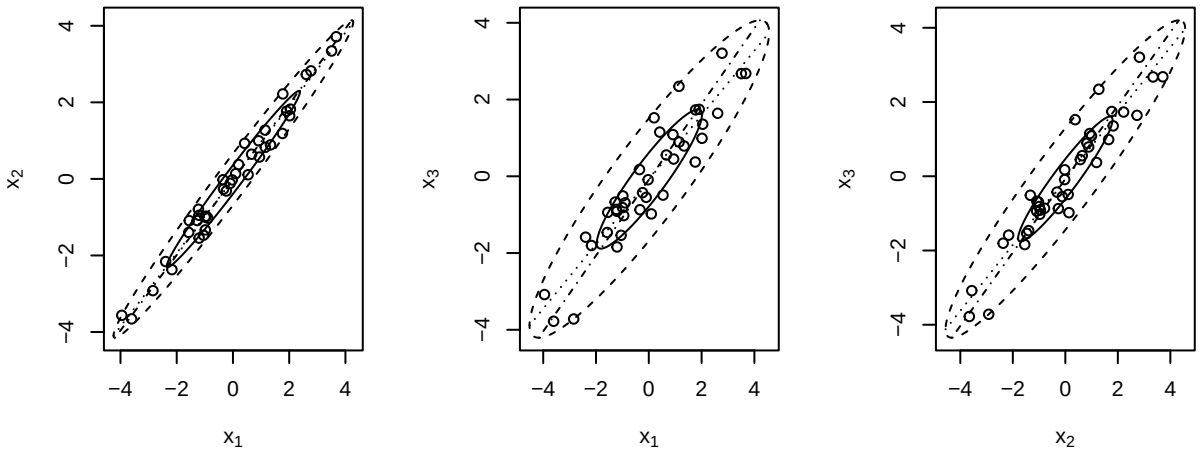
En la siguiente matriz de gráficos de dispersión, se visualizan las relaciones bivariadas entre las tres variables X1, X2 y X3. Cada gráfico individual representa un par de variables, donde los puntos muestran cómo una variable se relaciona con la otra.




Se observa una tendencia positiva en las relaciones entre las variables, indicando que a medida que una de las variables aumenta, la otra también tiende a aumentar. Esta correlación positiva es evidente en todos los pares de variables: X1 vs X2, X1 vs X3, y X2 vs X3. La alineación de los puntos en una dirección ascendente sugiere una relación lineal fuerte entre las variables.

## Prueba de normalidad multivariada informal

### Gráfico de dispersión con elipses de confianza




En los tres gráficos, las elipses de confianza abarcan la mayoría de los puntos, lo que indica que las relaciones bivariadas entre las variables (X1 vs X2, X1 vs X3, y X2 vs X3) son consistentes con la suposición de normalidad multivariada. Las elipses permiten identificar la concentración de los datos y sugiere que no hay desviaciones significativas de la normalidad en los pares de variables analizados. Esto es un indicio que existe una normal multivariada en esta muestra.

**QQplot normal multivariada**

En el gráfico Q-Q multivariado, comparamos las distancias de Mahalanobis calculadas a partir de los datos observados con los cuantiles teóricos de una distribución chi-cuadrado. Matemáticamente, la distancia de Mahalanobis para las observaciones es:

$$X : d_i^2 = (\mathbf{X} - \bar{\mathbf{X}})^{\top -1} (\mathbf{X} - \bar{\mathbf{X}})$$

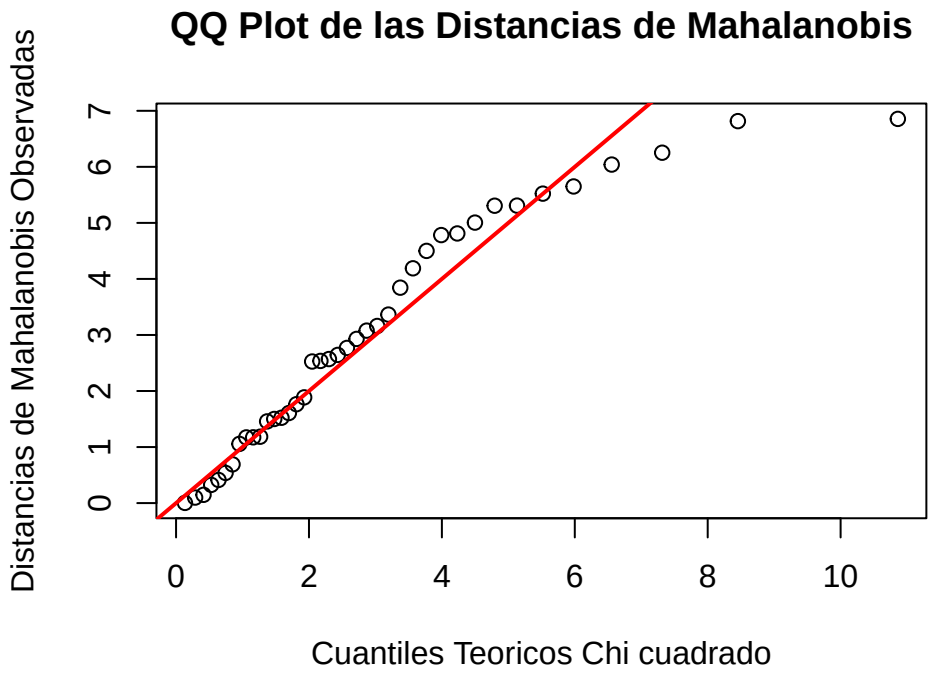
El objetivo es ver si las distancias de Mahalanobis siguen la distribución esperada bajo la normalidad multivariada. Si los datos son multivariadamente normales, las distancias deberían alinearse a lo largo de una línea recta en el gráfico Q-Q.

```
X <- as.matrix(datos)
Xbarra<-colMeans(X)
S<-cov(X)
dm <- mahalanobis(X,Xbarra,S)
cuantiles <- qchisq(ppoints(length(X)),df=4)
# Asignamos número de observaciones y dimensión
n <- length(dm)
p <- 3

# Calculamos los cuantiles teóricos de la distribución chi-cuadrado con p grados de libertad
teoricos <- qchisq(ppoints(n), df = p)

# Primero, generamos el gráfico
qqplot(teoricos, dm, main="QQ Plot de las Distancias de Mahalanobis",
       xlab="Cuantiles Teoricos Chi cuadrado",
       ylab="Distancias de Mahalanobis Observadas")

# Luego, agregamos la línea diagonal
abline(0, 1, col="red", lwd=2)
```




En el gráfico, se observa que los puntos se alinean bastante bien con la línea roja en las partes centrales, lo que sugiere que los datos en ese rango siguen aproximadamente una distribución normal multivariada. Sin embargo, en los

extremos, algunos puntos comienzan a desviarse de la línea. Estos puntos indican que las distancias de Mahalanobis en los extremos no siguen tan bien la distribución chi-cuadrado esperada. Este comportamiento podría ser una señal de la presencia de valores atípicos o de una distribución diferente a la normal en esos extremos.

## Prueba de normalidad multivariada formal

A diferencia de las pruebas informales, como los gráficos Q-Q multivariados, las pruebas formales proporcionan un criterio estadístico objetivo para determinar si se puede aceptar o rechazar la hipótesis de que los datos son multivariadamente normales.

Suponga que la muestra fue extraída de una población:

$$\text{Pob} \quad X_i = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

El interés es probar las hipótesis:

$$\mathrm{H}_0 : \mathbf{X} \sim N_3 \left( \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix} \right)$$

$$\mathrm{H}_1 : \mathbf{X}_{3\times 1} \nsim N_3 \left( \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix} \right)$$

Para resolver esta prueba de hipótesis, se usarán distintos métodos con el fin de comprobar el supuesto de normalidad multivariada, además de compararlos y observar si existe una diferencia entre ellos.

### Prueba De Shapiro

```
# ----- Prueba de Shapiro ----- #
require(mvShapiroTest)
s<-mvShapiro.Test(X)
# ----- Otras Pruebas ----- #
require(MVN)
m<-mvn(X, mvnTest="mardia") # test de Mardia
h<-mvn(X, mvnTest="hz") # test de Henze-Zirkler
r<-mvn(X, mvnTest="royston") # test de Royston
d<-mvn(X, mvnTest="dh") # test de Doornik-Hansen
```

```
#Mostramos los resultados
kable(resultados)
```

| Prueba | P_valor | Normalidad |
|---|---|---|
| Shapiro-Wilk | 0.4262779 | YES |
| Mardia | NA | YES |
| Henze-Zirkler | 0.6915232 | YES |
| Royston | 0.7599855 | YES |
| Doornik-Hansen | 0.1808371 | YES |

Al observar los test, se logra identificar que no se obtuvo suficiente evidencia para rechazar H0, por ende es una muestra con comportamiento normal multivariado, esta muestra de 3 variables y 40 unidades de investigacion, las pruebas demuestran suficiente potencia para ello.

Sin embargo, según Porras Cerron (2016), cuando se evalúa la potencia de las pruebas de normalidad multivariada, los resultados indican que no hay diferencias significativas entre las pruebas analizadas. Sin embargo, se observa que al aumentar el número de investigaciones o variables en el estudio, la potencia de algunas pruebas, como la de Mardia, tiende a disminuir ligeramente. Por otro lado, en pruebas como la de Royston, la potencia puede aumentar a medida que se incrementa el tamaño de la muestra o el número de variables. Esto sugiere que la elección de la prueba más adecuada puede depender de la estructura y tamaño de los datos utilizados en la investigación.

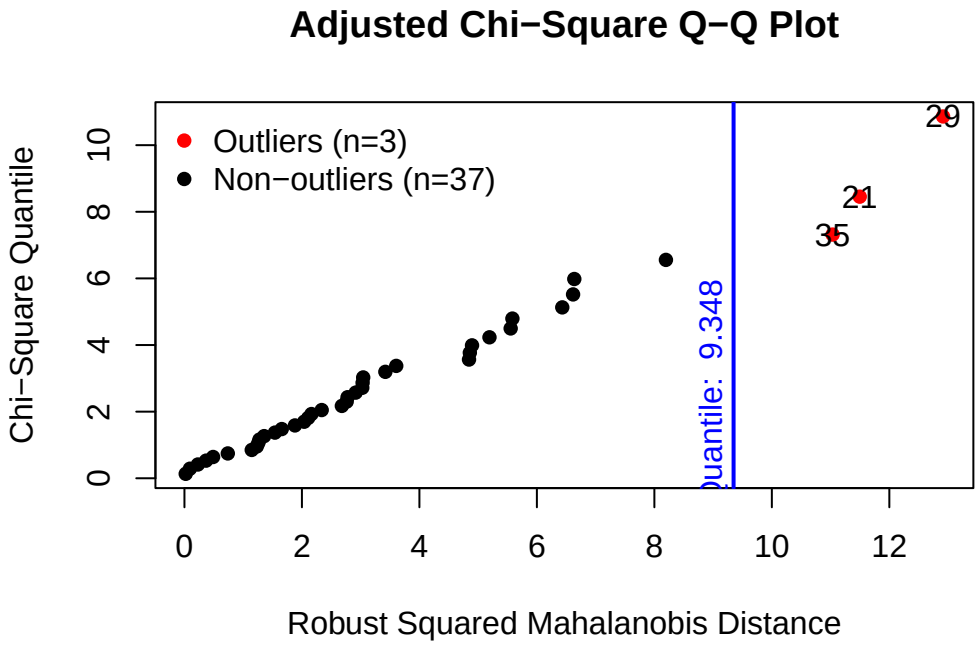
**b. Si el vector de medias poblacionales es $\mu = [0.1, -0.2, 0.05]^t$ y $S$ es la matriz de varianzas-covarianzas muestrales. ¿Cuál es la distribución aproximada de $40\left(\overline{X} - [0.1, -0.2, 0.05]^t\right)^t S^{-1}\left(\overline{X} - [0.1, -0.2, 0.05]^t\right)$?**

Para resolver este problema, primero se debe partir de la formula general de las distancias de Mahalanobis muestrales para cualquier vector p variado $X$: $d_i^2 = (\mathbf{X} - \overline{\mathbf{X}})^\top \mathbf{\Sigma}^{-1}(\mathbf{X} - \overline{\mathbf{X}})$, donde $\mathbf{X}$ es el vector de datos, $\overline{\mathbf{X}}$ es el vector de medias muestrales y $\mathbf{\Sigma}$ es la matriz de varianzas-covarianzas poblacionales.

Si consideramos la distribución del vector de medias muestrales $\overline{\mathbf{X}}$, se tiene que $\overline{\mathbf{X}} \sim N_p(\mu, \frac{\mathbf{\Sigma}}{n})$. Por lo tanto, la distribución de la distancia de Mahalanobis muestral es $d_i^2 = (\overline{\mathbf{X}} - \boldsymbol{\mu})^\top \left(\frac{\mathbf{\Sigma}}{n}\right)^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}) \sim \chi_p^2$, y considerando el teorema de limite central y que el vector de medias muestrales es un estimador insesgado de $\boldsymbol{\mu}$, se tiene que la varianza muestral es $\mathbf{S}$ tiende a la poblacional $\mathbf{\Sigma}$. Teniendo en cuenta esto, el vector de distancias de Mahalanobis muestrales para la media se podria escribir como $d_i^2 = (\overline{\mathbf{X}} - \boldsymbol{\mu})^\top (\frac{\mathbf{S}}{n})^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu})$. Si se reescribe esta formula modificando el exponente de n la formula queda de la siguiente manera: $d_i^2 = n(\overline{\mathbf{X}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu})$, la cual corresponde a la formula que buscamos estudiar y que podemos concluir que corresponde a la distancia de Mahalanobis muestral para el vector de medias poblacionales. Por lo tanto, la distribución aproximada de $40\left(\overline{\mathbf{X}} - [0.1, -0.2, 0.05]^t\right)^t \mathbf{S}^{-1}\left(\overline{\mathbf{X}} - [0.1, -0.2, 0.05]^t\right)$ es $\chi_3^2$.

**c. Usando la distancia cuadrada generalizada establezca si existen valores atípicos.**

```
X<-as.matrix(datos)
Xbarra<-colMeans(X)
S<-cov(X)
dm<-mahalanobis(X,Xbarra,S)
cuantiles<-qchisq(ppoints(length(X)),df=4)
Out<-mvn(datos, mvnTest = "mardia", multivariateOutlierMethod ="adj")
```



Se logra identificar 3 datos atípicos (21, 29 y 35), de los cuales se encuentran a la derecha del cuartil de la chi cuadrado ajustada (9.348), indicada por la línea azul. Esto significa que la distancia de Mahalanobis de estas unidades de investigación son inusualmente altas.
Al observar estos resultados, se puede deducir que la cantidad de valores atípicos en datos multivariados aumenta con el número de variables, ya que la probabilidad de que una observación se aleje significativamente de la media en alguna dimensión aumenta. Sin embargo, esta relación no es estrictamente proporcional dependerá de factores como la correlación y estructura de los datos.

#### Referencia

Porras Cerron, J. C. (2016). Comparación de pruebas de normalidad multivariada. Anales Científicos, 77(2), 141-146. https://doi.org/10.21704/ac.v77i2.483

# Punto 2

## Realizamos la carga de datos

**MultBiplotR** provee a traves de los datos *Protein* la información de datos nutricionales para habitantes de 25 paises de europa en la decadada de los 70s. Las variables presentes son:

- **Comunist:** Presencia del comunismo en el país
- **Region:** Nombre de la región que se encuentra el país
- **RedMeat:** Consumo de proteínas provenientes de carnes rojas.
- **WhiteMeat:** Consumo de proteínas provenientes de carnes blancas.
- **Eggs:** Consumo de proteínas del huevo.
- **Milk:**Consumo de proteínas de la leche.
- **Fish:** Consumo de proteínas provenientes del pescado.
- **Cereals:** Consumo de proteínas procedentes de cereales.
- **Starch:** Consumo de proteínas provenientes de carbohidratos.
- **Nuts:** Consumo de proteínas procedentes de cereales, frutos secos y semillas oleaginosas.
- **FruitVeg:** Consumo de proteínas procedentes de frutas y verduras.

Como se puede observar, nueve de las variables representan diferentes fuentes de proteína. Asumimos que todas están expresadas en **gramos por persona por día**. Aunque esto no es explícito en la documentación, se consultaron trabajos como los del programa de inteligencia de negocios la Universidad de Texas, donde se adoptó la misma suposición.

Dado lo anterior, procedemos a instalar el paquete **MultBiplotR**, renombrando y convirtiendo a DataFrame el dataset *Protein*.

```
#install.packages("MultBiplotR")
require(MultBiplotR)

# Renombramos y convertimos a Data Frame
Data <- Protein  %>% as.data.frame()

# Hacemos unas modificaciones para mostrar el nombre del país
Data$Country <- rownames(Data)
rownames(Data) <- NULL
Data <- Data[, c("Country", colnames(Data)[-ncol(Data)])]

# Mostramos los datos
kable(Data) %>% kable_styling(font_size = 6, full_width = FALSE)
```

| Country | Comunist | Region | Red_Meat | White_Meat | Eggs | Milk | Fish | Cereal | Starch | Nuts | Fruits_Vegetables |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Albania | Yes | South | 10.1 | 1.4 | 0.5 | 8.9 | 0.2 | 42.3 | 0.6 | 5.5 | 1.7 |
| Austria | No | Center | 8.9 | 14.0 | 4.3 | 19.9 | 2.1 | 28.0 | 3.6 | 1.3 | 4.3 |
| Belgium | No | Center | 13.5 | 9.3 | 4.1 | 17.5 | 4.5 | 26.6 | 5.7 | 2.1 | 4.0 |
| Bulgaria | Yes | South | 7.8 | 6.0 | 1.6 | 8.3 | 1.2 | 56.7 | 1.1 | 3.7 | 4.2 |
| Czechoslovakia | Yes | Center | 9.7 | 11.4 | 2.8 | 12.5 | 2.0 | 34.3 | 5.0 | 1.1 | 4.0 |
| Denmark | No | North | 10.6 | 10.8 | 3.7 | 25.0 | 9.9 | 21.9 | 4.8 | 0.7 | 2.4 |
| E_Germany | Yes | Center | 8.4 | 11.6 | 3.7 | 11.1 | 5.4 | 24.6 | 6.5 | 0.8 | 3.6 |
| Finland | No | North | 9.5 | 4.9 | 2.7 | 33.7 | 5.8 | 26.3 | 5.1 | 1.0 | 1.4 |
| France | No | Center | 18.0 | 9.9 | 3.3 | 19.5 | 5.7 | 28.1 | 4.8 | 2.4 | 6.5 |
| Greece | No | South | 10.2 | 3.0 | 2.8 | 17.6 | 5.9 | 41.7 | 2.2 | 7.8 | 6.5 |
| Hungary | Yes | Center | 5.3 | 12.4 | 2.9 | 9.7 | 0.3 | 40.1 | 4.0 | 5.4 | 4.2 |
| Ireland | No | Center | 13.9 | 10.0 | 4.7 | 25.8 | 2.2 | 24.0 | 6.2 | 1.6 | 2.9 |
| Italy | No | South | 9.0 | 5.1 | 2.9 | 13.7 | 3.4 | 36.8 | 2.1 | 4.3 | 6.7 |
| Holand | No | Center | 9.5 | 13.6 | 3.6 | 23.4 | 2.5 | 22.4 | 4.2 | 1.8 | 3.7 |
| Norway | No | North | 9.4 | 4.7 | 2.7 | 23.3 | 9.7 | 23.0 | 4.6 | 1.6 | 2.7 |
| Poland | Yes | Center | 6.9 | 10.2 | 2.7 | 19.3 | 3.0 | 36.1 | 5.9 | 2.0 | 6.6 |
| Portugal | No | South | 6.2 | 3.7 | 1.1 | 4.9 | 14.2 | 27.0 | 5.9 | 4.7 | 7.9 |
| Romania | Yes | South | 6.2 | 6.3 | 1.5 | 11.1 | 1.0 | 49.6 | 3.1 | 5.3 | 2.8 |
| Spain | No | South | 7.1 | 3.4 | 3.1 | 8.6 | 7.0 | 29.2 | 5.7 | 5.9 | 7.2 |
| Sweden | No | North | 9.9 | 7.8 | 3.5 | 24.7 | 7.5 | 19.5 | 3.7 | 1.4 | 2.0 |
| Switzerland | No | Center | 13.1 | 10.1 | 3.1 | 23.8 | 2.3 | 25.6 | 2.8 | 2.4 | 4.9 |
| UK | No | Center | 17.4 | 5.7 | 4.7 | 20.6 | 4.3 | 24.3 | 4.7 | 3.4 | 3.3 |
| USSR | Yes | Center | 9.3 | 4.6 | 2.1 | 16.6 | 3.0 | 43.6 | 6.4 | 3.4 | 2.9 |
| W_Germany | No | Center | 11.4 | 12.5 | 4.1 | 18.8 | 3.4 | 18.6 | 5.2 | 1.5 | 3.8 |
| Yugoslavia | Yes | South | 4.4 | 5.0 | 1.2 | 9.5 | 0.6 | 55.9 | 3.0 | 5.7 | 3.2 |

## a. Determine y analice el vector de medias y la matriz de covarianzas muestrales para las diferentes regiones.

### Vector De Medias

Dado que tenemos la muestra, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{25}$, nuestro objetivo es calcular y analizar el vector de medias $\overline{\mathbf{x}} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_9 \end{pmatrix}$, para ello calculamos las medias de cada una de las 9 variables númericas, exceptuando *Comunist*, *Country* y *Region*.

```
#Filtramos para quitar Region y Comunist
F_Data <-  Data[, !(colnames(Data) %in% c("Country", "Comunist", "Region"))]

# Calculamos y mostramos el vector de medias
kable(colMeans(F_Data), col.names = "$\\overline{X}$")
```

| | $\overline{X}$ |
|---|---|
| Red_Meat | 9.828 |
| White_Meat | 7.896 |
| Eggs | 2.936 |
| Milk | 17.112 |
| Fish | 4.284 |
| Cereal | 32.248 |
| Starch | 4.276 |
| Nuts | 3.072 |
| Fruits_Vegetables | 4.136 |

De acuerdo con los resultados, los cereales son, en promedio, la principal fuente de proteína diaria por persona, seguidos por la leche. En contraste, los huevos y las nueces proporcionan la menor cantidad de proteína. Las demás fuentes contribuyen con entre 4 y 9 gramos de proteína al día por persona.

**Matriz Covarianzas**

Para nuestro otro objetivo que es la matriz de covarianzas $\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1,9} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2,9} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \cdots & \sigma_{3,9} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{9,1} & \sigma_{9,2} & \sigma_{9,3} & \cdots & \sigma_{9,9} \end{pmatrix}$ necesitamos excluir de nuevo a las variables *Comunist*, *Country* y *Region*, y calcular la información que proporciona la matriz para las 9 variables.

```
# Calculamos y mostramos el vector de medias
kable(cov(F_Data), digits = 3) %>%  kable_styling(font_size = 8, full_width = FALSE)
```

| | Red_Meat | White_Meat | Eggs | Milk | Fish | Cereal | Starch | Nuts | Fruits_Vegetables |
|---|---|---|---|---|---|---|---|---|---|
| Red_Meat | 11.203 | 1.892 | 2.191 | 11.961 | 0.694 | -18.362 | 0.741 | -2.323 | -0.448 |
| White_Meat | 1.892 | 13.646 | 2.561 | 7.388 | -2.941 | -16.776 | 1.894 | -4.658 | -0.409 |
| Eggs | 2.191 | 2.561 | 1.249 | 4.570 | 0.249 | -8.738 | 0.826 | -1.242 | -0.092 |
| Milk | 11.961 | 7.388 | 4.570 | 50.487 | 3.334 | -46.222 | 2.582 | -8.763 | -5.234 |
| Fish | 0.694 | -2.941 | 0.249 | 3.334 | 11.577 | -19.576 | 2.245 | -0.994 | 1.634 |
| Cereal | -18.362 | -16.776 | -8.738 | -46.222 | -19.576 | 120.446 | -9.563 | 14.187 | 0.922 |
| Starch | 0.741 | 1.894 | 0.826 | 2.582 | 2.245 | -9.563 | 2.670 | -1.539 | 0.249 |
| Nuts | -2.323 | -4.658 | -1.242 | -8.763 | -0.994 | 14.187 | -1.539 | 3.943 | 1.343 |
| Fruits_Vegetables | -0.448 | -0.409 | -0.092 | -5.234 | 1.634 | 0.922 | 0.249 | 1.343 | 3.254 |

En la figura anterior se presenta la matriz de varianzas y covarianzas, donde en la diagonal se pueden identificar las varianzas de las variables correspondientes. Se observa que la variable *Cereal* tiene una varianza aproximadamente de 120, lo que, en comparación con otras variables, resulta considerablemente alta. De manera similar, la columna *Milk*, aunque no tan extrema con un valor de 50, también muestra una varianza relativamente grande, seguida de *Fish*, *Red Meat* y *White Meat*, con un valor al entre 11 y 12. Las demás variables mantienen varianzas en un rango de 1 a 3.5.

En cuanto a las covarianzas, es importante recordar que esta matriz revela la dirección de la relación entre las variables (positiva o negativa) sin indicar su magnitud absoluta. Se destacan relaciones negativas significativas entre *Cereal* y *Milk*, así como entre *Cereal* con *Red Meat* y *White Meat*. Por otro lado, las relaciones positivas más notables se observan entre *Milk* y *Red Meat*.

## b. Calcule la media de las variables por regiones ¿que puede decir al respecto?

Con el fin de calcular las medias por región realizamos un filtro que permita agrupar los países por regiones, y luego procedemos a realizar el cálculo de la media de todas las variables, lo que resulta en la siguiente matriz:

```
# Agrupamos los datos por región
G_Data <- Data %>% group_by(Region)

# Realizamos el calculo de las medias a lo largo de las variables números
MediaxRegion <- G_Data %>% summarize(
    across(Red_Meat:Fruits_Vegetables, mean, na.rm = TRUE)
  )

#Mostramos las medias por region
kable(MediaxRegion, digits = 3, caption = "Vectores De Medias Por Región") %>%
  kable_styling(font_size = 10, full_width = FALSE)
```

Table 6: Vectores De Medias Por Región

| Region | Red_Meat | White_Meat | Eggs | Milk | Fish | Cereal | Starch | Nuts | Fruits_Vegetables |
|---|---|---|---|---|---|---|---|---|---|
| North | 9.850 | 7.050 | 3.150 | 26.675 | 8.225 | 22.675 | 4.550 | 1.175 | 2.125 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Center | 11.177 | 10.408 | 3.546 | 18.346 | 3.131 | 28.946 | 5.000 | 2.246 | 4.208 |
| South | 7.625 | 4.237 | 1.837 | 10.325 | 4.188 | 42.400 | 2.963 | 5.362 | 5.025 |

En la tabla se observa que, en la región del norte, la proteína más consumida es la leche, mientras que en el centro y el sur predomina el cereal. Las nueces son en promedio las menos consumidas en el norte y el centro, mientras que, en el sur, los huevos registran el menor consumo.

Por otro lado, no se identifica un patrón de consumo promedio similar para ninguna proteína específica entre las distintas regiones. Sin embargo, en la zona sur, el consumo promedio de carnes blancas y pescados es similar, al igual que el de nueces y frutas y vegetales. De manera similar ocurre en la región central respecto a el consumo promedio de huevos y pescado.

### c. Intente construir grupos de paises usando representaciones pictóricas (gráficos de estrellas o caras de Chernoff).

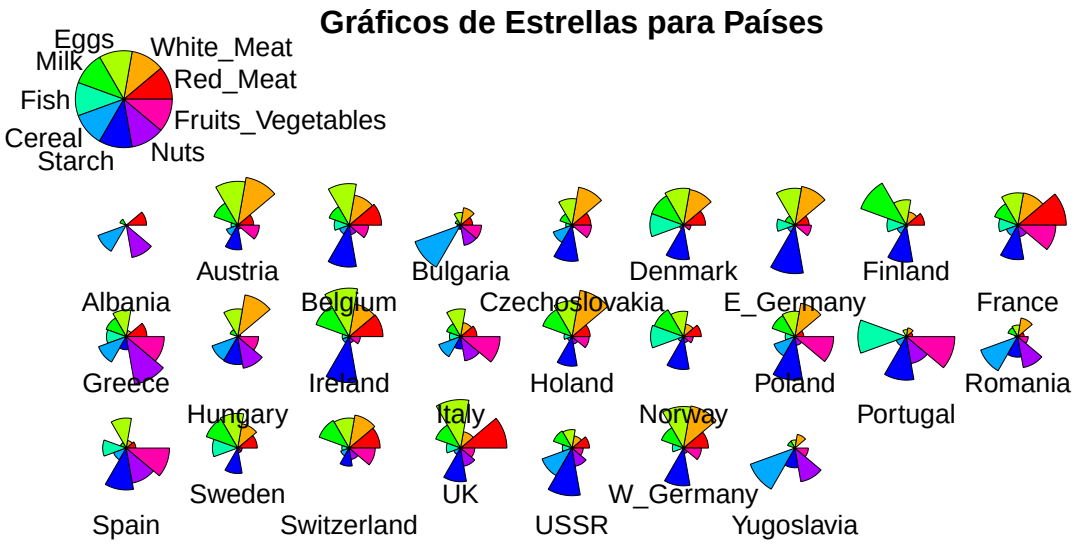
Con el fin realizar los posibles grupos de países vamos a observar cómo se materializas nuestro gráfico de estrellas ya que es un gráfico más reconocible para analizar que países pueden ser agrupados mediante una puntuación alta de las variables compartidas:

```
# Hacemos la función de escalado de Min-Max
min_max_norm <- function(x) {
  (x - min(x)) / (max(x) - min(x))}

# Escalamos las variables usando Min-Max
F_Data_std <- F_Data %>% mutate(across(everything(), min_max_norm)) %>% as.data.frame()

# Añadimos la columna para indentificar los países
F_Data_std$Country <- Data$Country

par(mar = c(10, 4, 4, 2))
# Creamos y mostramos el gráfico de estrellas con ajustes
stars(F_Data_std[, -10], labels = F_Data_std$Country,key.loc = c(2.25, 9.5),
main = "Gráficos de Estrellas para Países", draw.segments = TRUE,
col.segments = rainbow(ncol(F_Data_std) - 1), cex.main = 1.0,
cex.lab = 0.5, lty = 1.5, ncol = 9)
```



**Analisis Grupos**


Figure 1: Grupo Con Alto Consumo de Huevos, Leche y Carnohidratos

En el primer grupo encontramos 3 países pertenecientes al norte de Europa, donde debido a las bajas temperaturas la siembra de varios cultivos es muy complicada así que su consumo de proteína se evidencia en otros productos, relacionados al mundo de la ganadería y proteínas de origen animal, siendo la leche, los huevos y otros carbohidratos los principales, aunque también destacan las carnes rojas, blancas y el pescado.

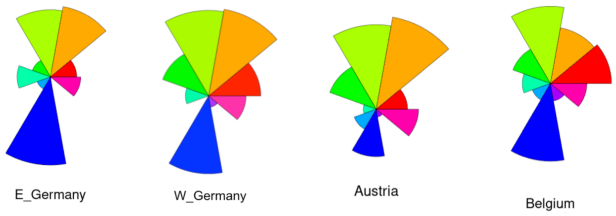


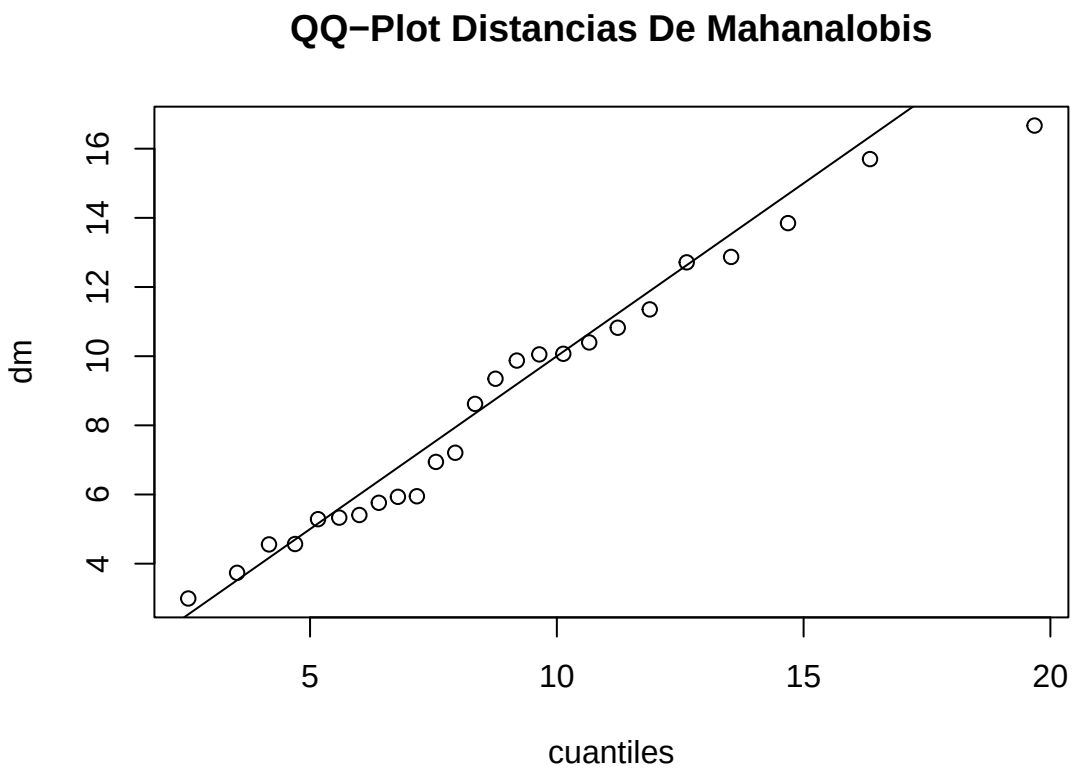

Figure 2: Grupo Con Alto Consumo de Huevos, Carne Blanca y Carnohidratos

El consumo de proteína de este grupo de países del centro de Europa también proviene principalmente de origen animal, sin embargo, a diferencia del primer grupo el consumo de pescado y leche disminuye bastante, pero se ve que hay un mayor consumo de carnes rojas y blancas (principalmente las blancas) aunque si se mantiene el alto consumo de huevos y carbohidratos.


Figure 3: Grupo Con Alto de Cereales y Frutos Secos

Este último grupo situado al este de Europa es el más distinto de los 3 porque, excepto el pescado, el consumo de proteínas animal disminuye mucho (el de carbohidratos también disminuyó) y los frutos secos se alzan como una de las principales fuentes de proteínas, junto con el pescado, para esta región.

### d. Utilice las herramientas de gráficas más adecuadas para verificar normalidad multivariada.




La gráfica sugiere que los datos transformados (escalados) están alineados en su mayoría con una distribución chi-cuadrado, lo que sugiere que podrían provenir de una distribución normal multivariada. No obstante, las desviaciones en los extremos podrían indicar la presencia de valores atípicos, lo que justifica una investigación más detallada (como una prueba formal) para confirmar esta hipótesis.

### e. Realice la prueba de Mardia para verificar las hipótesis:

**H0 : Los datos provienen de una poblacion Normal Multivariada**

**H1 : Los datos NO provienen de una poblacion Normal Multivariada**

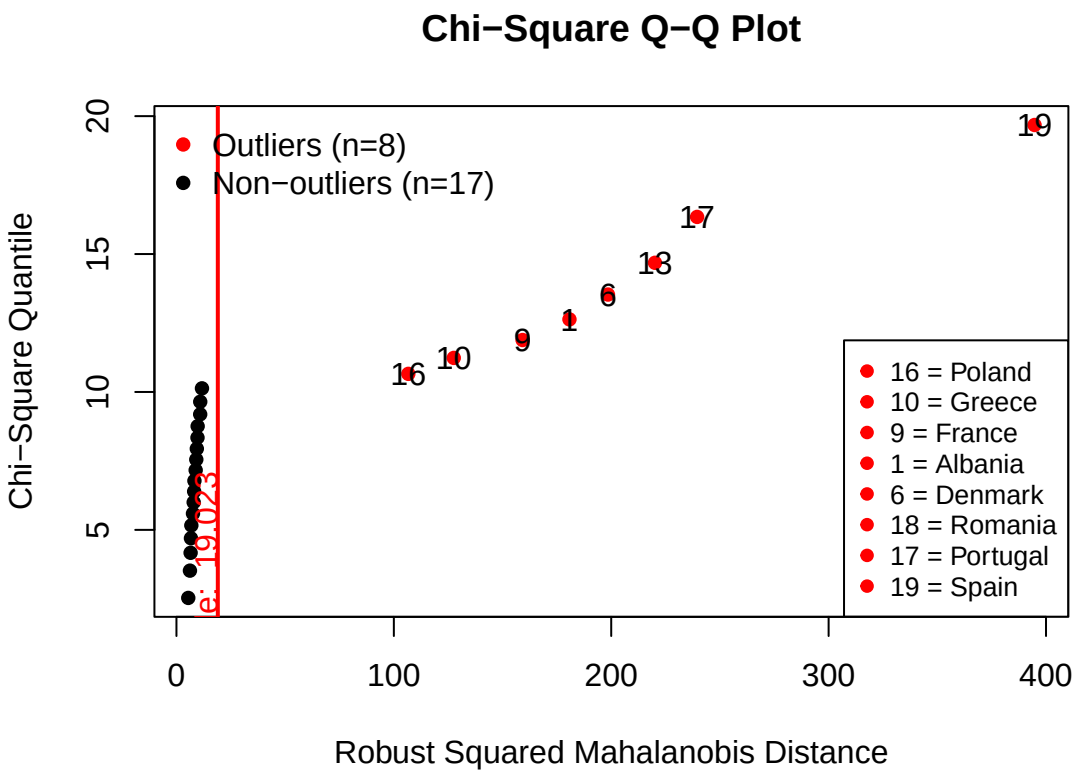
| Test | Variable | p-value | Result/Normality |
|---|---|---|---|
| Mardia Skewness | NA | 0.419 | YES |
| Mardia Kurtosis | NA | 0.601 | YES |
| MVN | NA | NA | YES |
| Anderson-Darling | Red_Meat | 0.074 | YES |
| Anderson-Darling | White_Meat | 0.093 | YES |
| Anderson-Darling | Eggs | 0.351 | YES |
| Anderson-Darling | Milk | 0.344 | YES |

| | | | |
|---|---|---|---|
| Anderson-Darling | Fish | 0.051 | YES |
| Anderson-Darling | Cereal | 0.01 | NO |
| Anderson-Darling | Starch | 0.292 | YES |
| Anderson-Darling | Nuts | 0.016 | NO |
| Anderson-Darling | Fruits_Vegetables | 0.038 | NO |

Podemos observar que la prueba de Mardia no rechaza la hipótesis nula (H0), lo que indica que los datos podrían provenir de una población normal multivariada, con un p-valor aproximado de 0.42 para la asimetría y 0.60 para la curtosis.

En las pruebas de normalidad univariadas, notamos que algunas variables no cumplen con el supuesto de normalidad. Específicamente, las variables *Cereal*, *Nuts*, y *Fruits_Vegetables* presentan p-valores de 0.0101, 0.0155, y 0.0380, respectivamente, lo que sugiere que estas distribuciones se desvían significativamente de la normalidad. No obstante, los p-valores no son extremadamente bajos, lo que podría explicar por qué la normalidad multivariada conjunta no fue rechazada.

### f. Verifique si hay outliers (multivariados) e identifíquelos.




Como podemos observar en el gráfico de distancia de mahalanobis robusta, vemos que son un total de 8 países que cuentan como outliers multivariados, entre ellos vemos un grupo conformado por 7 países que dentro de todo si se alejan en gran medida de la línea trazada en (19.023) unidades de distancia de mahalanobis robusta, entre ellos se encuentran países como **Polonia** que abre el grupo, **Albania** que es la mitad de este conjunto de países y **Portugal** que cierra este colectivo de puntos.

Por último, no podemos dejar de ver a **España** que es el país más alejado de todos, con una diferencia notable en cuanto a los no outliers, y más aún frente al monto de los otros 7 outliers, llegando a estar casi 150 unidades más lejos en cuanto a la distancia de mahalanobis robusta.

## g. Pruebe las hipótesis

Dado el vecto de medias $\mu_0 = [9, \quad 7, \quad 2, \quad 15, \quad 5, \quad 30, \quad 4, \quad 3, \quad 4]^\top$, se tiene el interés de comprabar los casos **Univariado** y **Multivariado** para saber si dicho vector es plausible para $\boldsymbol{\mu}$.

### i. De forma Univariada

Se tienen las siguientes hipotesis para las 9 variables de los datos:

*Variables* RedMeat$(\mu_1)$, WhiteMeat$(\mu_2)$, Eggs$(\mu_3)$, Milk$(\mu_4)$ *y* Fish$(\mu_5)$

$$\begin{array}{lllll} H_0 : \mu_1 = 9 & H_0 : \mu_2 = 7 & H_0 : \mu_3 = 2 & H_0 : \mu_4 = 15 & H_0 : \mu_5 = 5 \\ H_1 : \mu_1 \neq 9 & H_1 : \mu_2 \neq 7 & H_1 : \mu_3 \neq 2 & H_1 : \mu_4 \neq 15 & H_1 : \mu_5 \neq 5 \end{array}$$

*Variables* Cereals$(\mu_6)$, Starch$(\mu_7)$, Nuts$(\mu_8)$, FruitVeg$(\mu_9)$

$$\begin{array}{llll} H_0 : \mu_6 = 30 & H_0 : \mu_7 = 4 & H_0 : \mu_8 = 3 & H_0 : \mu_9 = 4 \\ H_1 : \mu_6 \neq 30 & H_1 : \mu_7 \neq 4 & H_1 : \mu_8 \neq 3 & H_1 : \mu_9 \neq 4 \end{array}$$

```r
# Obtenemos los nombres de las columnas
variables <- colnames(F_Data)

# Creamos los valores de mu para las prueba t
mu_value <- c(9, 7, 2, 15, 5, 30, 4, 3, 4)

# Creamos un DataFrame vacío para almacenar los resultados
results <- data.frame(Variable = character(), P_Value = numeric(), stringsAsFactors = FALSE)

# Recorremos las columnas y realizamos el t.test para cada una
for (i in 1:length(variables)) {
  variable_name <- variables[i]
  test_result <- t.test(Data[[variable_name]], mu = mu_value[i], conf.level = 0.95)

  # Añadimos los resultados al DataFrame
  results <- rbind(results, data.frame(Variable = variable_name, P_Value = test_result$p.value))
}

# Mostramos los resultados de los p-valores
kable(results)
```

| Variable | P_Value |
|---|---|
| Red_Meat | 0.2280914 |
| White_Meat | 0.2370278 |
| Eggs | 0.0003278 |
| Milk | 0.1502482 |
| Fish | 0.3032063 |
| Cereal | 0.3159699 |
| Starch | 0.4067248 |
| Nuts | 0.8576562 |
| Fruits_Vegetables | 0.7095156 |

Después de realizar pruebas t univariadas para cada variable, utilizando valores específicos de $\mu$ como referencia para verificar las sospechas sobre la cantidad de gramos promedio de una determinada proteína consumida por persona al día en 25 países, se interpretaron los p-valores obtenidos bajo niveles de significancia de 0.01, 0.05 y 0.10.

A un nivel de significancia de 0.01, no se rechaza la hipótesis nula para ninguna de las variables, lo que sugiere que las medias no difieren significativamente de los valores propuestos para .

Al aplicar un nivel de significancia de 0.05, solo la variable "Eggs" mostró un p-valor menor a 0.05, lo que permite rechazar la hipótesis nula e indica que su media difiere significativamente del valor .

Al aumentar el nivel de significancia a 0.10, el resultado permanece constante, y "Eggs" sigue siendo la única variable para la cual se puede rechazar la hipótesis nula. *Sin embargo, es importante señalar que podrían existir otros valores de $\mu$ que también sean consistentes con los datos, lo que introduce cierta incertidumbre en la interpretación de los resultados por lo que debemos optar por realizar una prueba para el caso multivariado.*

### ii. De forma Multivariada

Dado que nuestro objetivo es contrastar si (en conjunto) el vector $\boldsymbol{\mu_0}$ dado es un valor plausible para $\boldsymbol{\mu}$ buscamos probar la siguiente hipotesis:

$$\mathrm{H}_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0}$$
$$\mathrm{H}_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu_0}$$

Usaremos la **Estadística de $T^2$ de Hoteling** que es una generalización de la prueba t univariada, con lo cual obtuvimos el siguiente resultado:

| Valor.Estadístico | Valor.Crítico |
|---:|---:|
| 64.84185 | 34.2585 |

Como el valor estadístico (64.84185) es superior al valor crítico (34.2585), se rechaza la hipótesis nula $H_0$. Esto sugiere que hay suficiente evidencia para concluir que el promedio de gramos de distintos proteína consumidos no coincide con los valores propuestos en $\boldsymbol{\mu}_0$.