Actividad Integradora 1 - Variable Agua

Nallely Serna

2024-08-20

##AGUA

```
# Cargar el dataset
data <- read.csv("food_data_g.csv")

# Imprimir los primeros datos
head(data)
```

```
##   X Unnamed..0                              food Caloric.Value  Fat
## 1 0          0                     cream cheese            51  5.0
## 2 1          1                 neufchatel cheese           215 19.4
## 3 2          2 requeijao cremoso light catupiry            49  3.6
## 4 3          3                   ricotta cheese            30  2.0
## 5 4          4              cream cheese low fat            30  2.3
## 6 5          5             cream cheese fat free            19  0.2
##   Saturated.Fats Monounsaturated.Fats Polyunsaturated.Fats
Carbohydrates Sugars
## 1            2.9                1.300                0.200
0.8  0.500
## 2           10.9                4.900                0.800
3.1  2.700
## 3            2.3                0.900                0.000
0.9  3.400
## 4            1.3                0.500                0.002
1.5  0.091
## 5            1.4                0.600                0.042
1.2  0.900
## 6            0.1                0.091                0.075
1.4  1.000
##   Protein Dietary.Fiber Cholesterol Sodium Water Vitamin.A Vitamin.B1
## 1     0.9           0.0        14.6  0.016   7.6     0.200      0.033
## 2     7.8           0.0        62.9  0.300  53.6     0.200      0.099
## 3     0.8           0.1         0.0  0.000   0.0     0.000      0.000
## 4     1.5           0.0         9.8  0.017  14.7     0.075      0.019
## 5     1.2           0.0         8.1  0.046  10.0     0.016      0.080
## 6     2.8           0.0         2.2  0.100  12.9     0.063      0.020
##   Vitamin.B11 Vitamin.B12 Vitamin.B2 Vitamin.B3 Vitamin.B5 Vitamin.B6
Vitamin.C
## 1       0.064       0.092      0.097      0.084      0.052      0.096
0.004
## 2       0.079       0.090      0.100      0.200      0.500      0.078
0.000
```

```
## 3       0.000       0.000       0.000       0.000       0.000       0.000
0.000
## 4       0.079       0.091       0.027       0.041       0.016       0.007
0.006
## 5       0.062       0.049       0.026       0.080       0.100       0.003
0.000
## 6       0.089       0.092       0.021       0.025       0.200       0.038
0.000
##   Vitamin.D Vitamin.E Vitamin.K Calcium Copper  Iron Magnesium
Manganese
## 1     0.000     0.000     0.100   0.008 14.100 0.082     0.027
1.300
## 2     0.000     0.300     0.045  99.500  0.034 0.100     8.500
0.088
## 3     0.000     0.000     0.000   0.000  0.000 0.000     0.000
0.000
## 4     0.000     0.001     0.011   0.097 41.200 0.097     0.096
4.000
## 5     0.036     0.009     0.019  22.200  0.072 0.008     1.200
0.098
## 6     0.000     0.049     0.059  63.200  0.039 0.053     4.000
0.028
##   Phosphorus Potassium Selenium  Zinc Nutrition.Density
## 1      0.091      15.5   19.100 0.039             7.070
## 2    117.300     129.2    0.054 0.700           130.100
## 3      0.000       0.0    0.000 0.000             5.400
## 4      0.024      30.8   43.800 0.035             5.196
## 5     22.800      37.1    0.034 0.053            27.007
## 6     94.100      50.0    0.013 0.300            67.679
```

```
# Seleccionar la variable "Agua"
agua <- data$Water

# Ver las primeras observaciones de la variable
head(agua)
```

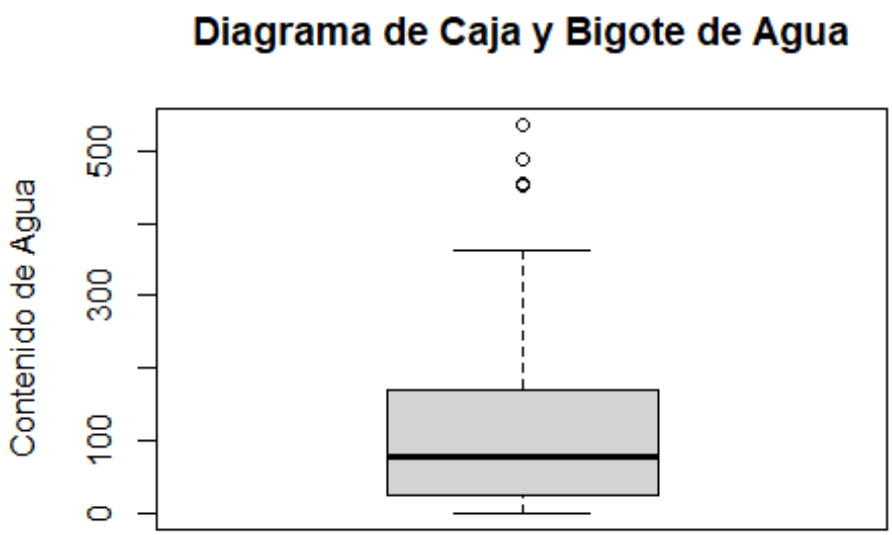
```
## [1]  7.6 53.6  0.0 14.7 10.0 12.9
```

```
cat(head(agua), sep = " ")
```

```
## 7.6 53.6 0 14.7 10 12.9
```

##1. Análisis de Datos Atípicos

```
# Crear un boxplot para la variable "Agua"
boxplot(agua, main="Diagrama de Caja y Bigote de Agua", ylab="Contenido
de Agua")
```




```r
# Calcular medidas descriptivas
summary(agua)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    25.9    76.7   101.7   169.1   535.8
```

```r
IQR_agua <- IQR(agua) # Rango intercuartílico
sd_agua <- sd(agua)    # Desviación estándar

# Imprimir resultados
cat("Rango Intercuartílico: ", IQR_agua, "\n")
```

```
## Rango Intercuartílico:  143.15
```

```r
cat("Desviación Estándar: ", sd_agua, "\n")
```

```
## Desviación Estándar:  88.50171
```

```r
# Calcular los límites para datos atípicos con 1.5 IQR
Q1 <- quantile(agua, 0.25)
Q3 <- quantile(agua, 0.75)
lim_inf <- Q1 - 1.5 * IQR_agua
lim_sup <- Q3 + 1.5 * IQR_agua

# Identificar datos atípicos
outliers_1.5 <- agua[agua < lim_inf | agua > lim_sup]
cat("Número de datos atípicos (1.5 IQR): ", length(outliers_1.5), "\n")
```

```
## Número de datos atípicos (1.5 IQR):  4
```

```
# Calcular los límites para datos atípicos con 3 desviaciones estándar
media_agua <- mean(agua)
lim_inf_sd <- media_agua - 3 * sd_agua
lim_sup_sd <- media_agua + 3 * sd_agua

# Identificar datos atípicos
outliers_3sd <- agua[agua < lim_inf_sd | agua > lim_sup_sd]
cat("Número de datos atípicos (3 Desviaciones Estándar): ",
length(outliers_3sd), "\n")
```

```
## Número de datos atípicos (3 Desviaciones Estándar):  4
```

```
# Calcular los límites para datos extremos con 3 IQR
lim_inf_ext <- Q1 - 3 * IQR_agua
lim_sup_ext <- Q3 + 3 * IQR_agua

# Identificar datos extremos
extremos_3IQR <- agua[agua < lim_inf_ext | agua > lim_sup_ext]
cat("Número de datos extremos (3 IQR): ", length(extremos_3IQR), "\n")
```

```
## Número de datos extremos (3 IQR):  0
```

##Interpretación:

Los datos atípicos identificados en ambos métodos (1.5 IQR y 3 desviaciones estándar) sugieren que hay algunas observaciones que se desvían significativamente de la mayoría de los valores en la variable Agua. Aunque no se encontraron datos extremos (según el criterio de 3 IQR).

##2. Análisis de Normalidad

```
# Prueba de Anderson-Darling
ad_test <- ad.test(agua)
cat("Anderson-Darling p-valor: ", ad_test$p.value, "\n")
```

```
## Anderson-Darling p-valor:  3.7e-24
```

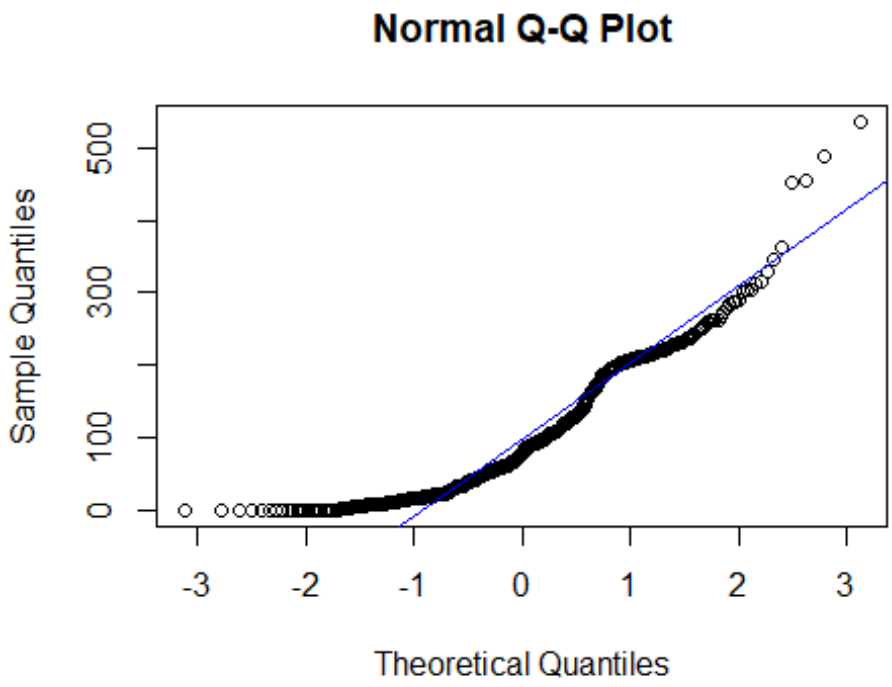
```
# Prueba de Jarque-Bera
jb_test <- jarque.bera.test(agua)
cat("Jarque-Bera p-valor: ", jb_test$p.value, "\n")
```

```
## Jarque-Bera p-valor:  0
```

##Hipótesis:

H0: Los datos siguen una distribución normal. H1: Los datos no siguen una distribución normal. Ambos p-valores indican que se rechaza la hipótesis nula, por lo que los datos no siguen una distribución normal.

```r
# Gráfico Q-Q plot
qqnorm(agua)
qqline(agua, col = "blue")
```



##Gráfico QQ Plot:

El gráfico QQ indica que los datos de Agua no siguen la línea diagonal, lo cual refuerza la conclusión de que los datos no son normalmente distribuidos.

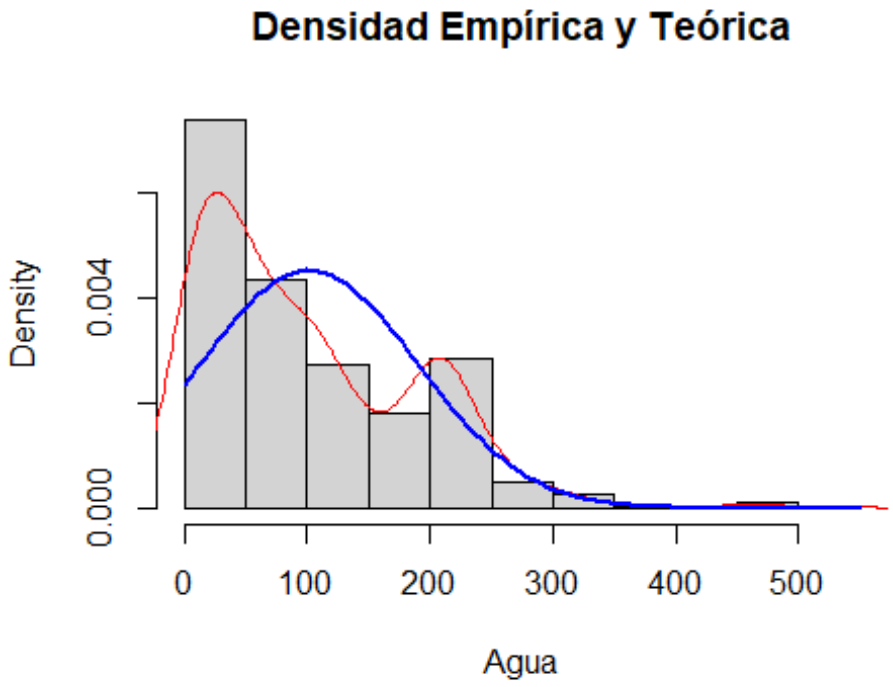
```r
# Calcular sesgo y curtosis
library(e1071)
sesgo <- skewness(agua)
curtosis <- kurtosis(agua)

cat("Coeficiente de Sesgo: ", sesgo, "\n")

## Coeficiente de Sesgo:  1.080845

cat("Coeficiente de Curtosis: ", curtosis, "\n")

## Coeficiente de Curtosis:  1.395062
```

##Sesgo y curtosis

Los valores indican que la distribución está sesgada a la derecha y tiene colas más pesadas que una distribución normal

```r
# Calcular medidas de tendencia central
mediana_agua <- median(agua)
```

```r
rango_medio <- (Q1 + Q3) / 2

cat("Media: ", media_agua, "\n")

## Media:  101.6587

cat("Mediana: ", mediana_agua, "\n")

## Mediana:  76.7

cat("Rango Medio: ", rango_medio, "\n")

## Rango Medio:  97.475

# Gráfico de densidad empírica y teórica
hist(agua, freq=FALSE, main="Densidad Empírica y Teórica", xlab="Agua")
lines(density(agua), col="red")
curve(dnorm(x, mean=mean(agua), sd=sd(agua)), add=TRUE, col="blue",
lwd=2)
```

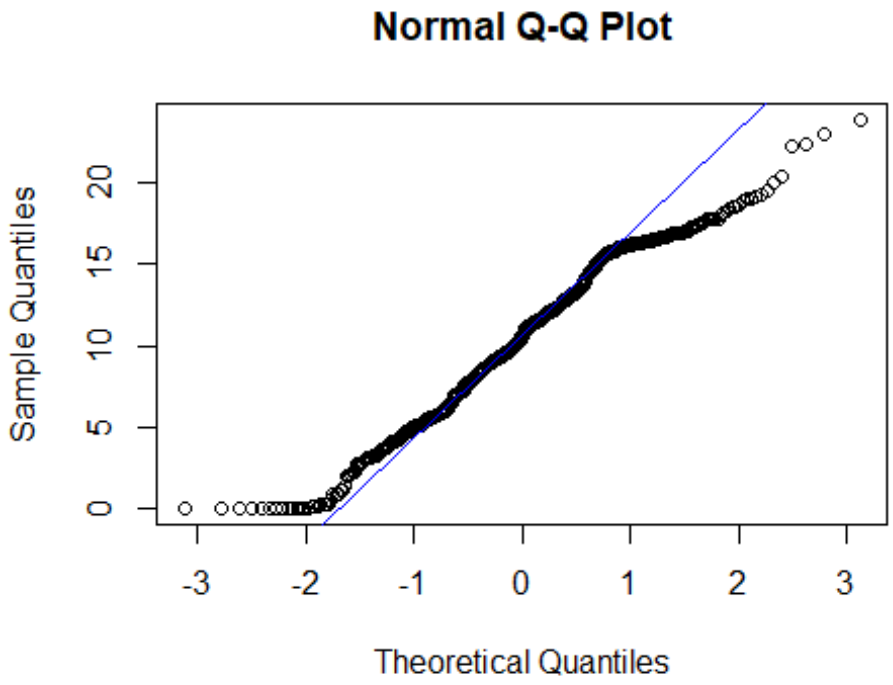


##Conclusion

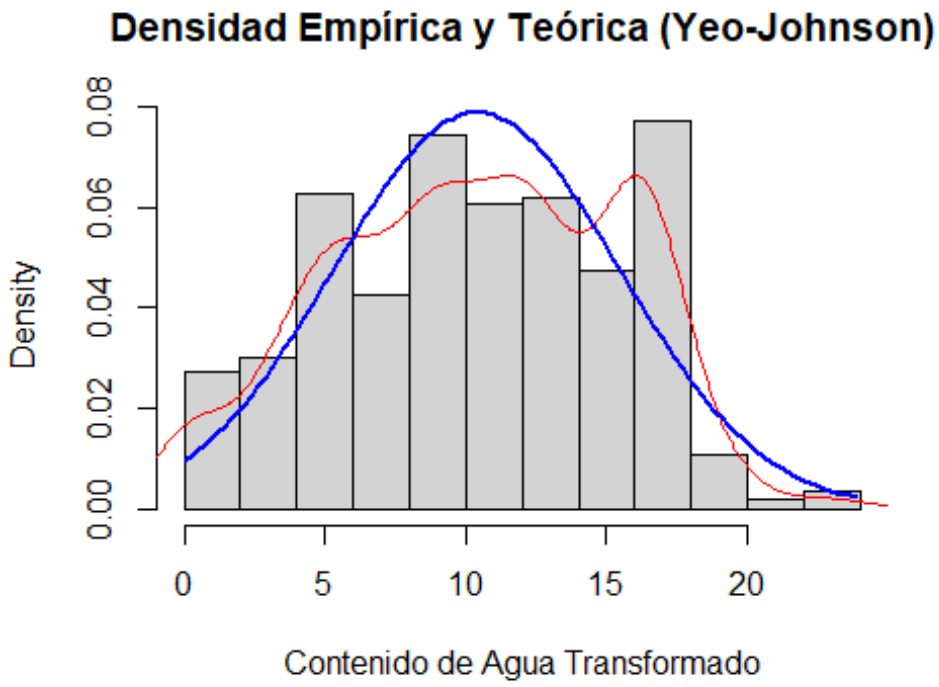
Los datos muestran un claro alejamiento de la normalidad, confirmado por las pruebas estadísticas, el gráfico QQ, y las medidas de sesgo y curtosis. Los datos atípicos pueden estar afectando la distribución, lo cual es evidente en los gráficos y pruebas.

##Transformación a Normalidad

```r
library(car)
yeo_johnson_agua <- powerTransform(agua, family = "yjPower")
lambda_yj <- yeo_johnson_agua$lambda
agua_transformada_yj <- yjPower(agua, lambda_yj)

# Analizar la normalidad de la variable transformada
# Pruebas de normalidad
ad_test_yj <- ad.test(agua_transformada_yj)
jarque_bera_test_yj <- jarque.bera.test(agua_transformada_yj)
##Graficos QQ
qqnorm(agua_transformada_yj)
qqline(agua_transformada_yj, col = "blue")
```




```r
# Ver sesgo y curtosis
skewness_yj <- mean((agua_transformada_yj -
mean(agua_transformada_yj))^3) / sd(agua_transformada_yj)^3
kurtosis_yj <- mean((agua_transformada_yj -
mean(agua_transformada_yj))^4) / sd(agua_transformada_yj)^4 - 3

# Gráfico de densidad empírica y teórica
hist(agua_transformada_yj, freq=FALSE, main="Densidad Empírica y Teórica
(Yeo-Johnson)", xlab="Contenido de Agua Transformado")
lines(density(agua_transformada_yj), col="red")
curve(dnorm(x, mean=mean(agua_transformada_yj),
sd=sd(agua_transformada_yj)), from=min(agua_transformada_yj),
to=max(agua_transformada_yj), add=TRUE, col="blue", lwd=2)
```




```
# Medidas descriptivas de los datos originales
summary(agua)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    25.9    76.7   101.7   169.1   535.8
```

```
skewness_original <- mean((agua - mean(agua))^3) / sd(agua)^3
kurtosis_original <- mean((agua - mean(agua))^4) / sd(agua)^4 - 3

# Medidas descriptivas de los datos transformados
summary(agua_transformada_yj)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.305  10.525  10.392  14.854  23.883
```

```
skewness_yj
```

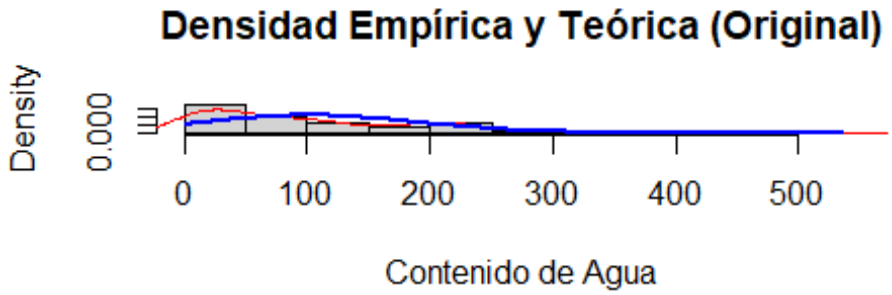
```
## [1] -0.1415837
```

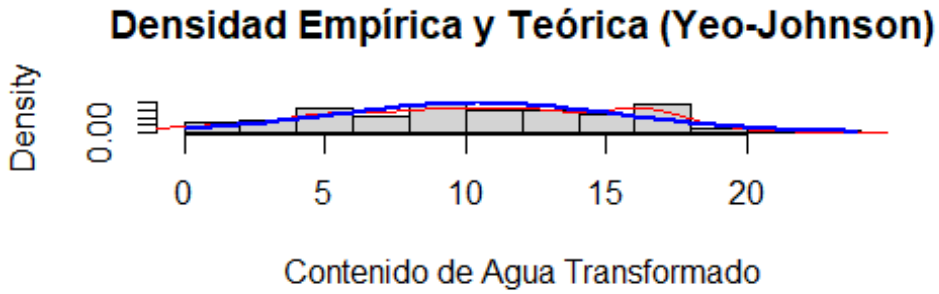
```
kurtosis_yj
```

```
## [1] -0.7614622
```

```
par(mfrow=c(2,1))  # Comparar en un mismo gráfico

# Datos originales
hist(agua, freq=FALSE, main="Densidad Empírica y Teórica (Original)",
xlab="Contenido de Agua")
```

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

# Datos transformados
hist(agua_transformada_yj, freq=FALSE, main="Densidad Empírica y Teórica
(Yeo-Johnson)", xlab="Contenido de Agua Transformado")
lines(density(agua_transformada_yj), col="red")
curve(dnorm(x, mean=mean(agua_transformada_yj),
sd=sd(agua_transformada_yj)), from=min(agua_transformada_yj),
to=max(agua_transformada_yj), add=TRUE, col="blue", lwd=2)
```






```
par(mfrow=c(1,1))  # Restablecer
```

##Conclusion

La transformación Yeo-Johnson ha mejorado significativamente la normalidad de la variable Agua. Los datos transformados presentan una distribución que se ajusta mejor a la normalidad en comparación con los datos originales, como lo evidencian las pruebas de normalidad, el gráfico QQ, y las medidas de sesgo y curtosis.

##Recomendación: La transformación Yeo-Johnson parece ser adecuada para normalizar los datos de Agua. La reducción en el sesgo y la curtosis, junto con una mejor alineación con la normalidad en los gráficos y las pruebas estadísticas, sugieren que esta transformación es efectiva para lograr una distribución más normal de los datos. Esta transformación permite un análisis más robusto y fiable en estudios futuros.