Base de Datos Plasma

Shamuel Manrrique NIP:802400

10/20/2020

Ejercicio propuesto 1

En el fichero de datos Base Datos Plasma Retinol.tex se tienen las variables cuantitativas CHOLESTEROL, BETADIET, RETDIET, BETAPLASMA y RETPLASMA. En este fichero de texto se comentan esas variables. Además, se dispone de la variable cualitativa (factor) GÉNERO que agrupa los datos en dos niveles: Hombre y Mujer.

Import libraries

```
library("RcmdrMisc")
library("ggplot2")
library("dplyr")          # load
```

Carga de dataset

Los datos que se encuentran en el archivo BaseDatosPlasmaCompleta.RData

```
load("C:/Users/smmanrrique/3D Objects/unizar/add/aad_practices/practice1/datasets/BaseDatosPlasmaComple
```

Resumen general de la información del dataset

```
summary(Dataset)
```

```
##       Edad     Género  Fumador    Quetelet     Vitamina    Calorias
##  41     : 13   1: 42   1:157   Min.   :16.33   1:122    Min.   : 445.2
##  46     : 12   2:273   2:115   1st Qu.:21.80   2: 82    1st Qu.:1338.0
##  49     : 12           3: 43   Median :24.74   3:111    Median :1666.8
##  37     : 11                   Mean   :26.16            Mean   :1796.7
##  44     : 11                   3rd Qu.:28.85            3rd Qu.:2100.4
##  43     : 10                   Max.   :50.40            Max.   :6662.2
##  (Other):246
##      Grasa            Fibra          Alcohol         Cholesterol
##  Min.   : 14.40   Min.   : 3.10   Min.   :  0.000   Min.   :  37.7
##  1st Qu.: 53.95   1st Qu.: 9.15   1st Qu.:  0.000   1st Qu.: 156.8
##  Median : 72.90   Median :12.10   Median :  0.500   Median : 211.7
##  Mean   : 77.03   Mean   :12.79   Mean   :  6.503   Mean   : 255.7
##  3rd Qu.: 95.25   3rd Qu.:15.60   3rd Qu.:  3.850   3rd Qu.: 325.9
##  Max.   :235.90   Max.   :36.80   Max.   :328.000   Max.   :1153.0
##
```

```
##     Betadiet       Retdiet         Betaplasma       Retplasma
##  Min.   : 141   Min.   :  27.0   Min.   :   0.0   Min.   : 179.0
##  1st Qu.:1106   1st Qu.: 465.5   1st Qu.:  93.0   1st Qu.: 466.0
##  Median :1802   Median : 706.0   Median : 143.0   Median : 564.0
##  Mean   :2180   Mean   : 823.7   Mean   : 199.4   Mean   : 603.2
##  3rd Qu.:2836   3rd Qu.:1026.5   3rd Qu.: 236.5   3rd Qu.: 721.0
##  Max.   :9642   Max.   :6901.0   Max.   :1415.0   Max.   :1727.0
##                                                   NA's   :6
```

Estadística descriptiva

Para la variable cuantitativa CHOLESTEROL se considera, en primer lugar, los datos de forma conjunta(sin distinguir por género) y, en segundo lugar, agrupando por la variable GÉNERO.

1.Datos conjuntos

Analizamos los datos tomando en cuenta el factor colesterol

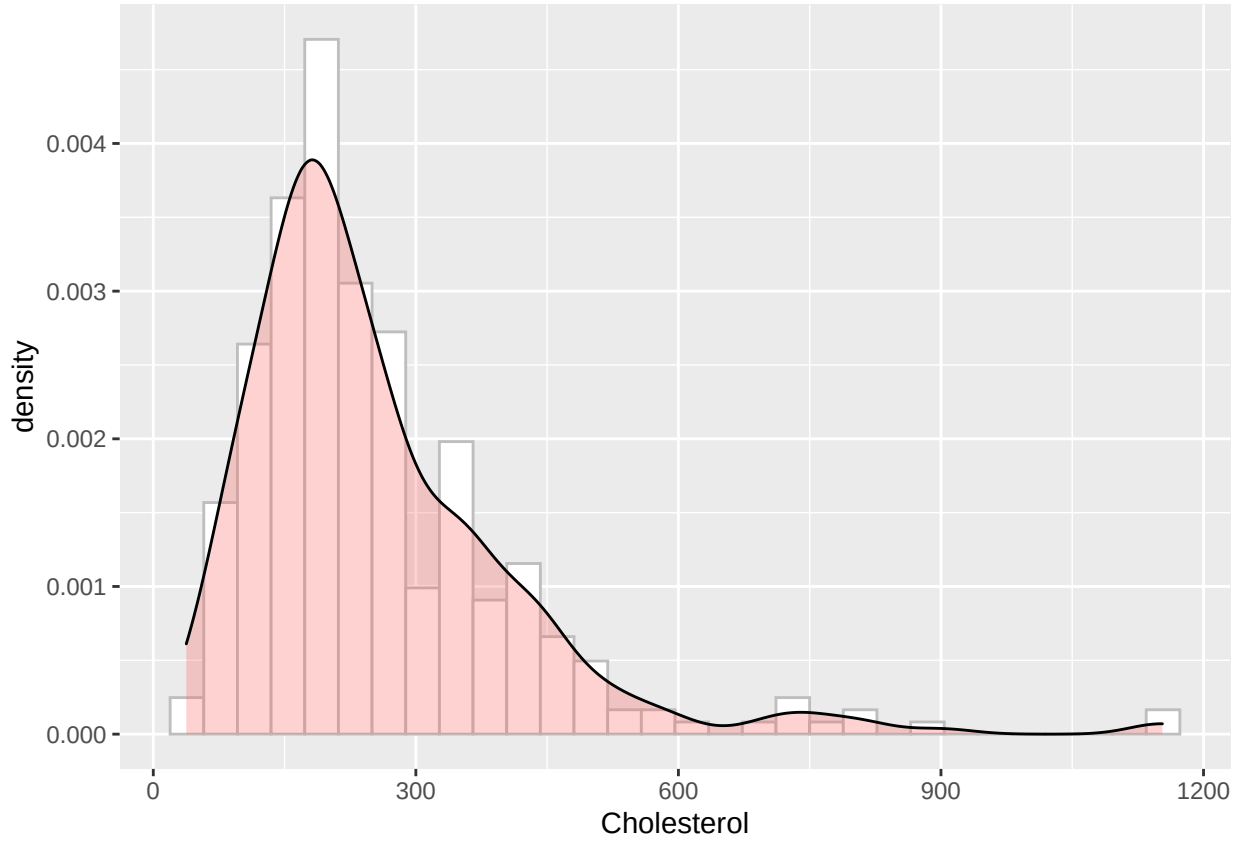
```
# Todas las variables cuantitativas
numSummary(Dataset[,c("Cholesterol"), drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles", "skew
```

```
##      mean       sd    IQR skewness kurtosis   0%   25%   50%    75% 100%   n
##  255.7057 159.1196 169.15 2.176161 7.418406 37.7 156.8 211.7 325.95 1153 315
```

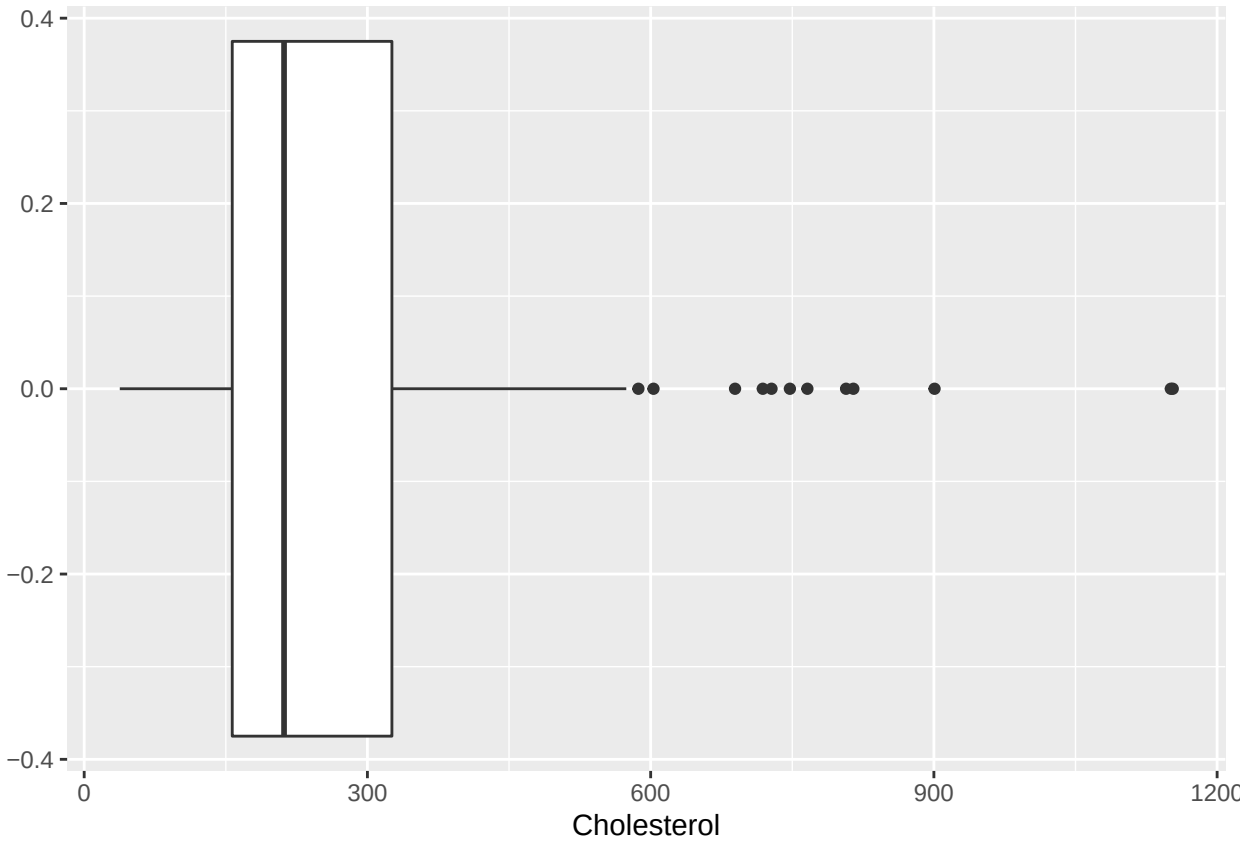
Observando los datos de la variable Colesterol se puede decir que la media es de 255.7057, con una desviación típica de 159.1196, rango en (37.7,1153). También obtenemos un valor de skewness positivo, esto indica que la distribución tiene una asimetría positiva. El 50% de los valores de colesterol en las personas supera el 211.7. La muestra es de tamaño n = 363.

Gráficas de histograma y box plot de la variable Colesterol:

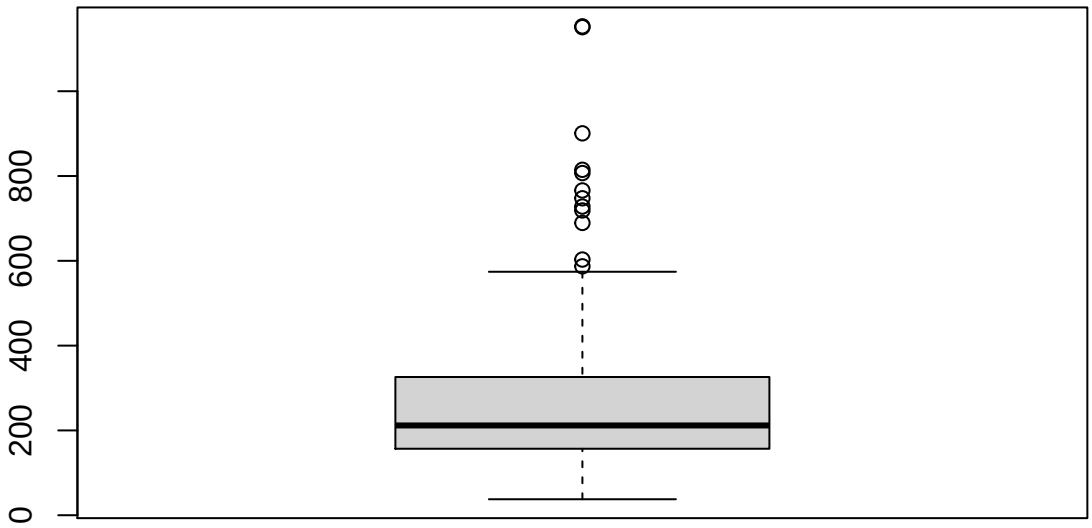
```
# Histogram with density plot
ggplot(Dataset, aes(x=Cholesterol)) +
 geom_histogram(aes(y=..density..), colour="gray", fill="white", bins = 30) +
 geom_density(alpha=.3, fill="#FF6666")
```




```
# Basic box plot
ggplot(Dataset, aes(x=Cholesterol)) +
  geom_boxplot()
```




```
# boxplot
boxplot(Dataset$Cholesterol)
```



Con la Gráfica se puede confirmar el comportamiento *asimétrico positivo*, deducido a partir del valor skewness retornado en el resumen de los datos. También al ver los diagramas de cajas se puede observar que los datos no siguen una distribución normal por la desproporción de sus cuartiles. Por lo que antes de buscar intervalos de confianza y demás datos de interes primero se validará la normalidad.

Comprobación de normalidad de los datos con distintas pruebas estadísticas

Contraste de Hipótesis:

1. H0: La distribución de la variable aleatoria sigue una distribución normal
2. H1: La distribución de la variable aleatoria no sigue una distribución normal

Test de normalidad

Aplicar distindos test de normalidad es una buena practica dado que cada una de las pruebas tienes distintos fundamentos teoricos. En este caso usaremos los test de normalidad de Shapiro, Anderson-Darling, Cramer.

```
normalityTest(~Cholesterol, test="shapiro.test", data=Dataset) # Shapiro
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Cholesterol
## W = 0.82583, p-value < 2.2e-16
```

```
normalityTest(~Cholesterol, test="ad.test", data=Dataset)   # Anderson-Darling
```

```
##
##  Anderson-Darling normality test
```

```
##
## data:  Cholesterol
## A = 11.667, p-value < 2.2e-16
```

```
normalityTest(~Cholesterol, test="lillie.test", data=Dataset)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Cholesterol
## D = 0.14001, p-value < 2.2e-16
```

```
normalityTest(~Cholesterol, test="cvm.test", data=Dataset)  # Cramer
```

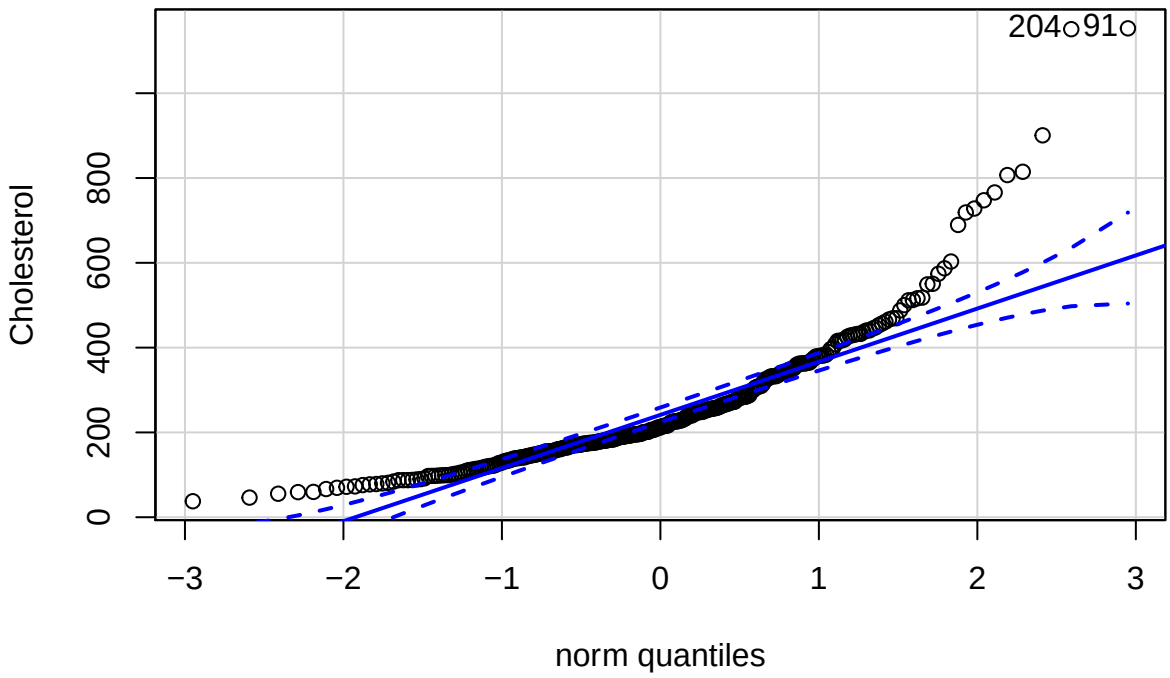
```
## Warning in cvm.test(x = c(170.3, 75.8, 257.9, 332.6, 170.8, 154.6, 255.1, : p-
## value is smaller than 7.37e-10, cannot be computed more accurately
```

```
##
##  Cramer-von Mises normality test
##
## data:  Cholesterol
## W = 1.9866, p-value = 7.37e-10
```

```
normalityTest(~Cholesterol, test="pearson.test", data=Dataset)
```

```
##
##  Pearson chi-square normality test
##
## data:  Cholesterol
## P = 101.06, p-value = 5.649e-14
```

```
# QQPlot
with(Dataset, qqPlot(Cholesterol, dist="norm", id=list(method="y", n=2, labels=rownames(Cholesterol))))
```




```
## [1]  91 204
```

Los valores arrojados por las pruebas indican que p-value $<$ 2.2e-16 y en el mejor de los casos p-value $=$ 7.37e-10(Cramer) por lo que rechazamos la hipótesis nula y podemos asumir que la función de distribución de la variable aleatoria Colesterol es distinta a la normal. De igual forma se puede corroborar con la gráfica qq-plot que los valores sobresalen en gran parte de los límites, lo que es un indicio de que no hay normalidad.

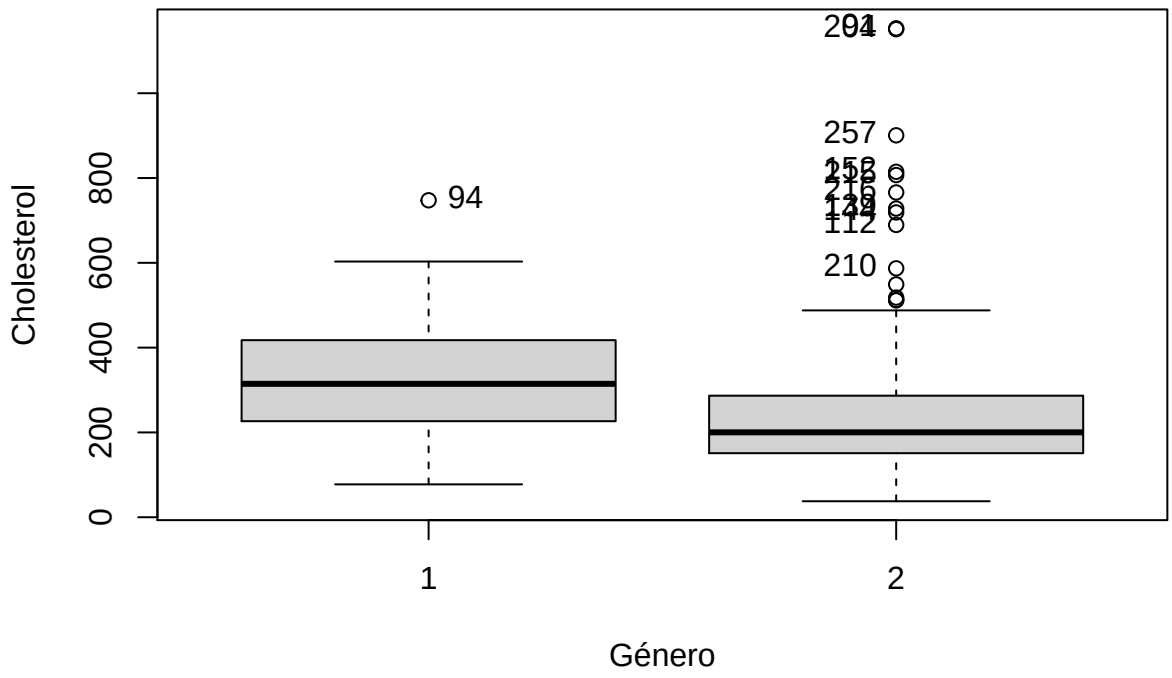
2.Datos Agrupados por género

Análisis de datos tomando en cuenta el factor género en la variable colesterol

```
# Resumen estadístico de la variable colesterol por genero
numSummary(Dataset[,c("Cholesterol"), drop=FALSE], groups = Dataset$Género, statistics=c("mean", "sd",
```

```
##       mean       sd   IQR  skewness  kurtosis   0% 25%   50%   75%   100%
## 1 328.1238 145.4333 190.2 0.6459659 0.4393445 77.5 227 314.6 417.2  747.5
## 2 244.5645 158.4489 135.7 2.4970890 9.3007085 37.7 151 200.2 286.7 1153.0
##   Cholesterol:n
## 1            42
## 2           273
```

```
# Grafica de diagrama de caja de Cholesterol~Género
Boxplot(Cholesterol~Género, data=Dataset, id=list(method="y"))
```




```
##  [1] "94"  "91"  "204" "257" "152" "215" "216" "139" "144" "112" "210"
```

Al generar un resumen con los datos agrupados por genero se aprecia que el género 1 cuenta con un total de 42 muestras que equivalen a un aproximado del 13.333% de la muestra total, estos datos tienen una media de 328.1238 con valor mínimo 77.5 y máximo 747.5. También se puede notar que el valor de skewness está en un rango aceptable entre cero y uno. Por otro lado, el género 2 tiene un total de 273 muestras que equivale a un 86.666% de la muestra total con una media de 244.5645 con valor mínimo 37.7 y máximo 1153.0. El valor de skewness es bastante elevado(mayor a dos) indica que la distribución es asimétrica positiva.

Validamos que los datos siguen una distribución Normal

```
# Prueba de normalidad por Región
normalityTest(Cholesterol~Género, test="shapiro.test", data=Dataset) # Shapiro
```

```
##
##  --------
##  Género = 1
##
##  Shapiro-Wilk normality test
##
## data:  Cholesterol
## W = 0.96649, p-value = 0.2506
##
##  --------
##  Género = 2
##
##  Shapiro-Wilk normality test
```

```
##
## data:  Cholesterol
## W = 0.78788, p-value < 2.2e-16
##
##  --------
##
##  p-values adjusted by the Holm method:
##   unadjusted adjusted
## 1 0.25062    0.25062
## 2 < 2e-16    < 2e-16
```

```
normalityTest(Cholesterol~Género, test="ad.test", data=Dataset)   # Anderson-Darling
```

```
##
##  --------
##  Género = 1
##
##  Anderson-Darling normality test
##
## data:  Cholesterol
## A = 0.44132, p-value = 0.2764
##
##  --------
##  Género = 2
##
##  Anderson-Darling normality test
##
## data:  Cholesterol
## A = 12.58, p-value < 2.2e-16
##
##  --------
##
##  p-values adjusted by the Holm method:
##   unadjusted adjusted
## 1 0.27638    0.27638
## 2 < 2e-16    < 2e-16
```

```
normalityTest(Cholesterol~Género, test="lillie.test", data=Dataset)
```

```
##
##  --------
##  Género = 1
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Cholesterol
## D = 0.11392, p-value = 0.1858
##
##  --------
##  Género = 2
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Cholesterol
## D = 0.14748, p-value = 9.158e-16
```

```
##
##  --------
##
##  p-values adjusted by the Holm method:
##    unadjusted adjusted
## 1 0.18576    0.18576
## 2 9.1582e-16 1.8316e-15
```

```
normalityTest(Cholesterol~Género, test="cvm.test", data=Dataset)  # Cramer
```

```
##
##  --------
##  Género = 1
##
##  Cramer-von Mises normality test
##
## data:  Cholesterol
## W = 0.073658, p-value = 0.2441
##
##  --------
```

```
## Warning in cvm.test(x = c(170.3, 75.8, 257.9, 332.6, 170.8, 154.6, 255.1, : p-
## value is smaller than 7.37e-10, cannot be computed more accurately
```

```
##
##  Género = 2
##
##  Cramer-von Mises normality test
##
## data:  Cholesterol
## W = 2.1513, p-value = 7.37e-10
##
##  --------
##
##  p-values adjusted by the Holm method:
##    unadjusted adjusted
## 1 0.24408    0.24408
## 2 7.37e-10   1.474e-09
```

Al realizar las pruebas de normalidad agrupando los datos por Género se obtuvo que el género uno sigue una distribución normal con un p-valor mayor a 0.05. Sin embargo, el género dos no sigue una distribución normal. Por lo que se tendrán que usar otros métodos que no requieran la normalidad de los datos para realizar comparaciones de media y varianza de estos dos conjuntos.

Determinar si influye el factor GÉNERO en la variable CHOLESTEROL

Para determinar si el factor genero influye en la variable CHOLESTEROL validaremos que ambos grupos tienen varianza y media igual, teniendo en cuenta que se deben usar validaciones como Levene (comprobar igualda de varinaza) y Wilcox (comprobar igualdad de media) dado que una de cuando una de las muestras no sigue una distribución normal.

Validamos si las dos varianzas son iguales o no

```
# Test Levene para dos varianzas con distribuciones distintas a la normal
leveneTest(Cholesterol~Género, data=Dataset)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.5507 0.4586
##       313
```

```
leveneTest(Cholesterol~Género, data=Dataset, center=mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value Pr(>F)
## group   1  0.2281 0.6333
##       313
```

```
# Test Fligner-Killeen para dos varianzas con distribuciones distintas a la normal
fligner.test(Cholesterol~Género, data=Dataset)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Cholesterol by Género
## Fligner-Killeen:med chi-squared = 3.7732, df = 1, p-value = 0.05208
```

El valor obtenido por el p-value es mayor a 0.05 lo que significa que aceptamos la hipótesis nula y asumimos que no hay una diferencia significativa entre las varianzas del colesterol agrupada por género.

Validamos si las dos medias son iguales o no

```
wilcox.test(Cholesterol~Género, data = Dataset)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Cholesterol by Género
## W = 8037, p-value = 2.764e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
#wilcox.test(Cholesterol~Género, data = Dataset, paired = TRUE)
#wilcox.test(Dataset~Homicidios, Dataset~Región, alternative = "g")
```

El valor del p-value es menor a 0.05 por lo que rechazo la hipótesis nula y se determina que el factor genero influye en el colesterol.