03GIAR Probabilidad y Estadística.

Actividad 1. Evidencia 2

2024@LUCIA, RODRIGUEZ, Y8678805E

Fecha de entrega: 21/12/2024

Ejercicio 1 (Momentos centrales y dispersión sobre datos)

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```
head(diamonds) ##Instalo libreria necesaria para analizar variable
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

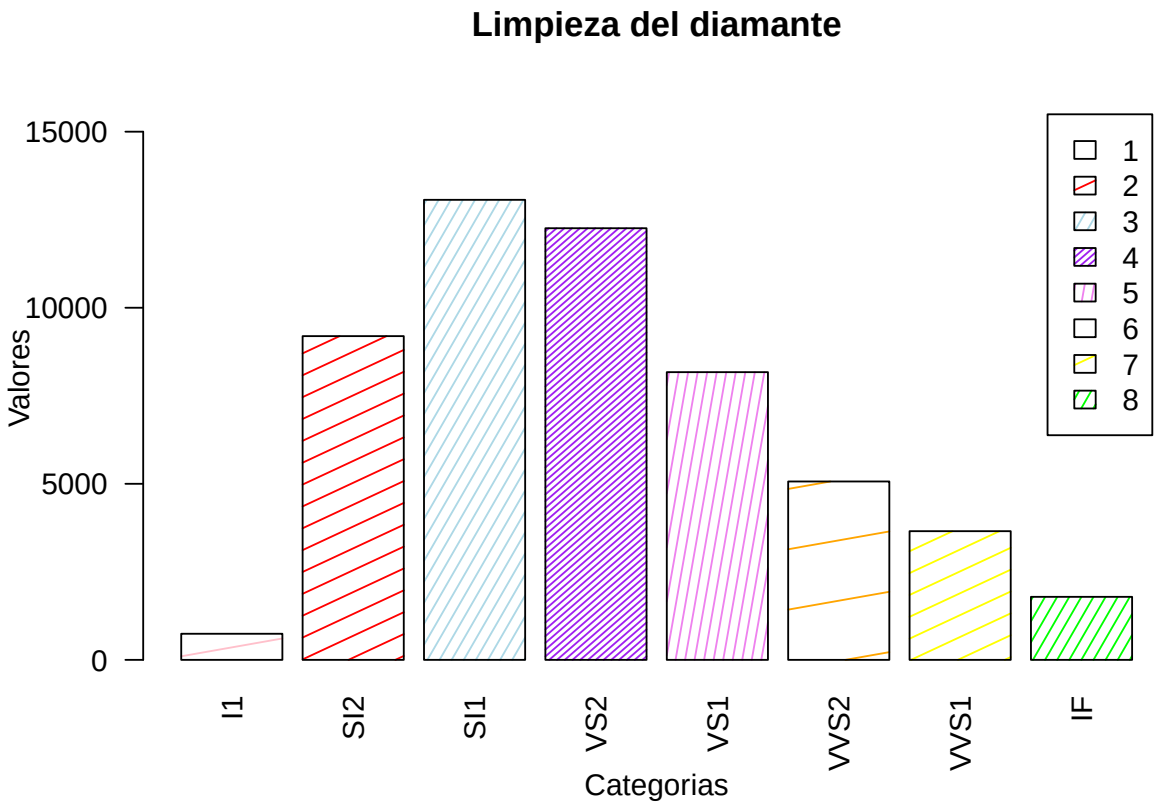
```
vc<-diamonds$clarity
#La variable "Clarity" es una variables discreta,
#ya que representa la claridad de un diamante.
#Representa niveles categoricos.
v2<-table(vc)    #Creo una tabla de valores
v2
```

```
## vc
##    I1   SI2   SI1   VS2   VS1  VVS2  VVS1    IF
##   741  9194 13065 12258  8171  5066  3655  1790
```

```
# Defino vector con etiquetas para la variable "Clarity"
lv2<-c("I1","SI2","SI1","VS2","VS1","VVS2","VVS1","IF")
df1<-data.frame(lv2,v2)  #Creo un data frame
```

```
# Diagrama de barras, uso este grafico ya que es ideal para representar datos
# discretos. Permite visualizar frecuencias por categorias y comparar valores.
barplot(height=df1$Freq, names=df1$lv2,las=2,
```

```
        main="Limpieza del diamante",#titulo del grafico
        ylab="Valores",
        xlab="Categorias",
        ylim=c(0,16000),
        ##4 opciones para las barras y leyenda
        col=c("pink","red","lightblue","purple","violet","orange","yellow","green"),
        legend=rownames(df1),
        density=c(3,9,17,35,19),
        angle=c(10,25,60,41,80),)
```




```
#El siguiente diagrama de barras nos muestra como se distribuyen los datos por los niveles según
#el tipo indicado.Nos muestra que las categorias SI1 y VS2, que representan diamantes, de calidad
#moderada son los mas frecuentes, con valores entre 10000-15000, y por tanto podemos observar
#que para este conjunto de datos, las categorias I1, que representan a aquellos con calidad baja,
#y a los incluidos en la categoria IF, que representan a los de mas alta calidad, que son los menos fre
#Realizo calculo de estadisticos, en el caso de esta variable por su definicion discreta, no seria
#aplicable calcular, por ejemplo, medidas como la media aritmetica, para este conjunto de datos.
#Por esto evaluare su moda.
```

```
## moda de clarity: SI1
```

```
## [1] "SI1"
```

```
#Como podemos ver definiendo la moda de esta variable, nos confirma lo visto en el grafico anterior,
#donde la mayor frecuencia esta en la categoria SI1. Podemos tambien verla graficamente con el
#Diagrama de Pareto, por ejemplo:
```

```
## [1] "SI1"
```

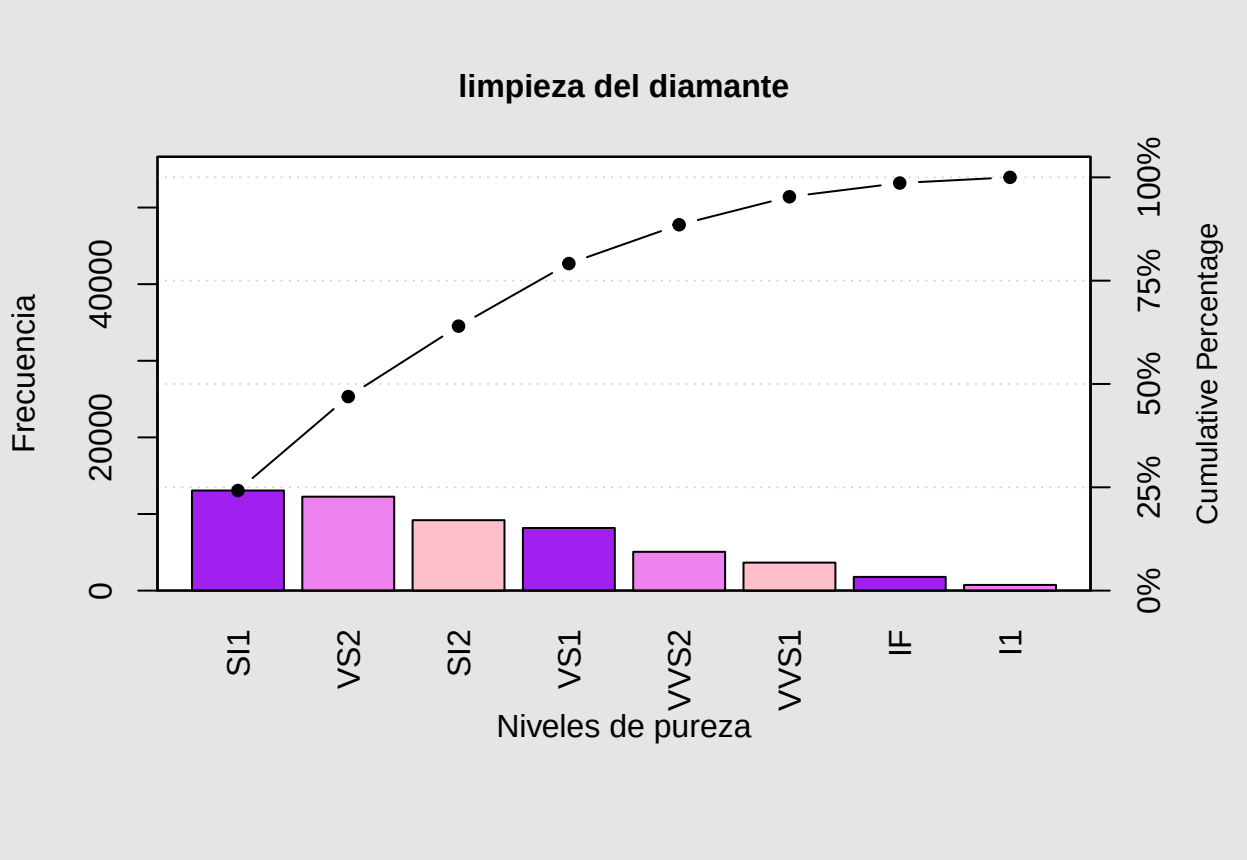
```
## Installing package into 'C:/Users/lulia/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)
```

```
## package 'qcc' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\lulia\AppData\Local\Temp\RtmpSOYOIy\downloaded_packages
```

```
## Warning: package 'qcc' was built under R version 4.4.2
```

```
## Package 'qcc' version 2.7
```

```
## Type 'citation("qcc")' for citing this R package in publications.
```




```
##
## Pareto chart analysis for table(diamonds$clarity)
##           Frequency     Cum.Freq.    Percentage Cum.Percent.
```

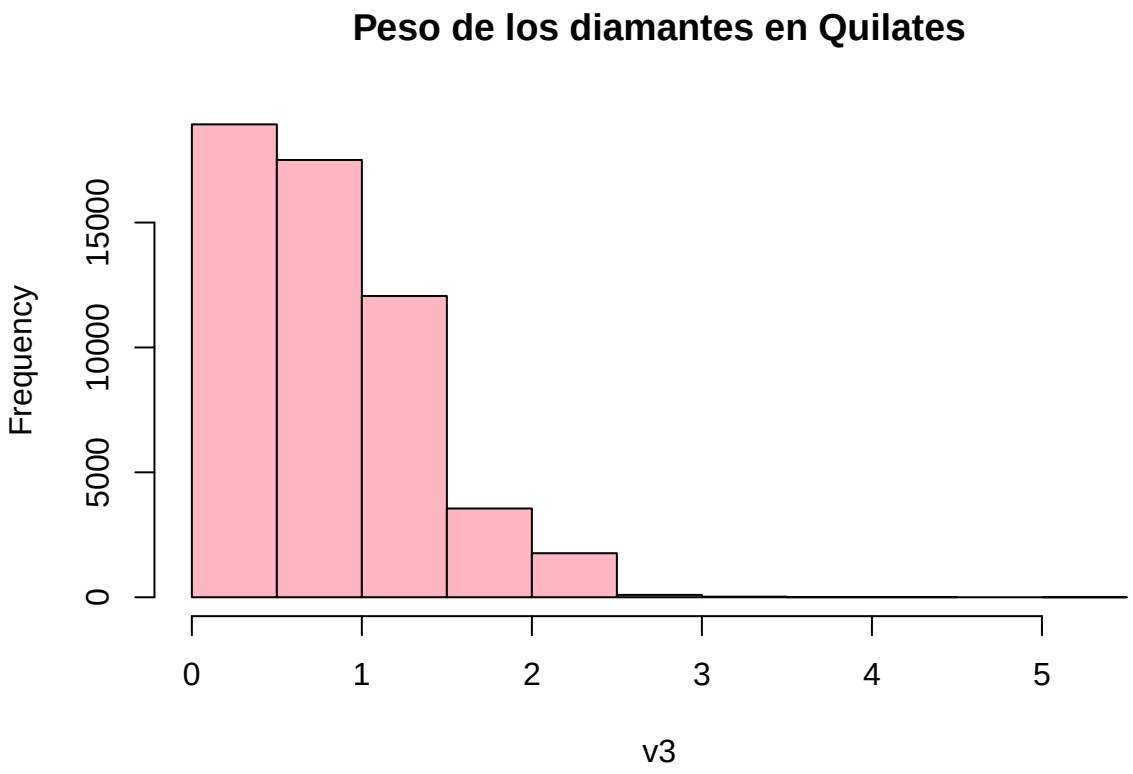
```
##   SI1  13065.000000 13065.000000    24.221357    24.221357
##   VS2  12258.000000 25323.000000    22.725250    46.946607
##   SI2   9194.000000 34517.000000    17.044865    63.991472
##   VS1   8171.000000 42688.000000    15.148313    79.139785
##   VVS2  5066.000000 47754.000000     9.391917    88.531702
##   VVS1  3655.000000 51409.000000     6.776047    95.307749
##   IF    1790.000000 53199.000000     3.318502    98.626251
##   I1     741.000000 53940.000000     1.373749   100.000000
```

```
#Graficamente confirmamos, lo expuesto anteriormente, siendo los diamantes con caracteristicas SI1,
#los de mayor frecuencia.
```

```
#Variable aleatoria continua "Carat"
v3<-diamonds$carat
```

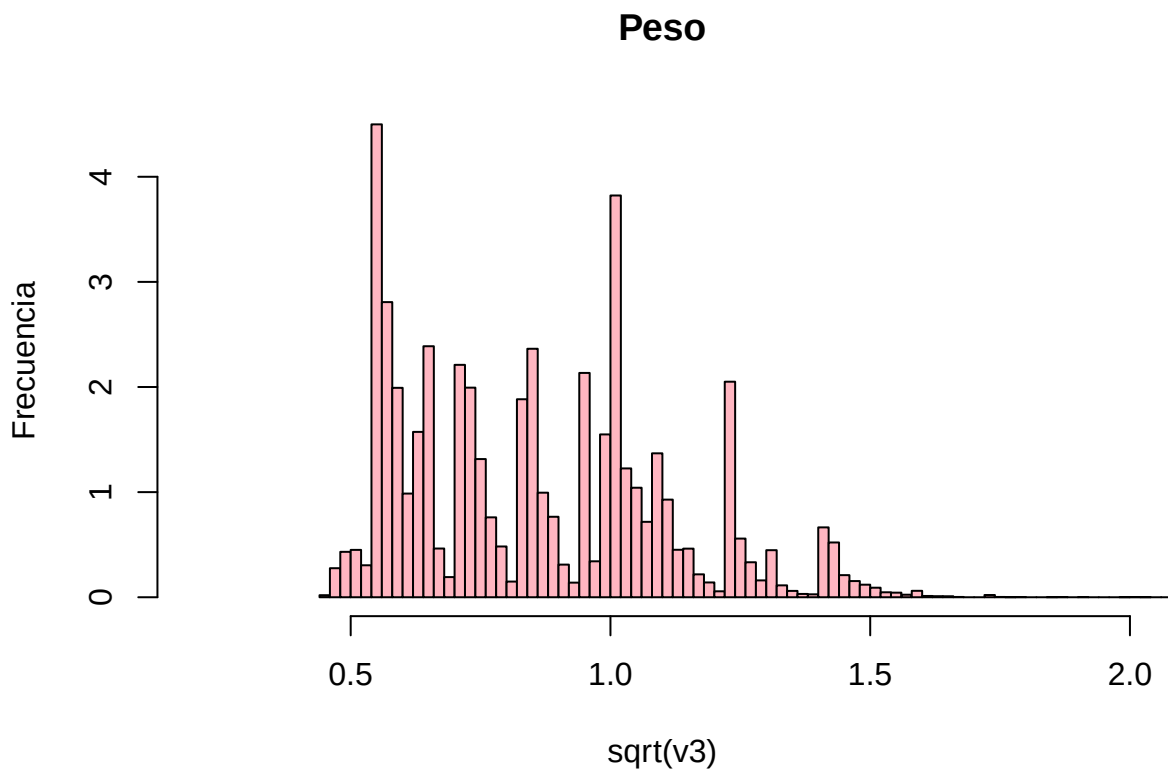
```
#Si realizamos un histograma los datos se muestran como una distribucion que refleja la frecuencia
#de los distintos pesos, valorados en Quilates, de la variable "Carat", a simple vista podemos
#observar que aquellos que presentan mayor frecuencia, seran los valorados, entre 0 y 1.
```

```
#Hacemos un histograma basico
hist (v3, main = "Peso de los diamantes en Quilates",
     #xlab = "$",
     #ylab="Frecuencia",
     #prob = TRUE,
     col="lightpink",
     #ylim = c(0,0.44)
)
```




```
#Si realizamos una transformacion a los datos con la raiz cuadrada, se expone una reduccion
#en la asimetria de la distribucion, lo que facilita el analisis y la interpretacion de los
#datos al acercarlos mas a una distribucion simetrica. Esto es util, ya que al aplicar
#la raiz cuadrada, los valores grandes se reducen proporcionalmente mas que los pequeños,
#obteniendo un sesgo disminuido. Y asi creando graficamente, una opcion visualmente mas clara.
```

```
#Hacemos un histograma pero aplicamos raiz cuadrada a "v3"

hist (sqrt(v3),
     breaks=85,main = "Peso",
     #xlab = "$",
     ylab="Frecuencia",
     prob = TRUE, col="lightpink",
     xlim=c(0.2,2)
     #ylim = c(0,0.44)
     )
```




```
#Defino la moda
mean(v3)
```

```
## [1] 0.7979397
```

```
library(multimode)
```

```
## Warning: package 'multimode' was built under R version 4.4.2
```

```
modas<- locmodes(sqrt(v3),
                 mod0 =5) # Localiza la moda
```

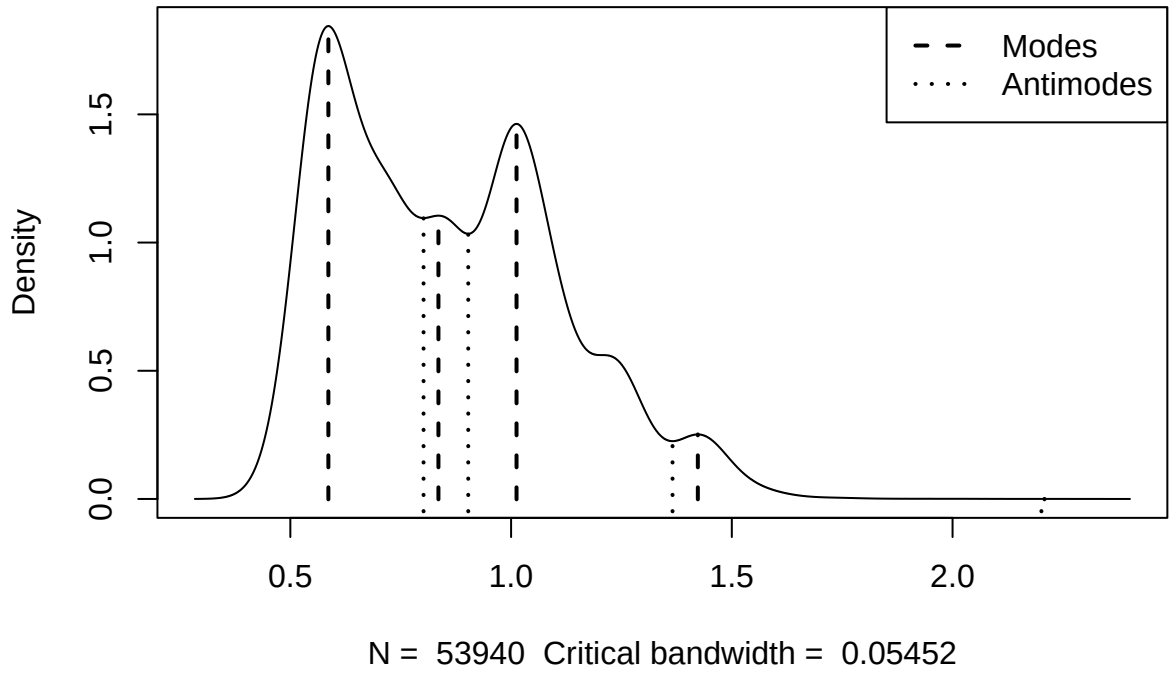
```
## Warning in locmodes(sqrt(v3), mod0 = 5): If the density function has an
## unbounded support, artificial modes may have been created in the tails
```

```
modas
```

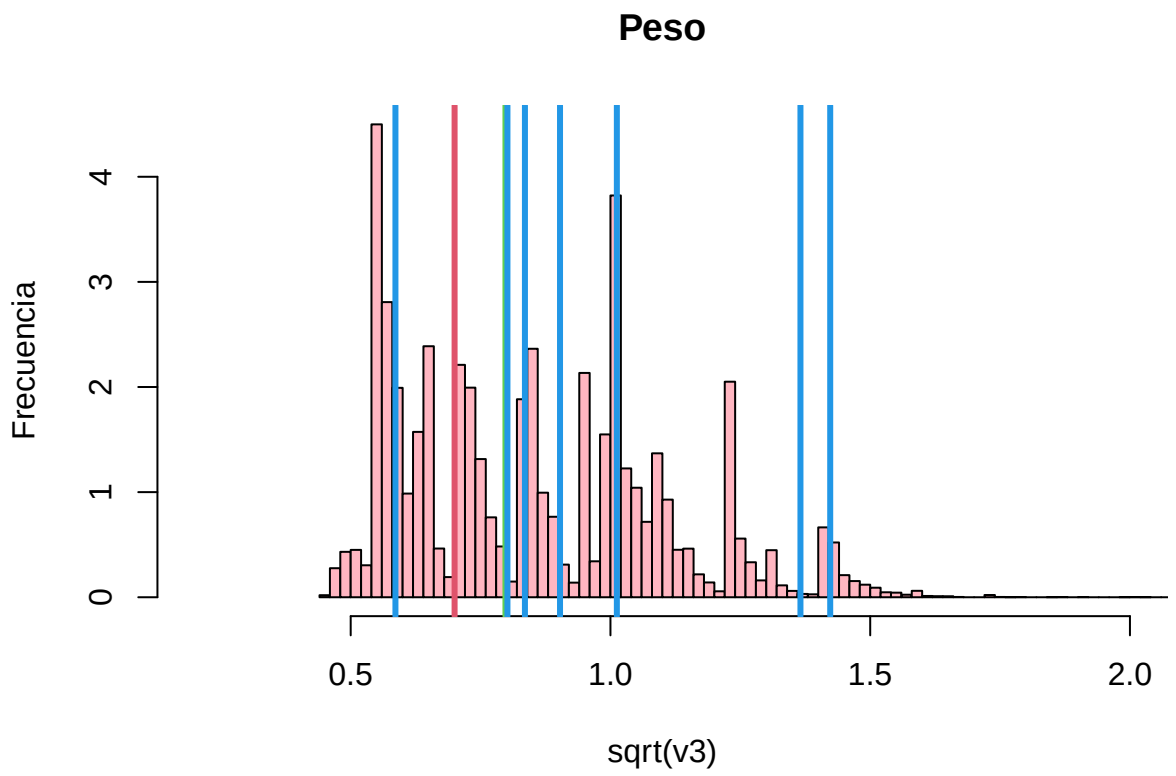
```
##
## Estimated location
## Modes: 0.5861264  0.8353957  1.012328  1.423014  2.208252
## Antimodes: 0.8017159  0.9029492  1.365739   2.20127
##
## Estimated value of the density
```

```
## Modes: 1.844465  1.104948   1.46284  0.2516582  0.0001556416
## Antimodes: 1.095612  1.034233  0.2254498  0.000155611
##
## Critical bandwidth: 0.05451965
```

```r
plot(modas)
```

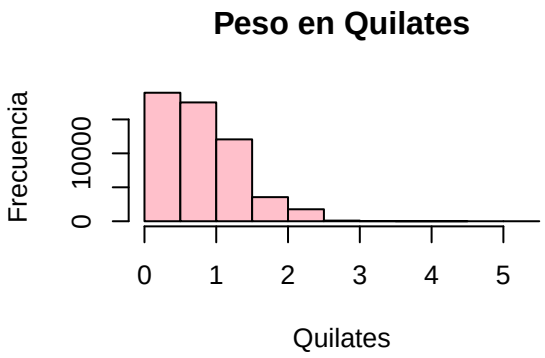


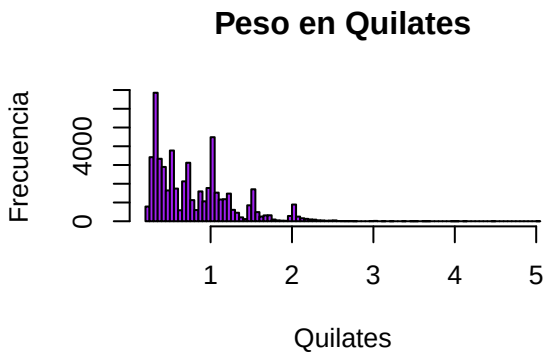

```r
#Subjetivamente ¿ Dónde estarían los estadísticos clásicos?

hist(sqrt(v3),
     breaks=85,main = "Peso",
     #xlab = "$",
     ylab="Frecuencia",
     prob = TRUE,
     col="lightpink",
     xlim=c(0.2,2)
     #ylim = c(0,0.44)
     )
#Graficamos cada estadistico

abline(v = mean(v3), col = 3, lwd = 3) # Línea para media aritmética.
abline(v = median(v3), col = 2, lwd = 3) # Línea para mediana.
abline(v=modas[[1]], col = 4, lwd = 3) # Línea para la Moda
```

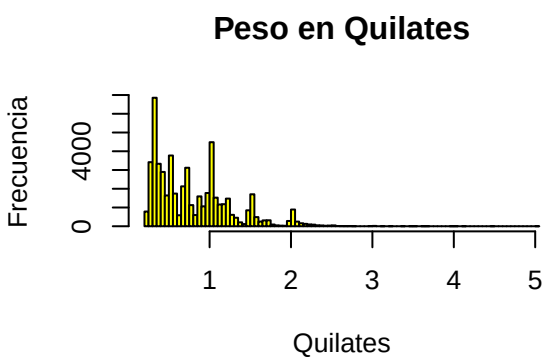



```r
#Apliquemos algunos métodos para los intervalos: #"Sturges", "Scott" y "freedman-diaconis"
#Metodo Sturges
par(mfrow = c(2, 2)) # Desplegamos 4 paneles
hist(v3,breaks="sturges", main = "Peso en Quilates",
     xlab = "Quilates",
     ylab="Frecuencia",
     col="pink", #ylim = c(0,0.44)
     )
#Freedman-Diaconis
hist(v3, breaks="fd", main = "Peso en Quilates",
     xlab = "Quilates",
     ylab="Frecuencia",
     col="purple",
     #ylim = c(0,0.44)
     )
#Scott
hist(v3, breaks="scott", main = "Peso en Quilates",
     xlab = "Quilates",
     ylab="Frecuencia",
     col="yellow",
     #ylim = c(0,0.44)
     )
#¿Con cuál nos quedamos?
#Teniendo en cuanta los datos de esta variable, podemos ver que tanto en metodo Scott o el
#Freedman-Diaconis, reflejan la informacion, graficamente de manera similar, por tanto
#cualquiera de estos, cumplirian, mejor con la muestra grafica, que la opcion de Sturges.
```

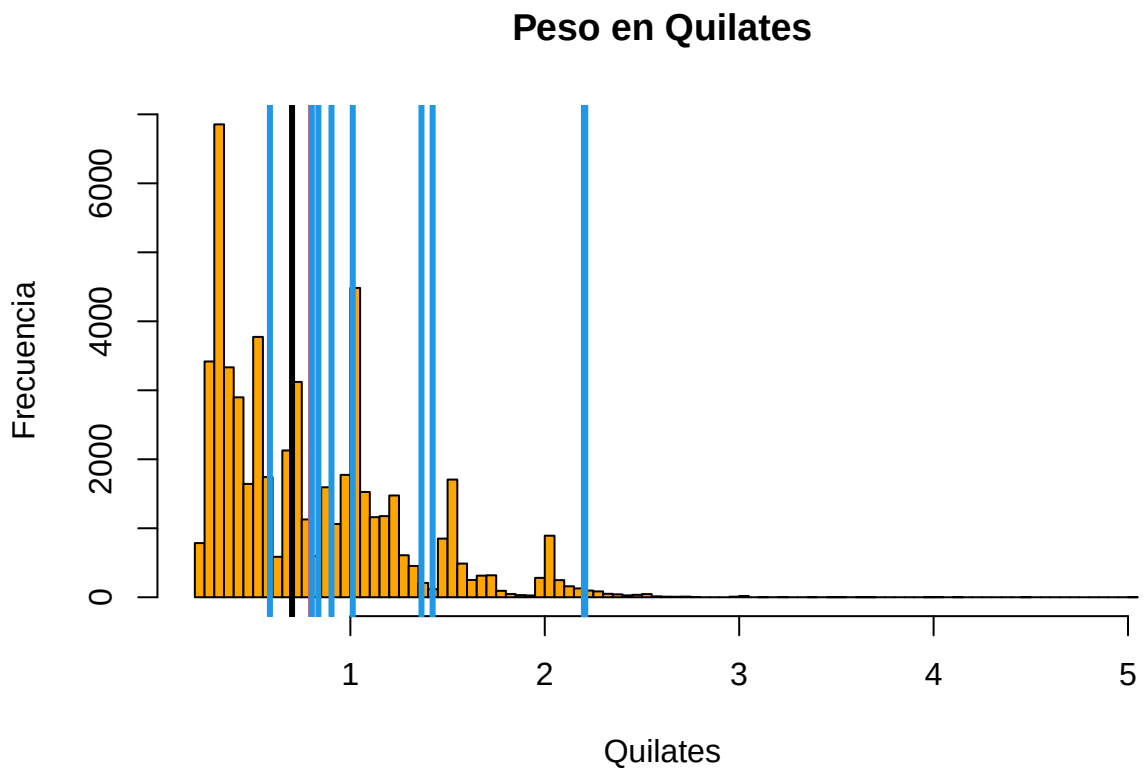
```r
#Por ejemplo si optamos por el metodo Scott:
par(mfrow = c(1,1))
```





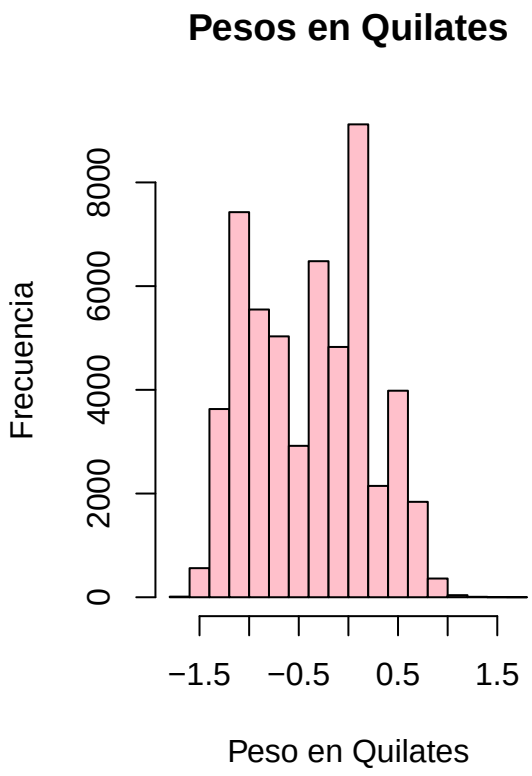


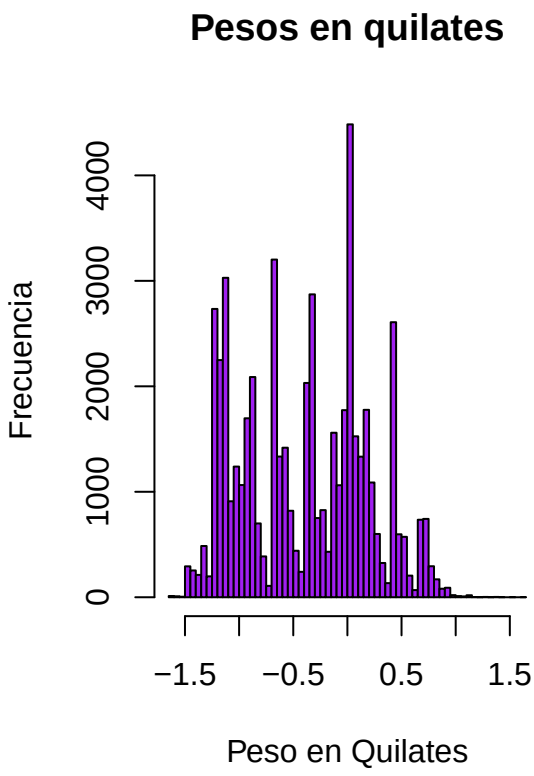

```r
hist(v3, breaks="scott", main = "Peso en Quilates",
     xlab = "Quilates",
     ylab="Frecuencia",
     col="orange",
     xlim = c(min(v3),max(v3))
     #ylim = c(0,0.44)
     )
abline(v = mean(v3), col = 2, lwd = 3) # Línea para media aritmética.
abline(v = median(v3), col = 1, lwd = 3) # Línea para mediana. modas
abline(v =modas[[1]], col = 4, lwd = 3) # Línea para la Moda.
```



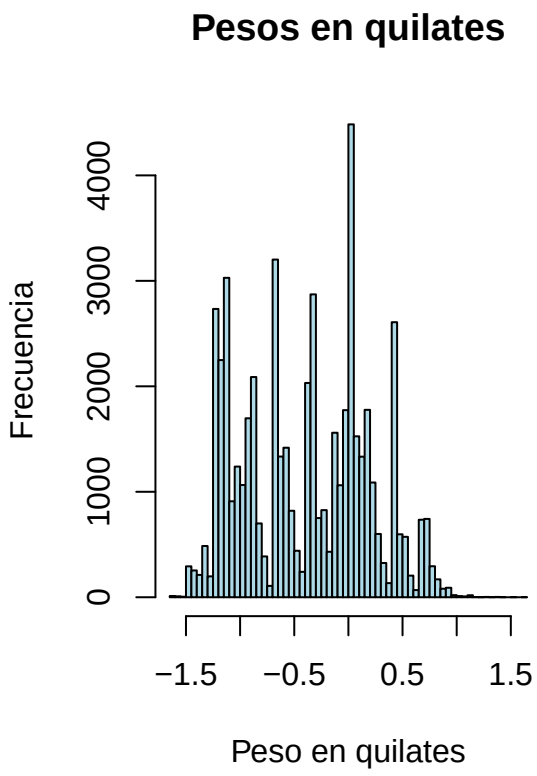

```
#En este punto podemos ver que ni la media, ni la mediana ni 1 sola moda, representan correctamente,
#lo que necesitamos. Veamos si podemos exponer mas informacion, de estos datos, aplicando logaritmo.
```

```
par(mfrow=c(1,2))
hist(log(v3), breaks="sturges",
    main = "Pesos en Quilates",
    xlab = "Peso en Quilates",
    ylab="Frecuencia", col="pink",
    #ylim = c(0,0.44)
    )

hist(log(v3), breaks="fd",
     main = "Pesos en quilates",
     xlab = "Peso en Quilates",
     ylab="Frecuencia",
     col="purple",
     #ylim = c(0,0.44)
     )
```






```r
hist(log(v3), breaks="scott",
     main = "Pesos en quilates",
     xlab = "Peso en quilates",
     ylab="Frecuencia",
     col="lightblue",
     #ylim = c(0,0.44)
     )
```




```
# Definitivamente, ni la media ni la mediana y una sola moda no la representa. Sera multimoda,
#por ejemplo. Calculemos las más significativas.
```

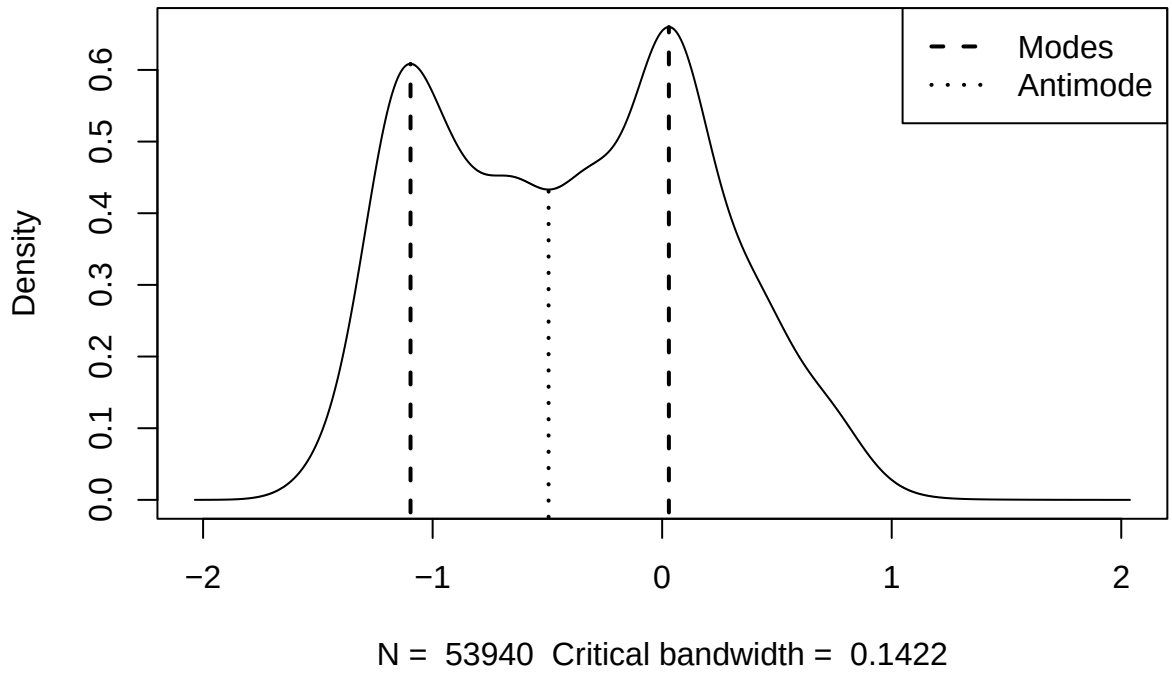
```
par(mfrow=c(1,1))
modas<-locmodes(log(v3), mod0 = 2)     #Localiza la moda
```

```
## Warning in locmodes(log(v3), mod0 = 2): If the density function has an
## unbounded support, artificial modes may have been created in the tails
```

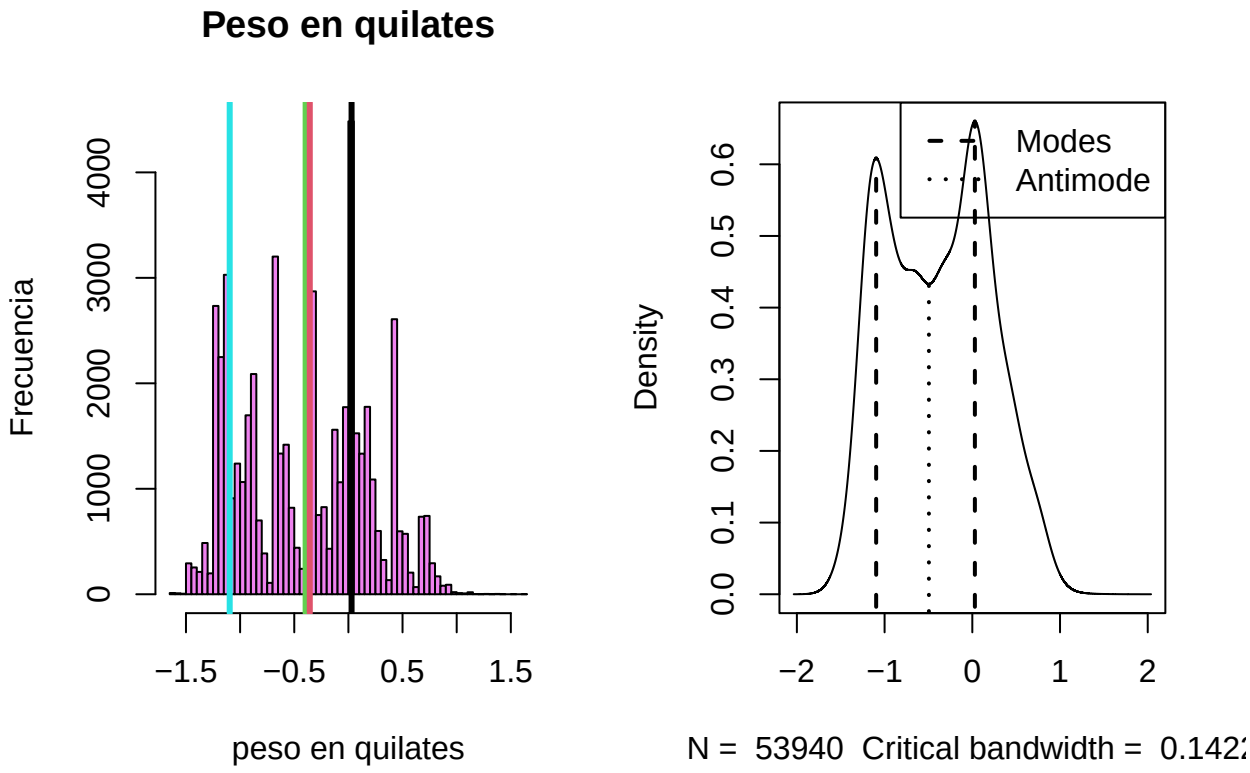
```
modas
```

```
##
## Estimated location
## Modes: -1.096237  0.02928309
## Antimode: -0.4948739
##
## Estimated value of the density
## Modes: 0.6085034  0.6599266
## Antimode: 0.4330636
##
## Critical bandwidth: 0.1421509
```

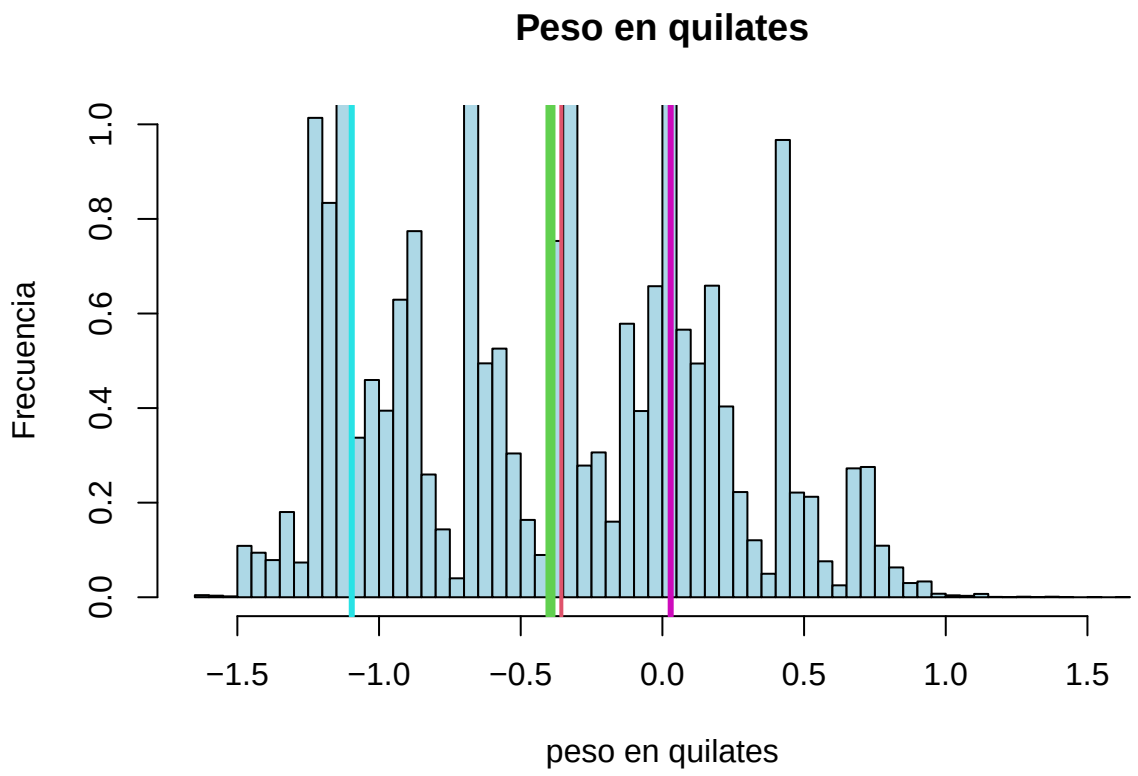
```r
plot(modas)
```



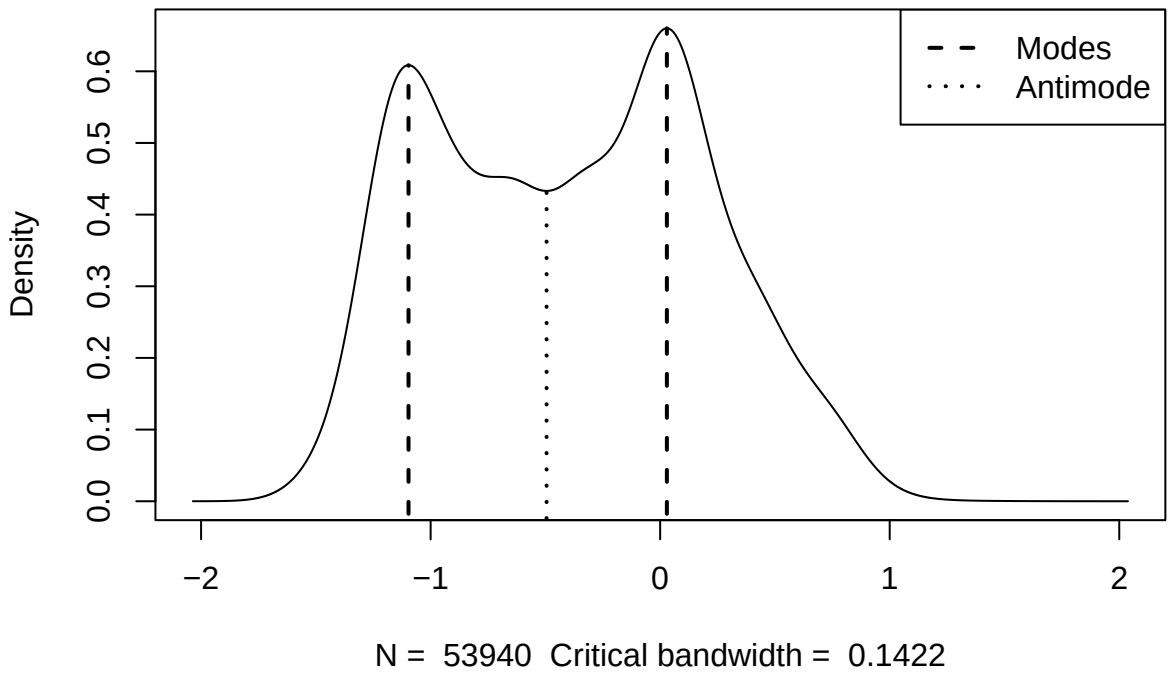

```r
par(mfrow = c(1, 2)) #Desplegamos 2 paneles
hist(log(v3),breaks="scott",
     main = "Peso en quilates",
     xlab = "peso en quilates",
     ylab="Frecuencia",
     col="violet",
     #ylim = c(0,0.44)
     )
abline(v = mean(log(v3)), col = 3, lwd = 3) #Línea para media aritmética.
abline(v = median(log(v3)), col = 2, lwd = 3) #Línea para mediana.
abline(v =modas[[1]][1], col = 5, lwd = 3) #Línea para la Moda 1.
abline(v =modas[[1]][3], col = 1, lwd = 3) #Línea para la Moda 2.
plot(modas)
```




```
# Pero para comparar, mejor en la misma escala los dos gráficos.
hist(log(v3), breaks="scott",
     main = "Peso en quilates",
     xlab = "peso en quilates",
     ylab="Frecuencia",
     prob=TRUE, #Añadimos esta función.
     col="lightblue",
     ylim = c(0,1)
     )
abline(v = mean(log(v3)),col = 3, lwd = 5) # Línea para media aritmética.
abline(v = median(log(v3)), col = 2, lwd = 2) # Línea para mediana.
abline(v = modas[[1]][1], col = 5, lwd = 3) # Línea para la Moda 1.
abline(v = modas[[1]][3], col = 6, lwd = 3) # Línea para la Moda 2.
```


Peso en quilates
Frecuencia
0.0
0.2
0.4
0.6
0.8
1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
peso en quilates

```
plot(modas)
```




```
par(mfrow=c(1,1))
plot(modas) #Como estadísticos de tendencia central tenemos dos valores:
# según la moda:
modas
```

```
##
## Estimated location
## Modes: -1.096237  0.02928309
## Antimode: -0.4948739
##
## Estimated value of the density
## Modes: 0.6085034  0.6599266
## Antimode: 0.4330636
##
## Critical bandwidth: 0.1421509
```

```
mean(log(v3))
```

```
## [1] -0.394967
```

```
# Calculemos otros estadísticos de dispersión.
rg=max(log(v3))-min(log(v3)) # Rango
rg
```

```
## [1] 3.220874
```

```r
var(log(v3))  # Varianza
```

```
## [1] 0.3420232
```

```r
sd(log(v3)) # Desviación estándar
```

```
## [1] 0.5848275
```

```r
library(moments) # Librería para el cálculo de los momentos.
skewness(log(v3))   #Asimetría
```

```
## [1] 0.09610032
```

```r
kurtosis(log(v3)) #Curtosis
```

```
## [1] 1.935102
```

```r
#Obtenidos estos valores, podemos decir
#1)Segun el valor de Varianza, indica que los datos estan relativamente dispersos, aunque
#no de forma excesiva. Estan alejados de la media, por tanto descartamos dicho estadistico.
#2)La desviacion estandar, es la raiz cuadrada de la varianza. En este caso sugiere una
#dispersion moderada.
#3)En el caso, de la Asimetria,podemos decir que en este caso, es casi simetrica, porque
#el valor es cercano a cero, con una leve inclinacion, a derecha (asimetria positiva).
#4)Por ultimo, la curtosis, en este caso, es inferior a 3,por tanto indica una curtosis
#leptocurtica, esto sugiere que hay menos datos extremos.

#Hasta aqui podemos decir, que esta variable parece tener una distribucion moderadamente
#dispersa, con ligera asimetria positiva, lo que indica que los valores estan distribuidos
#de manera bastante equilibrada, hay menos valores extremos. # Elijamos una muestra del
#total de valores disponibles, aplicando un método de muestreo y repitamos los cálculos
#previos y verificamos si son consistentes o no.
```

Muestreo Simple.

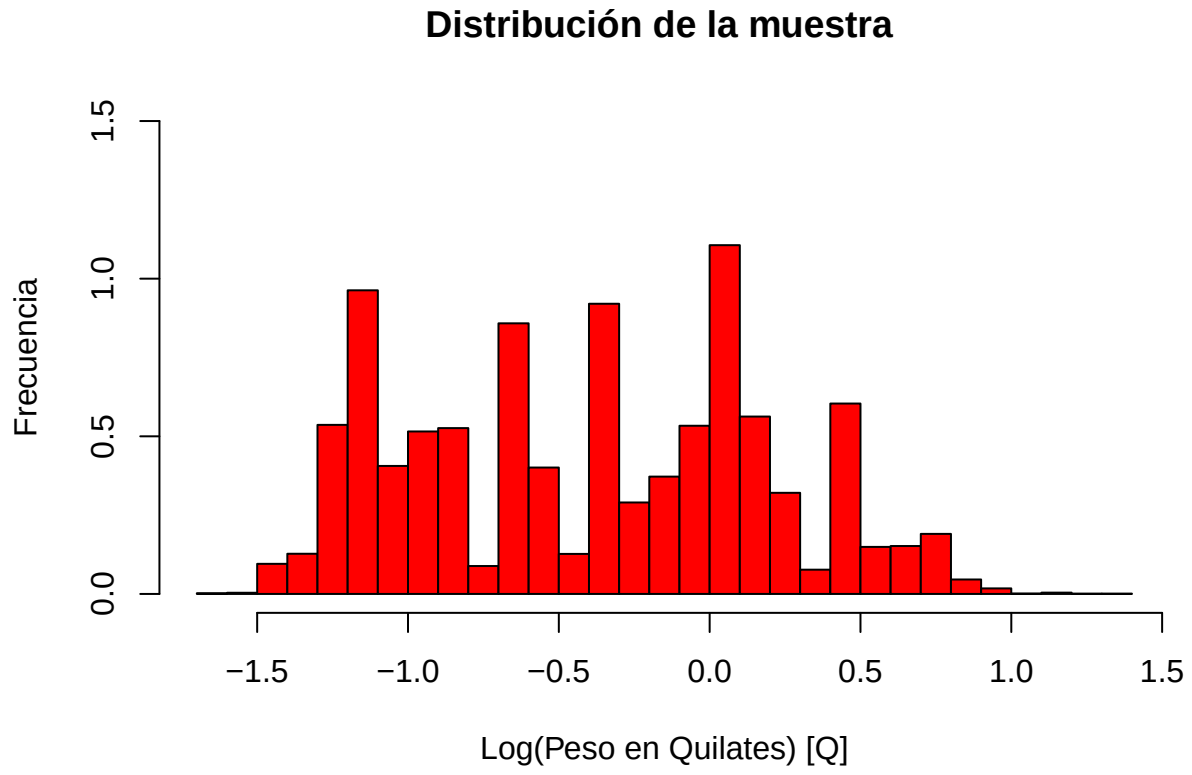
```r
length(v3) # Total de datos disponibles: 53940.
```

```
## [1] 53940
```

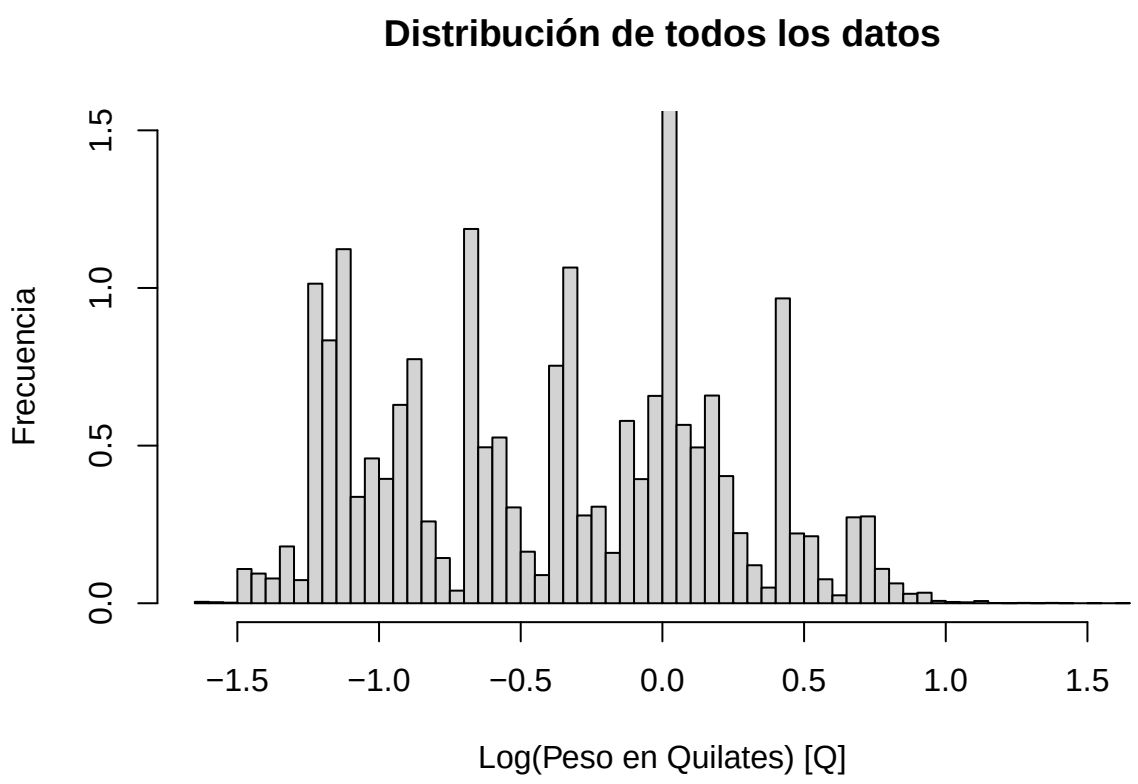
```r
# Selección de la muestra
n<-length(v3)*0.4 # Tamaño de la muestra. 40%
n
```

```
## [1] 21576
```

```r
# Elegimos una muestra de forma aleatoria de los 53940. Para ello, hacemos uso de
v4<- sample(log(v3),size=n,replace=FALSE)

# Gráfico de la muestra "v4" y comparamos con el original.
hist(v4, breaks="scott",
     main = "Distribución de la muestra ",
     xlab = "Log(Peso en Quilates) [Q]",
     ylab = "Frecuencia",
     ylim=c(0,1.5), prob=TRUE,          #Añadimos esta función.
     col="red")    # Nuestro gráfico original, con toda la muestra.
```




```r
hist(log(v3), breaks="scott",
     main = "Distribución de todos los datos",
     xlab = "Log(Peso en Quilates) [Q]",
     ylab="Frecuencia", prob=TRUE,      #Añadimos esta función.      col="blue",
     ylim = c(0,1.5)
     )
```




```
#En esta caso, podemos quedarnos con la muestra, ya que al ver la comparacion,

#Definimos los estadisticos, nuevamente con respecto a la muestra.
mean(v4)  # Media aritmética.
```

```
## [1] -0.3942149
```

```
median(v4) # Mediana.
```

```
## [1] -0.3566749
```

```
library(multimode)
modas <- locmodes(sqrt(v4), mod0 = 2) #Localiza las modas
```

```
## Warning in sqrt(v4): Se han producido NaNs
```

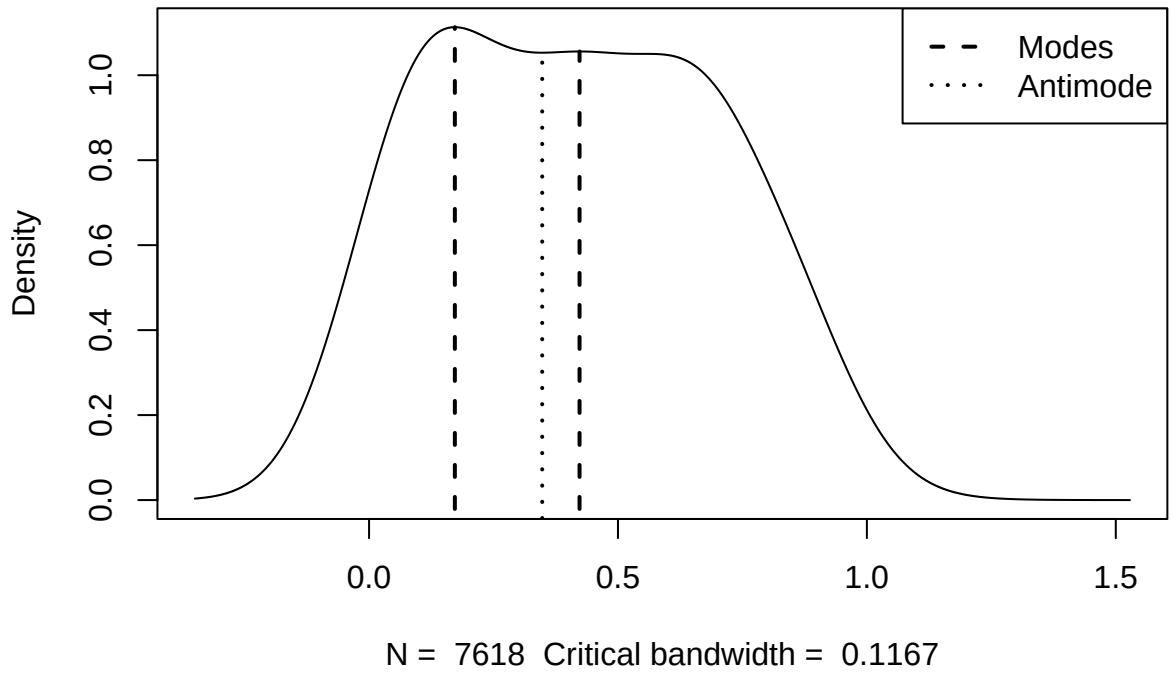
```
## Warning in locmodes(sqrt(v4), mod0 = 2): Missing values were removed
```

```
## Warning in locmodes(sqrt(v4), mod0 = 2): If the density function has an
## unbounded support, artificial modes may have been created in the tails
```

```
modas
```

```
##
## Estimated location
## Modes: 0.1723136  0.4227817
## Antimode: 0.3477961
##
## Estimated value of the density
## Modes: 1.113053  1.055863
## Antimode: 1.052975
##
## Critical bandwidth: 0.1166687
```

```
plot(modas)
```




```
var(v4) # Varianza.
```

```
## [1] 0.3400066
```

```
sd(v4) #Desviación estándar.
```

```
## [1] 0.5831009
```

```r
skewness(v4) #Asimetría.
```

```
## [1] 0.09467696
```

```r
kurtosis(v4) #Curtosis.
```

```
## [1] 1.928556
```

Conclusiones Ejercicio 1:

```r
# Con respecto a la Variable discreta "Clarity" SI1 es el nivel de claridad más
#común en el conjunto de datos. Esto sugiere que una proporción significativa de
#los diamantes tiene pequeños defectos que son "ligeramente incluidas" (Slightly Included 1).
#SI1 suele encontrarse en una calidad intermedia entre diamantes de muy alta calidad
#(como IF, VVS1/VVS2) y aquellos con imperfecciones más visibles (como I1, I2).
#Esto podría indicar que los diamantes del dataset están orientados a un mercado de
#calidad media, donde SI1 ofrece una buena relación entre precio y apariencia.
#Con la herramienta del grafico con el metodo de Pareto, podemos confirmar,que
#la mayor concentracion de diamantes, son aquellos que estan catalogados, como SI1,
#es decir, para este grupo de datos analizados, podemos decir,que la mayor cantidad
#de diamantes son aquellos, de calidad media, o con defectos, poco visibles, seguido
#de aquellos agrupados en la categoria VS2, que como los anteriores, representan a aquellos,
#con ligeros defectos.

# Con respecto a la Variable continua "Carat" : Las modas representan los valores más
#frecuentes. En este caso, se encuentran en dos puntos diferentes, lo que sugiere que
#la distribución podría ser bimodal, es decir, hay dos valores en los que los datos
#se agrupan con mayor frecuencia.  En este caso, ambos son bastante cercanos, lo que
#sugiere que las modas son puntos con alta concentracion de datos. Análisis de la otros
#estadisticos: Varianza (0.3441411): En este caso, una varianza moderada indica que
#los valores de "Carat" no están excesivamente dispersos ni muy concentrados alrededor
#de la media.Desviación estándar (0.5866354): es relativamente baja en comparación con
#el rango. Esto indica que los datos no se desvían demasiado de la media.
#Asimetría(0.09669699): Es cercana a 0, lo que indica que la distribución es casi simétrica.
#No hay una tendencia clara hacia un sesgo positivo o negativo en los datos.
#Curtosis (1.921081): La curtosis es menor que 3, lo que indica que la distribución
#tiene colas más ligeras que una distribución normal. Esto puede sugerir que los valores
#extremos (outliers) son menos frecuentes en comparación con una distribución normal.
#Análisis de Region crítica(0.1091385): Podría indicar la suavidad de la estimación.
#Una region crítica relativamente pequeña sugiere que la estimación de la densidad
#está más ajustada a los puntos de datos y podría no ser tan suave.  Es decir, podría
#haber dos grupos distintos o características con mayor frecuencia en los datos.
#Los datos están relativamente dispersos alrededor de un valor central.
#La curtosis sugiere que la distribución tiene colas ligeras, lo que significa que
#no hay una presencia significativa de valores extremos. El valor de la region crítica
#podría indicar que la estimación de la densidad se ajusta bien a los datos. En conclusion,
#la variable "Carat" muestra una distribución con dos grupos diferentes de datos mas
#frecuentes, no muestra una alta concentración en torno a un único valor,
#sino más bien una dispersión moderada con pocas colas extremas.
```

Ejercicio 2 (L-moments)

```
#Los L-Moments, son otros estadísticos alternativos que nos permiten estudiar ciertas
#distribuciones que se pueden considerar con colas muy pronunciadas, donde los momentos
#clásicos no logran adaptarse bien. Es decir, caracterizan mejor la distribución ya que
#son menos sensibles en los extremos. Para una variable estadística aleatoria $X$ están
#definidos en términos de los ordenes estadísticos ($p-$ésimos) tal que:

install.packages("evd")
```

```
## Installing package into 'C:/Users/lulia/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'evd' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\lulia\AppData\Local\Temp\RtmpSOYOIy\downloaded_packages
```

```
install.packages("lmom") #Cargo librerias necesarias
```

```
## Installing package into 'C:/Users/lulia/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'lmom' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\lulia\AppData\Local\Temp\RtmpSOYOIy\downloaded_packages
```

```
library(lmom)
library(Lmoments)
```

```
## Warning: package 'Lmoments' was built under R version 4.4.2
```

```
library(evd)
```

```
## Warning: package 'evd' was built under R version 4.4.2
```

```
#Simulo un conjunto de datos, de ejemplo, para la distribucion GUMBEL, que se encuentra
#en ambas librerias.
set.seed(123) #Reproducibilidad
n<-100
data_gumbel<- rgumbel(n,loc=0, scale=1)

#Calculo los L-Moments con cada libreria
#Con la libreria lmom:
lmom_lmom<-samlmu(data_gumbel)
lmom_lmom
```

```
##       l_1       l_2       t_3       t_4
## 0.4802932 0.6613717 0.2200150 0.1761222
```

```r
#Con la libreria L-moments
lmom_Lmoments<-Lmoments(data_gumbel)
lmom_Lmoments
```

```
##             L1        L2        L3        L4
## [1,] 0.4802932 0.6613717 0.1455117 0.1164822
```

```r
#Creo un dataframe para comparar los resultados
comparacion<-data.frame(Momentos = c("L1 (media)","L2(varianza)","L3(Asimetria)","L4(curtosis)"),
                        lmom=lmom_lmom,
                        Lmoments=lmom_Lmoments)
print(comparacion)
```

```
##          Momentos      lmom Lmoments.L1 Lmoments.L2 Lmoments.L3 Lmoments.L4
## l_1    L1 (media) 0.4802932   0.4802932   0.6613717   0.1455117   0.1164822
## l_2  L2(varianza) 0.6613717   0.4802932   0.6613717   0.1455117   0.1164822
## t_3 L3(Asimetria) 0.2200150   0.4802932   0.6613717   0.1455117   0.1164822
## t_4  L4(curtosis) 0.1761222   0.4802932   0.6613717   0.1455117   0.1164822
```

```r
#Conclusion ejercicio 2:
#La distribucion Gumbel, es de alta presicion, por tanto en este caso elegiria trabajar
#su analisis con la libreria "lmom". Aun asi podemos ver que con ambas librerias,
#se obtienen datos similares y consistentes. Como momentos mas relevantes en este caso,
#podemos decir que serian el momento L3 (Asimetria), ya que permite evaluar el sesgo
#de la distribucion, y el L4 (Curtosis), que describe la prominencia de valores extremos.
#Por lo tanto, estos momentos son los mas utiles para representar adecuadamente
#las caracteristicas de esta distribucion.
```