Datos cuantitativos agrupados

Ramon Ceballos

3/2/2021

1. Histogramas

La mejor manera de representar datos agrupados es mediante unos *diagramas de barras especiales* conocidos como **histogramas**.

En ellos se dibuja sobre cada clase una barra cuya área representa su frecuencia (las barras van pegadas unas a otras). Podéis comprobar que el producto de la base por la altura de cada barra es igual a la frecuencia de la clase correspondiente.

1.1 Uso de histogramas

Si todas las clases tienen la misma amplitud, las alturas de estas barras son proporcionales a las frecuencias de sus clases, con lo cual podemos marcar sin ningún problema las frecuencias sobre el eje vertical. Pero si las amplitudes de las clases no son iguales, las alturas de las barras en un histograma no representan correctamente las frecuencias de las clases.

En este último caso, las alturas de las barras son las necesarias para que el área de cada barra sea igual a la frecuencia de la clase correspondiente y como las bases son de amplitudes diferentes, estas alturas no son proporcionales a las frecuencias de las clases, por lo que no tiene sentido marcar las frecuencias en el eje vertical.

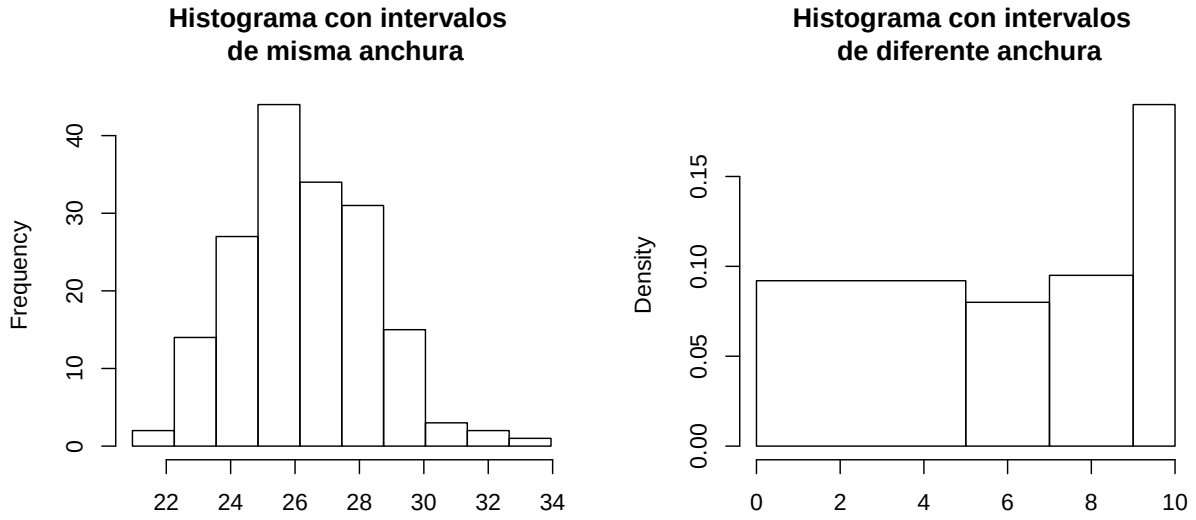
Los histogramas también son utilizados para representar frecuencias acumuladas de datos agrupados. En este caso, las alturas representan las frecuencias independientemente de la base debido a que éstas deben ir creciendo.

1.2 Interpretación de los histogramas

El eje de las abcisas representa los datos. Aquí marcamos los extremos de las clases y se dibuja una barra sobre cada una de ellas. Esta barra tiene significados diferentes en función del tipo de histograma, pero en general representa la *frecuencia de su clase*.

- **Histograma de frecuencias absolutas**: la altura de cada barra es la necesaria para que el área de la barra sea igual a la frecuencia absoluta de la clase. Las amplitudes de las clases pueden ser todas iguales o no. En el primer caso, las alturas son proporcionales a las frecuencias. En el segundo caso, no existe tal proporcionalidad. De todas formas, sea cual sea el caso, conviene indicar de alguna forma la frecuencia que representa cada barra.

```r
#Ejemplo de histogramas de frecuencias absolutas
crabs = read.table("../../../../data/datacrab.txt", header = TRUE)
cw = crabs$width
L1 = min(cw)-1/2*0.1
A=1.3
L = L1 + A*(0:10)

set.seed(144)
notas = sample(0:10,100, replace = TRUE)
set.seed(NULL)
Lnotas = c(0,5,7,9,10)

par(mfrow = c(1,2))

#Misma amplitud
hist(cw, breaks = L,
     right = FALSE,
     main = "Histograma con intervalos \n de misma anchura",
     xlab = "")

#Diferente amplitud
hist(notas, breaks = Lnotas,
     right = FALSE,
     include.lowest = TRUE,
     main = "Histograma con intervalos \n de diferente anchura",
     xlab = "")
```




```r
par(mfrow = c(1,1))
```

- **Histograma de frecuencias relativas**: la altura, **densidad**, de cada barra es la necesaria para que el área sea igual a la frecuencia relativa de la clase. La suma de todas las áreas debe ser 1. De nuevo, conviene indicar de alguna forma la frecuencia que representa cada barra.

- **Histogramas de frecuencias acumuladas**: las alturas de las barras son iguales a las frecuencias acumuladas de las clases, independientemente de su amplitud.

1.3 Frecuencias nulas

No es conveniente que en un histograma aparezcan clases con frecuencia nula, exceptuando el caso en que represente poblaciones muy diferentes y separadas sin individuos intermedios.

Si apareciesen clases vacías, convendría utilizar un número menor de clases, o bien unir las clases vacías con alguna de sus adyacentes. De este último modo romperíamos nuestro modo de trabajar con clases de la misma amplitud.

1.4 Dibujar histogramas en R

Lo hacemos con la función **`hist`**, la cual ya conocemos. Su sintaxis es

```
hist(x, breaks=..., freq=..., right=..., ...)
```

- **`x`**: vector de los datos.
- **`breaks`**: vector con los extremos de los intervalos o el número k de intervalos. Incluso podemos indicar, entre comillas, el método que deseemos para calcular el número de clases: `"Scott"`, `"Sturges"`... Eso sí, para cualquiera de las dos últimas opciones, no siempre obtendréis el número deseado de intervalos, puesto que R lo considerará solo como sugerencia. Además, recordad que el método para calcular los intervalos es diferente al de la función **`cut`**. Por tanto, se recomienda hacer uso de la primera opción (vector con los extremos de los intervalos).
- **`freq=TRUE`**, que es su valor por defecto, produce el histograma de frecuencias absolutas si los intervalos son todos de la misma amplitud y de frecuencias relativas en caso contrario. **`freq=FALSE`** nos produce siempre el de frecuencias relativas.
- **`right`** funciona exactamente igual que en la función `cut`. Para cerrar o abrir el intervalo por la derecha.
- **`include.lowest = TRUE`** también funciona exactamente igual que en la función `cut`. Incluye el último valor como cerrado.
- También podéis utilizar los parámetros de la función **`plot`** que tengan sentido.

La instrucción **`hist`** titula por defecto los histogramas del siguiente modo: "Histogram of" seguido del nombre del vector de datos. No suele quedar muy bien si no estamos haciendo nuestro análisis en inglés.

Recordemos que el parámetro **`plot`** igualado a **`FALSE`** no dibujaba, pero sí calculaba el histograma. Se podrían acceder a las variables del histograma.

La función **`hist`** contiene mucha información en su estructura interna de los datos. De la información que nos presenta citar:

- **`breaks`** contiene el vector de extremos de los intervalos: $L_1, \ldots, L_{k+1}$
- **`mids`** contiene los puntos medios de los intervalos, lo que nosotros consideramos las marcas de clase: $X_1, \ldots, X_k$
- **`counts`** contiene el vector de frecuencias absolutas de los intervalos: $n_1, \ldots, n_k$
- **`density`** contiene el vector de las densidades de los intervalos. Estas se corresponden con las alturas de las barras del histograma de frecuencias relativas. Recordemos, la densidad de un intervalo es su frecuencia relativa divida por su amplitud.

1.4.1 Funciones para cálculo de histogramas de frecuencias absolutas

####**Histogramas de frecuencias absolutas**

Aquí os dejamos una función útil para calcular histogramas de frecuencias absolutas más completos. Se debe de facilitar a la función el vector de la variable y un vector que recoja los extremos de los intervalos.

```
histAbs = function(x,L) {
  h = hist(x, breaks = L,
           right = FALSE,
           freq = FALSE,
           xaxt = "n",
           yaxt = "n",
           col = "lightgray",
           main = "Histograma de frecuencias absolutas",
           xlab = "Intervalos y marcas de clase",
           ylab = "Frecuencias absolutas")

  axis(1, at=L)#En eje x se coloca las etiquetas de L

  text(h$mids, h$density/2, labels=h$counts, col="purple")
  }
```

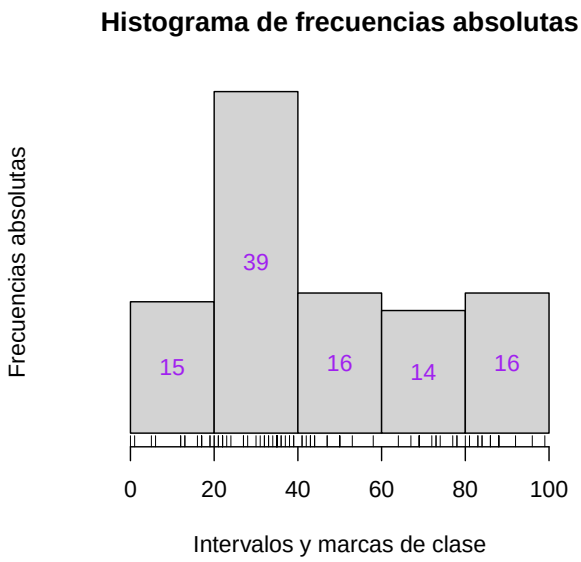
- **xaxt="n"** e **yaxt="n"** especifican que, por ahora, la función no dibuje los ejes de abcisas y ordenadas, respectivamente. La función **axis()** permite incorporar dichos ejes en las divisiones de L.
- **axis(i, at=...)** dibuja el eje correspondiente al valor de i con marcas en los lugares indicados por el vector definido mediante **at**. Si $i = 1$, el de abcisas; si $i = 2$, el de ordenadas.

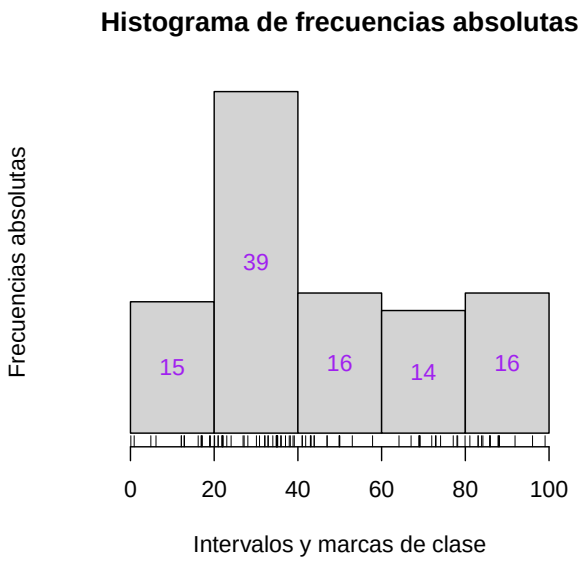
Os habréis fijado que con **freq = FALSE** en realidad hemos dibujado un histograma de frecuencias relativas, pero al haber omitido el eje de ordenadas, da lo mismo. En cambio, sí que nos ha sido útil para poder añadir, con la función **text**, la frecuencia absoluta de cada clase sobre el punto medio de su intervalo, los valores **h$mids** y a media algura de su barra, correspondiente a **h$density** gracias a que, con **freq = FALSE** estas alturas se corresponden con la densidad.

Otra forma de indicar las frecuencias absolutas de las barras es utilizar la función **rug**, la cual permite añadir al histograma una "alfombra" con marcas en todos los valores del vector, donde el grosor de cada marca es proporcional a la frecuencia del valor que representa.

Existe la posibilidad de añadir un poco de ruido a los datos de un vector para deshacer posibles empates. Esto lo conseguimos combinando la función **rug** con **jitter**.

```
set.seed(1)

edades = c(sample(0:99,80,replace = TRUE),rep(35,10),rep(22,5),rep(17,3),50,50)

extremos = c(0,20,40,60,80,100)

par(mfrow=c(1, 2))

#Ejemplo de la función del histograma
histAbs(edades, extremos)
rug(edades)
histAbs(edades, extremos)
rug(jitter(edades))
```






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

set.seed(NULL)
```

Histogramas de frecuencias absolutas acumuladas Aquí os dejamos una función útil para calcular histogramas de frecuencias absolutas acumuladas más completos:

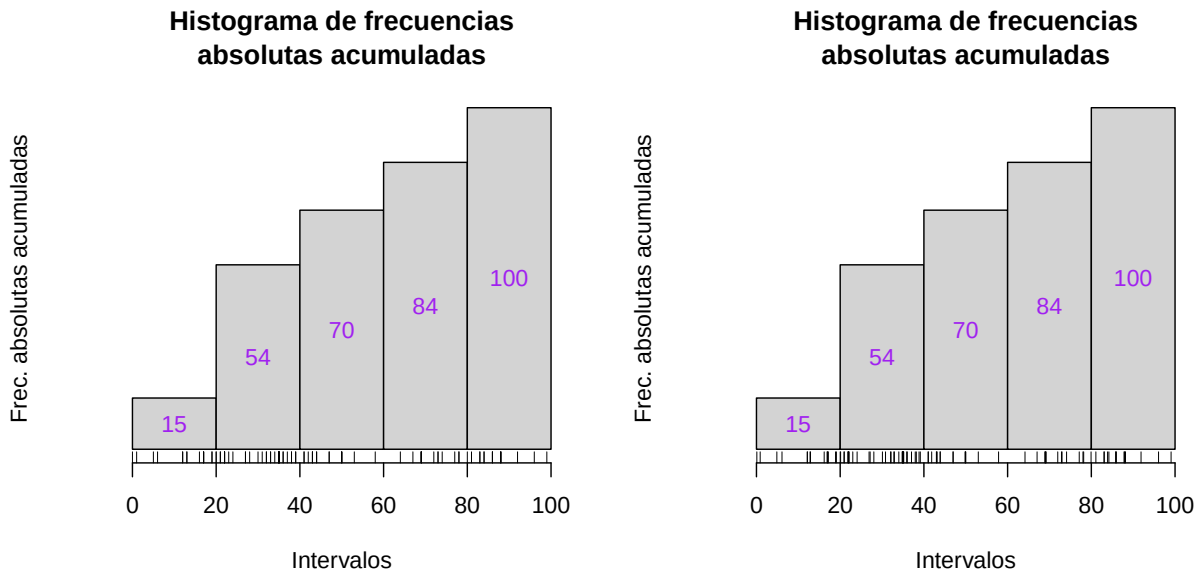
```
histAbsCum = function(x,L) {

  h = hist(x, breaks = L,
           right = FALSE ,
           plot = FALSE)

  #Las densidades serán las sumas acumuladas de las densidades
  #Será lo que se dibujara a posteriori
  h$density = cumsum(h$density)

  plot(h, freq = FALSE,
       xaxt = "n", yaxt = "n",
       col = "lightgray",
       main = "Histograma de frecuencias\nabsolutas acumuladas",
       xlab = "Intervalos",
       ylab = "Frec. absolutas acumuladas")

  axis(1, at=L)

  text(h$mids, h$density/2, labels = cumsum(h$counts), col = "purple")
  }
```

Con la función anterior, lo que hacemos es, en primer lugar, producir el histograma básico de los datos, sin dibujarlo para a continuación modificar la componente **`density`** para que contenga las sumas acumuladas de esta componente del histograma original.

Seguidamente, dibujamos el nuevo histograma resultante, aplicando la función **plot**. Es aquí donde debemos especificar los parámetros y no en el histograma original.

Finalmente, añadimos el eje de abcisas y las frecuencias acumuladas en color lila.

```
set.seed(1)

edades = c(sample(0:99,80,replace = TRUE),rep(35,10),
           rep(22,5),rep(17,3),50,50)
extremos = c(0,20,40,60,80,100)

#Histograma de la segunda función
par(mfrow=c(1, 2))
histAbsCum(edades, extremos)
rug(edades)
histAbsCum(edades, extremos)
rug(jitter(edades))
```




```
par(mfrow=c(1,1))
set.seed(NULL)
```

1.5 Histogramas de frecuencias relativas

En estos histogramas, es común superponer una curva que estime la densidad de la distribución de la variable cuantitativa definida por la característica que estamos midiendo.

La **densidad** de una variable es una curva cuya área comprendida entre el eje de las abcisas y la propia curva sobre un intervalo es igual a la fracción de individuos de la población que caen dentro de ese intervalo.

Para hacernos una idea visual, imaginad que vais aumentando el tamaño de la muestra a la vez que agrupáis los datos en un conjunto cada vez mayor de clases. Si el rango de los datos se mantiene constante, la amplitud de las clases del histograma irá menguando. Además, cuando n, el tamaño de la muestra, tiende a

infinito, los intervalos tienden a ser puntos y, a su vez, las barras tienden a ser líneas verticales. Pues bien, los extremos superiores de estas líneas serán los que dibujen la densidad de la variable.

1.5.1 Campana de Gauss

Es la densidad más famosa: la Campana de Gauss. Ésta se corresponde con una variable que siga una distribución nomal.

La forma de la campana depende de dos parámetros: el valor medio, μ, y su desviación típica, σ.

1.5.2 Dibujar la curva de densidad

Existen muchos métodos con los cuales estimar la densidad de distribución a partir de una muestra.

Una de ellas es mediante la función **`density`** de R. Al aplicarla a un conjunto de datos, produce una **`list`** que incluye los vectores **`x`** e **`y`** que continen la primera y segunda coordenadas, respectivamente, de 512 puntos de la forma (x, y) sobre la curva de densidad estimada. Es decir, se emplea una especie de método de interpolación para que se calcule una curva continua a partir de los datos originales.

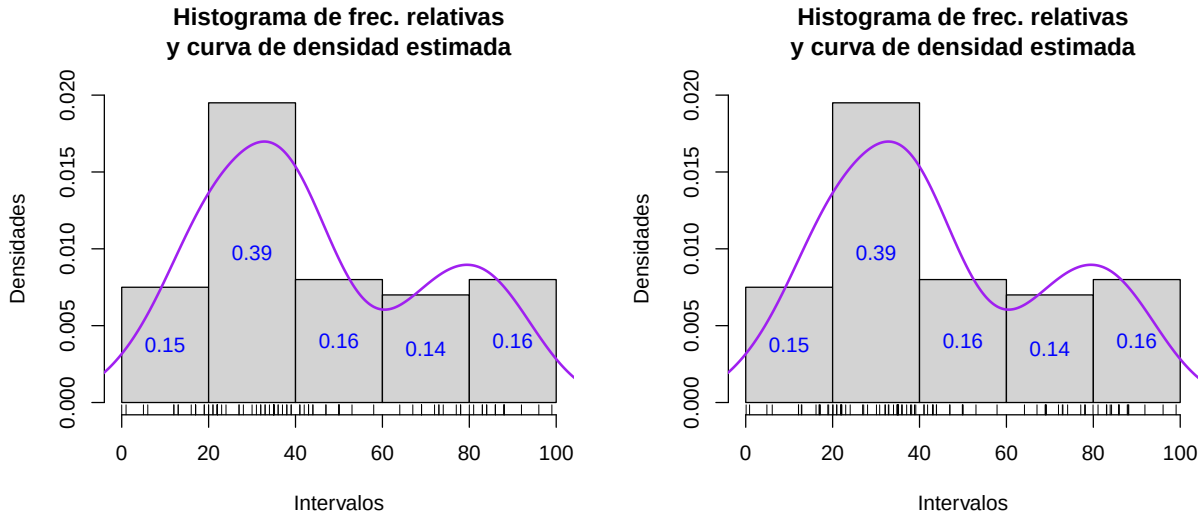
Aplicando **`plot`** o **`lines`** a este resultado según pertoque, obtenemos la representación gráfica de esta curva.

1.5.3 Funciones para cálculo de histogramas de frecuencias relativas

Función para calcular histogramas de frecuencias relativas Aquí os dejamos una función útil para calcular histogramas de frecuencias relativas más completos. Hay que suministrar la varibales de estudio y el vector de los extremos de los intervalos.

```
histRel = function(x,L) {
  h = hist(x, breaks=L, right=FALSE , plot=FALSE)

  #Para ajustar el eje de las "y" un 10% por encima del valor máximo
  t = round(1.1*max(max(density(x)[[2]]),h$density),2)

  plot(h, freq = FALSE,
       col = "lightgray",
       main = "Histograma de frec. relativas\ny curva de densidad estimada",
       xaxt="n", #Quita el eje "x"
       ylim=c(0,t),
       xlab="Intervalos",
       ylab="Densidades")

  axis(1, at = L) #eje "x" según L

  text(h$mids, #centro de las barras
       h$density/2, #altura media de cada barra
       labels = round(h$counts/length(x),2), #Valor F.Rel.
       col = "blue")

  #Pintar la función de densidad
  lines(density(x), col = "purple", lwd = 2)
  }
```

```r
set.seed(1)

edades = c(sample(0:99,80,replace = TRUE),rep(35,10),rep(22,5),rep(17,3),50,50)
extremos = c(0,20,40,60,80,100)

par(mfrow=c(1, 2))

#Histograma con la función anterior
histRel(edades, extremos)
rug(edades)
histRel(edades, extremos)
rug(jitter(edades))
```




```r
par(mfrow=c(1,1))
set.seed(NULL)
```

En este último tipo de histograma, se suele superponer una curva que estime la **función de distribución** de la variable definida por la característica que estamos midiendo.

Esta función de distribución, en cada punto nos da la fracción de individuos de la población que caen a la izquierda de este punto: su frecuencia relativa acumulada.

En general, la función de distribución en un valor determinado se obtiene hallando el área de la función de densidad que hay a la izquierda del valor.

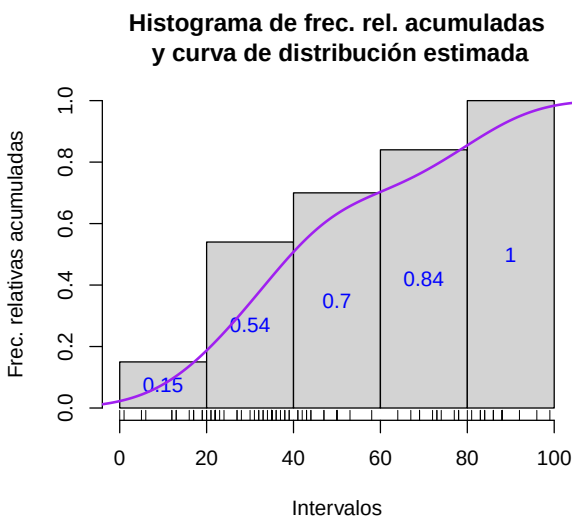
Por ello en muchas ocasiones, la función de distribución se defina como la *integral* de los valores menores o iguales que el que se está calculando.

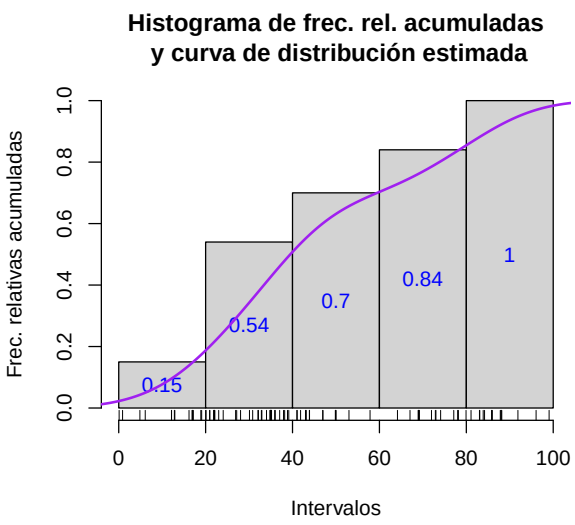
Función para calcular histogramas de frecuencias relativas acumuladas Aquí os dejamos una función útil para calcular histogramas de frecuencias relativas acumuladas más completos.

```r
histRelCum = function(x,L){
  h = hist(x, breaks = L,
```

```
          right = FALSE ,
          plot = FALSE)

  #FREC. REL ACUM.
   h$density = cumsum(h$counts)/length(x)

  plot(h, freq = FALSE,
      main = "Histograma de frec. rel. acumuladas\n y curva de distribución estimada",
      xaxt = "n",
      col = "lightgray",
      xlab = "Intervalos",
      ylab = "Frec. relativas acumuladas")

  axis(1, at = L)

  text(h$mids,
       h$density/2,
       labels = round(h$density ,2),
       col = "blue")

  #Calculo de la densidad de x
  #Igual que la curva representada ateriormente
  dens.x = density(x)

  #Se va acumulando el valor de la densidad de x
  #Se cambia el parámetro "y" para que se represente por las sumas acum.
  #Se conoce como función de distribución acumulada
  dens.x$y = cumsum(dens.x$y)*(dens.x$x[2]-dens.x$x[1])

  lines(dens.x,col = "purple",lwd = 2)
}
```

```
set.seed(1)

edades = c(sample(0:99,80,replace = TRUE),rep(35,10),rep(22,5),rep(17,3),50,50)
extremos = c(0,20,40,60,80,100)

par(mfrow=c(1, 2))

#Histograma Frec. Rel. Acum.
histRelCum(edades, extremos)
rug(edades)
histRelCum(edades, extremos)
rug(jitter(edades))
```






```
par(mfrow=c(1,1))
set.seed(NULL)
```