Histogramas

Santiago Pérez Moncada

15/7/2020

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. Podemos comprobar que el producto de la base por la altura de cada barra es igual a la frecuencia de la clase correspondiente.

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 sebre 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 ultimo 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.

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

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 aparecen clases vacías, convendria 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.

Dibujando histogramas con 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 obtienes el numero deseado de intervalos, puesto que R lo considerará solo como sugerencia.Ademas, recuerda que el método para calcular los intervalos es diferente al de la función cut, Por tanto, se recomienda hacer uso de la primera opción.
- **freq = TRUE** que su valor por defecto, produce el histograma de frecuencias absolutas si los intervalos son todos de la misma amplitud y el de frecuencias relativas en caso contrario, `freq=FALSE` nos produce el de frecuencias relativas.
- **right** funciona exactamente igual que en la función `cut`.
- **include.lowest = TRUE** también funciona exactamente igual que en la función `cut`.
- Tambien podemos utilizar los parámetros de la función `plot` que tengan sentido.

`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 dibuja, pero sí calculaba el histograma.

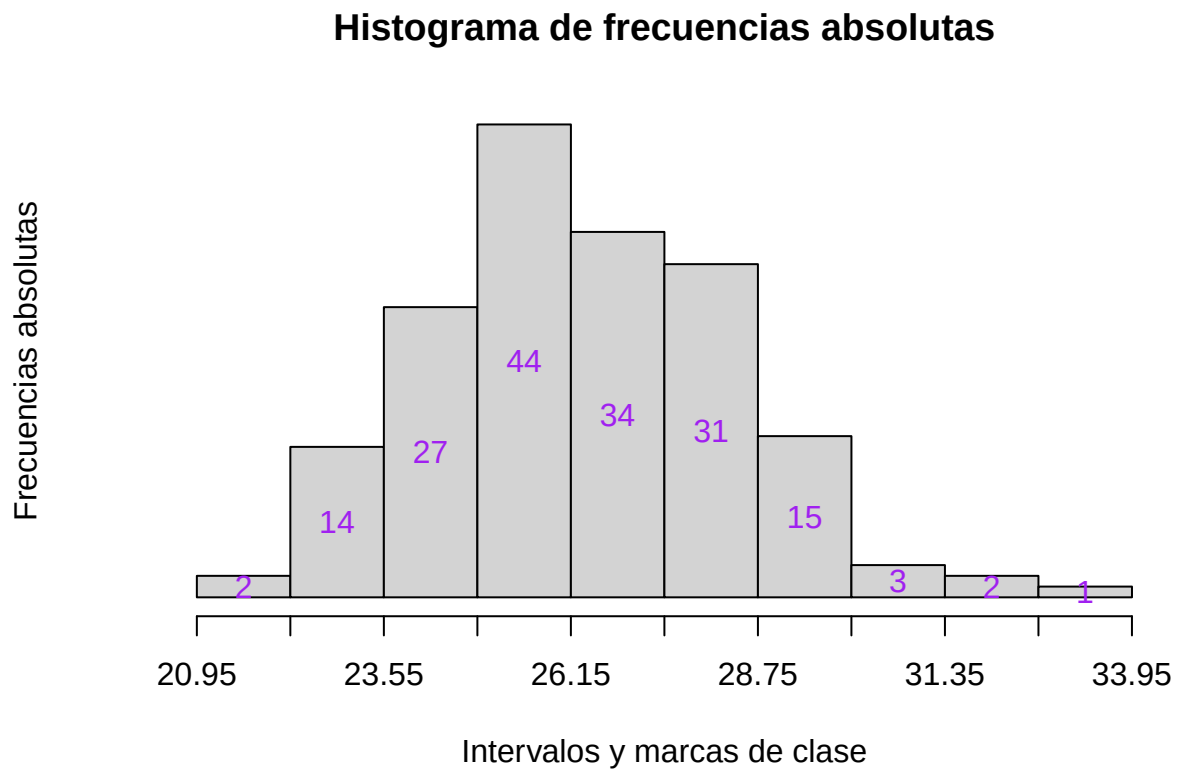
La función `hist` contiene mucha informacion en su estructura interna.

- **breaks** contiene el vector de los extremos de los intervalos $L_1, ..., L_{k+1}$
- **mids** contiene los puntos medios de los intervalos, lo que nosotros consideramos las marcas de clase: $X_1, ...X_k$
- **counts** contiene el vector de frecuencias absolutas de los intervalos: $n_1, ..., n_k$
- **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 dividida por su amplitud.

Aquí tenemos una función para calcular histogramas de frecuencias absolutas más completos:

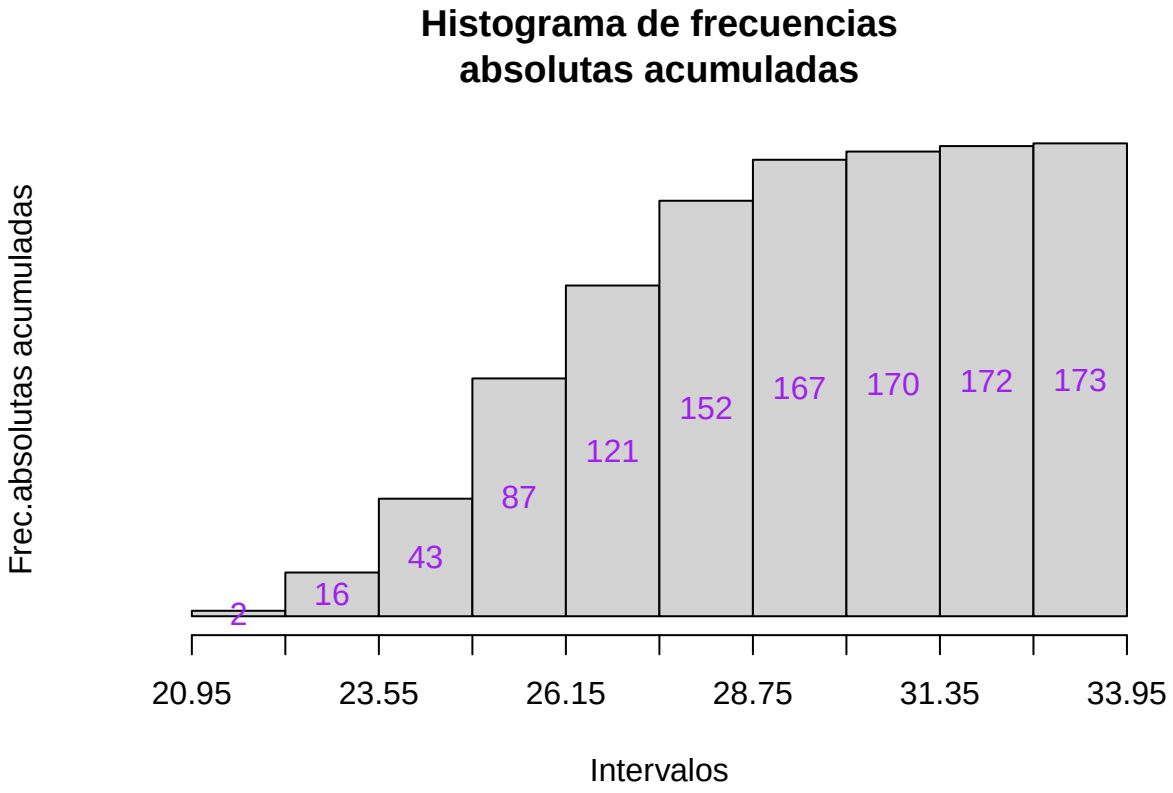
```
crabs = read.table("../data/datacrab.txt", header = TRUE)
cw = crabs$width
K_Scott = nclass.scott(cw) # 10
A = (max(cw)-min(cw))/K_Scott
A = 1.3
L1 = min(cw)-(1/2)*0.1
L = L1 + A*(0:10)

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

histAbs(cw,L)
```



Aqui 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)
  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" )
}

histAbsCum(cw,L)
```



Densidades y funciones de distribución

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.

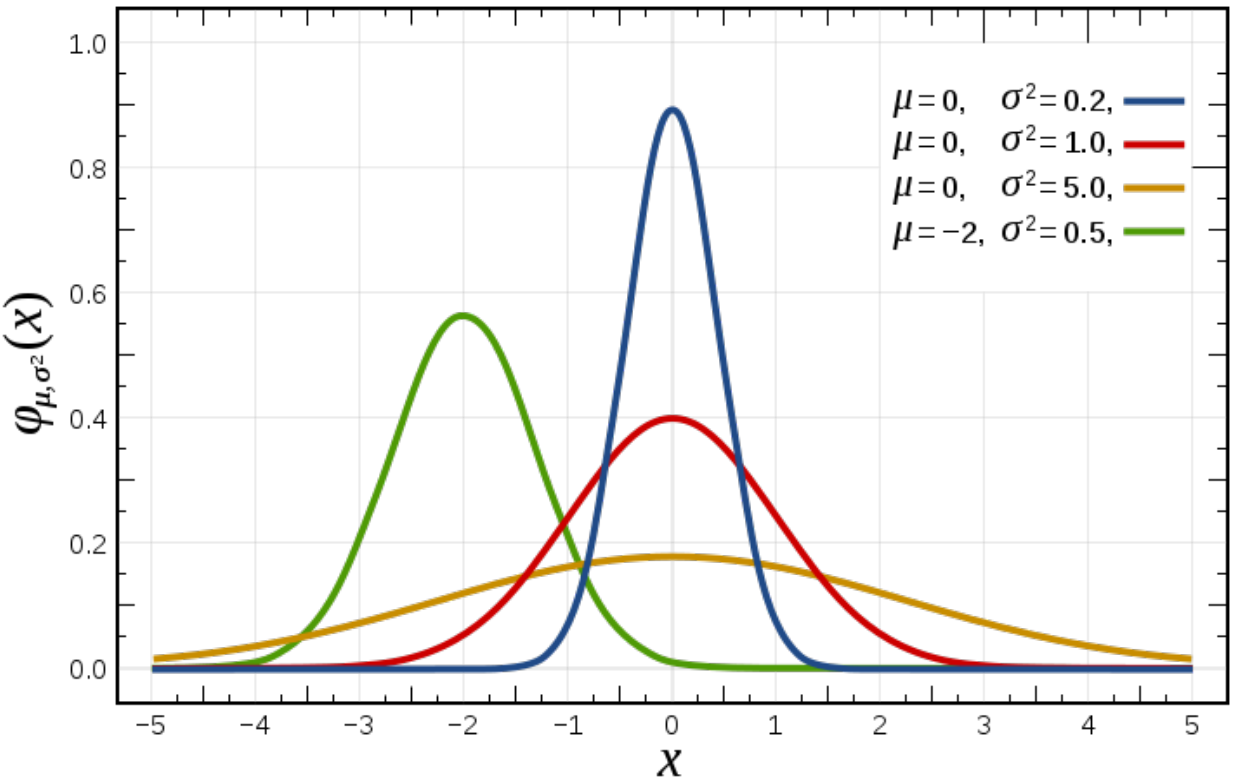
La **densidad** deuna variable es una curva cuya área comprendida entre el eje de las abcisas y la propia curva sobre un intervalo es igual a la fracción de los individuos de la población que caen dentro de ese intervalo.

Para hacernos una idea visual, imagina que vaz aumentando el tamaño de la muestra a la que estas agrupando los datos en un conjunto cada vez de mayor 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 vertivales. Pues bien, los extremos superiores de estas lineas serán los que dibujen la densidad de la variable.

Campana de Gauss

Es la densidad más famosa: La campana de gauss estase corresponde con una variable que siga una distribución normal.

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



Existen muchos métodos con los cuales se puede 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 contienen la primera y segunda coordenadas, respectivamente, de 512 puntos de la forma (x, y) sobre la curva de densidad estimada.

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

Aquí tenemos una función útiul para calcular histogramas de frecuencias relativas más completos.

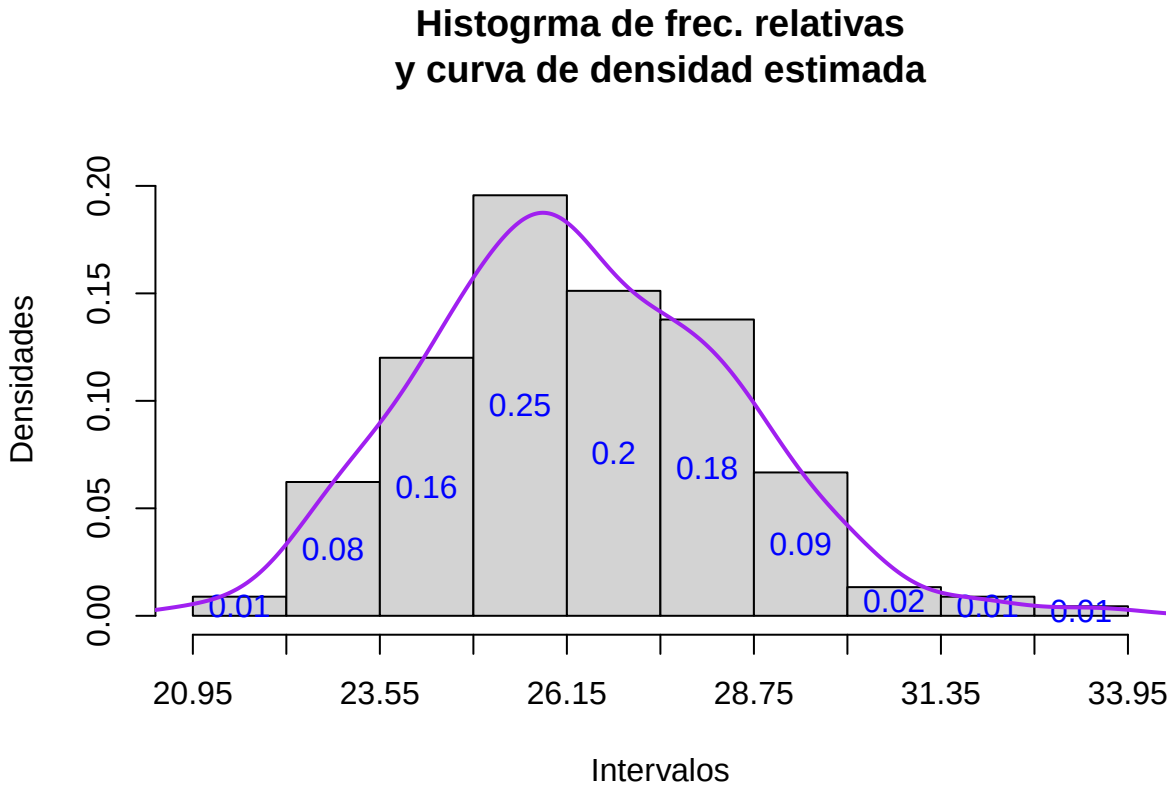
```
histRel <- function(x,L) {
  h =  hist(x, breaks = L, right = FALSE, plot=FALSE)
  t = round(1.1*max(max(density(x)[[2]]),h$density),2)
  plot(h,freq=FALSE,col="lightgray",
       main="Histogrma de frec. relativas\ny curva de densidad estimada",
       xaxt = "n", ylim = c(0,t), xlab = "Intervalos", ylab = "Densidades")
  axis(1,at = L)
  text(h$mids,h$density/2, labels = round(h$counts/length(x),2), col = "blue")
  lines(density(x),col="purple",lwd=2)
}
```

En este último tipo de histograma, se suele superponer una curva que estime la **función de densidad** 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 poblacion que caen a la izquierda de este punto: su frecuencia relativa acumulada.

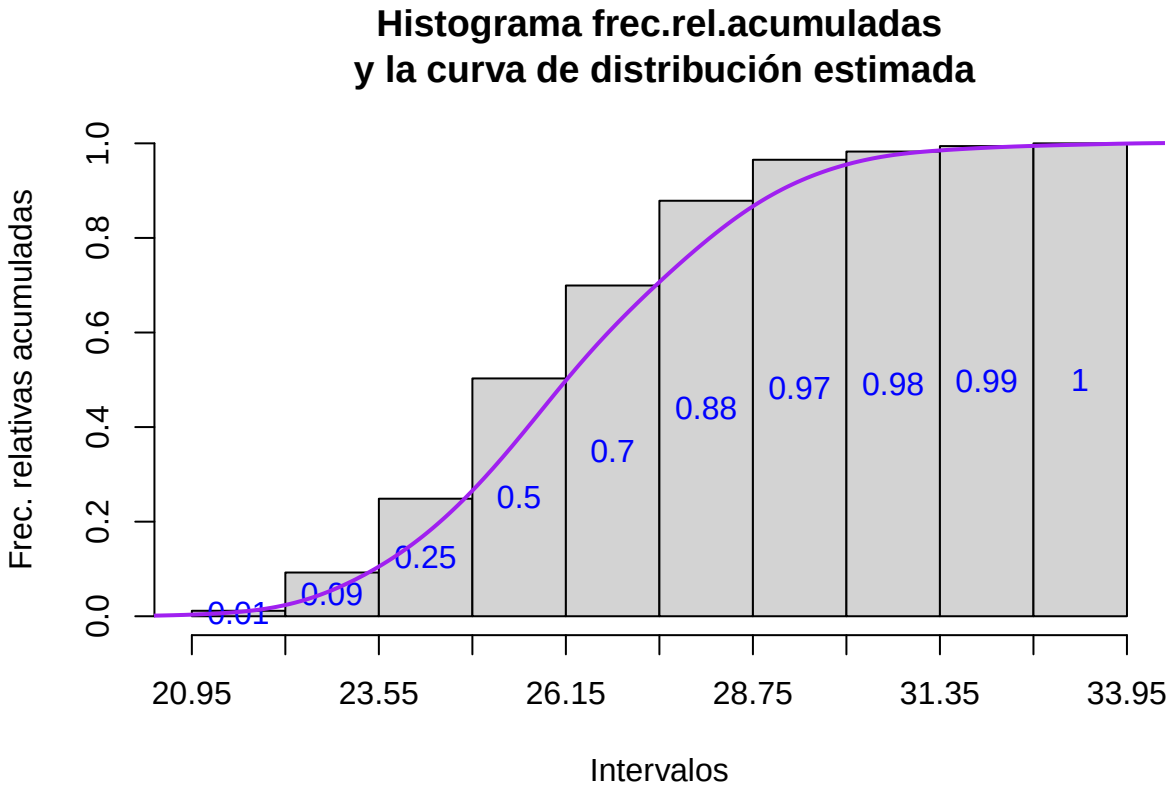
En general, la función de distribución en una valor determinado se obtiene hallando el área de la función de densidad que hay a la izquierda del valor.

```
histRel(cw,L)
```



Aqui dejamos una función útil para calcular histogramas de frecuencias relativas acumuladas más completos.

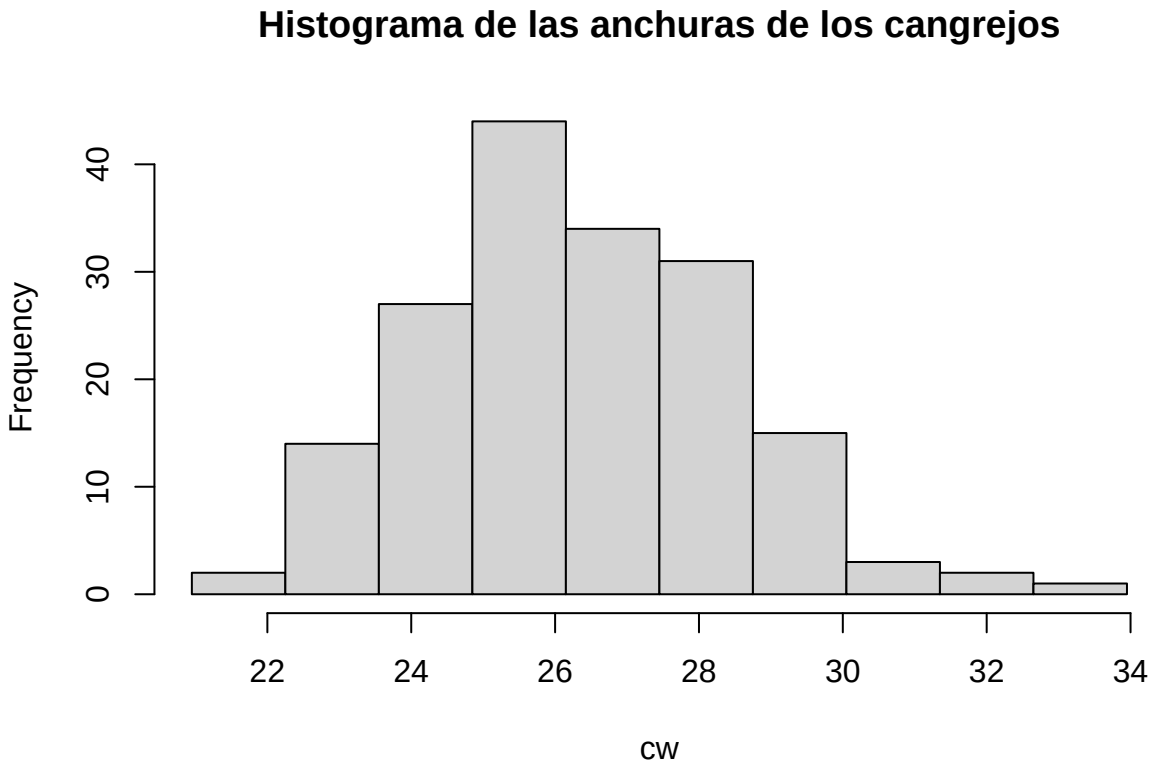
```
histRelCum = function(x,L){
  h = hist(x,breaks = L, right = FALSE, plot=FALSE)
  h$density = cumsum(h$counts)/length(x)
  plot(h, freq=FALSE,
       main = "Histograma frec.rel.acumuladas\n y la curva de distribución estimada",
       xaxt = "n", col = "lightgray", xlab = "Intervalos",
       ylab = "Frec. relativas acumuladas")
  axis(1, at = L)
  text(h$mids, h$density/2, labels = round(h$density,2), col = "blue")
  dens.x = density(x)
  dens.x$y = cumsum(dens.x$y)*(dens.x$x[2]-dens.x$x[1])
  lines(dens.x,col="purple",lwd=2)
}

histRelCum(cw,L)
```



Ejemplo

Vamos a seguir trabajando con nuestra variable `cw` y esta vez, lo que haremos será calcular histogramas de todas las formas explicadas anteriormente.

```
crabs = read.table("../data/datacrab.txt", header = TRUE)
cw = crabs$width
K_Scott = nclass.scott(cw) # 10
A = (max(cw)-min(cw))/K_Scott
A = 1.3
L1 = min(cw)-(1/2)*0.1
L = L1 + A*(0:10)

hist(cw, breaks = L, right = FALSE, main = "Histograma de las anchuras de los cangrejos")
```

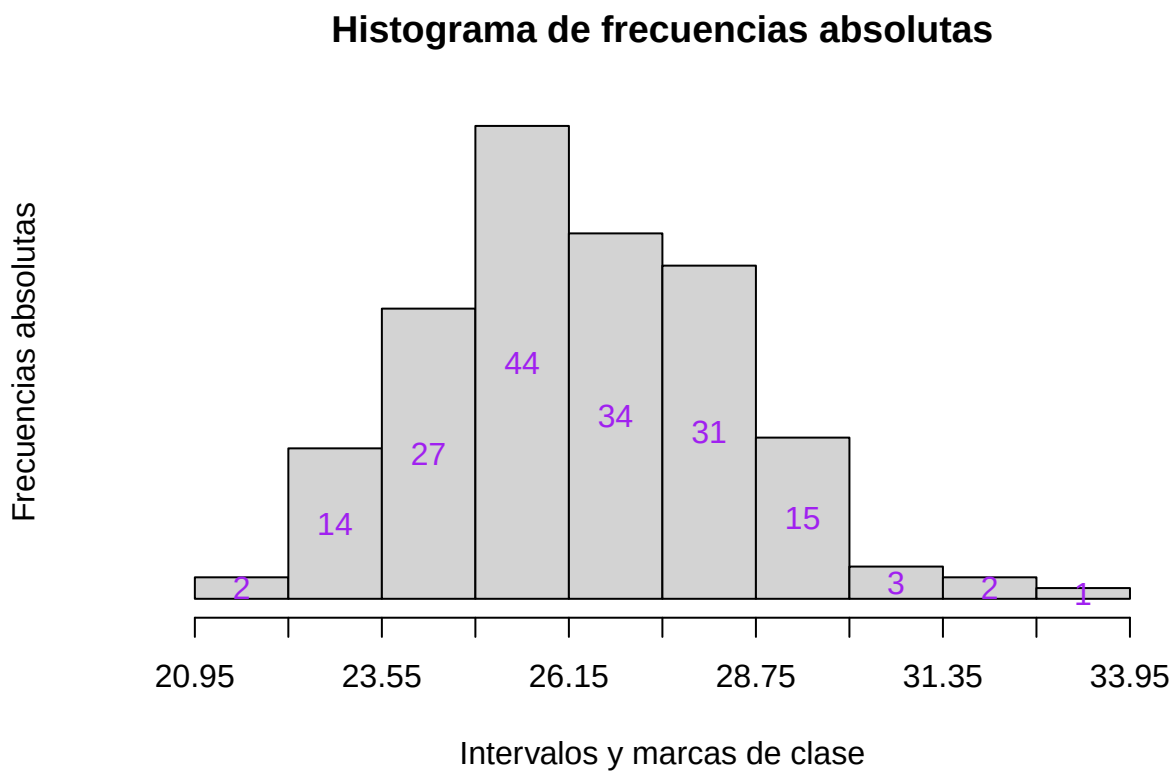



```
hist(cw, breaks = L, right = FALSE, plot = FALSE)
```

```
## $breaks
##  [1] 20.95 22.25 23.55 24.85 26.15 27.45 28.75 30.05 31.35 32.65 33.95
##
## $counts
##  [1]  2 14 27 44 34 31 15  3  2  1
##
## $density
##  [1] 0.008892841 0.062249889 0.120053357 0.195642508 0.151178301 0.137839040
##  [7] 0.066696309 0.013339262 0.008892841 0.004446421
##
## $mids
##  [1] 21.6 22.9 24.2 25.5 26.8 28.1 29.4 30.7 32.0 33.3
##
## $xname
## [1] "cw"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

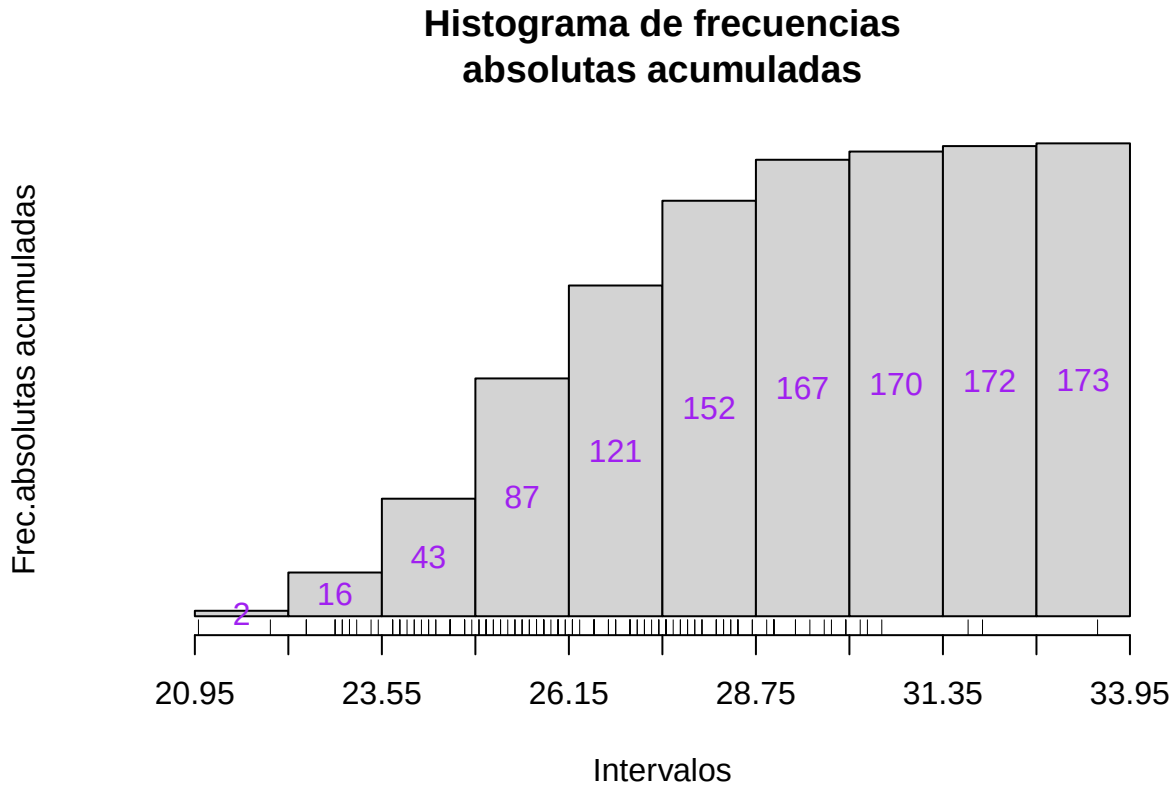
dibujamos el histograma con `histAbs`

```
histAbs(cw,L)
```



Dibujamos el histograma con `histAbsCum`

```
histAbsCum(cw,L)
rug(cw)
```

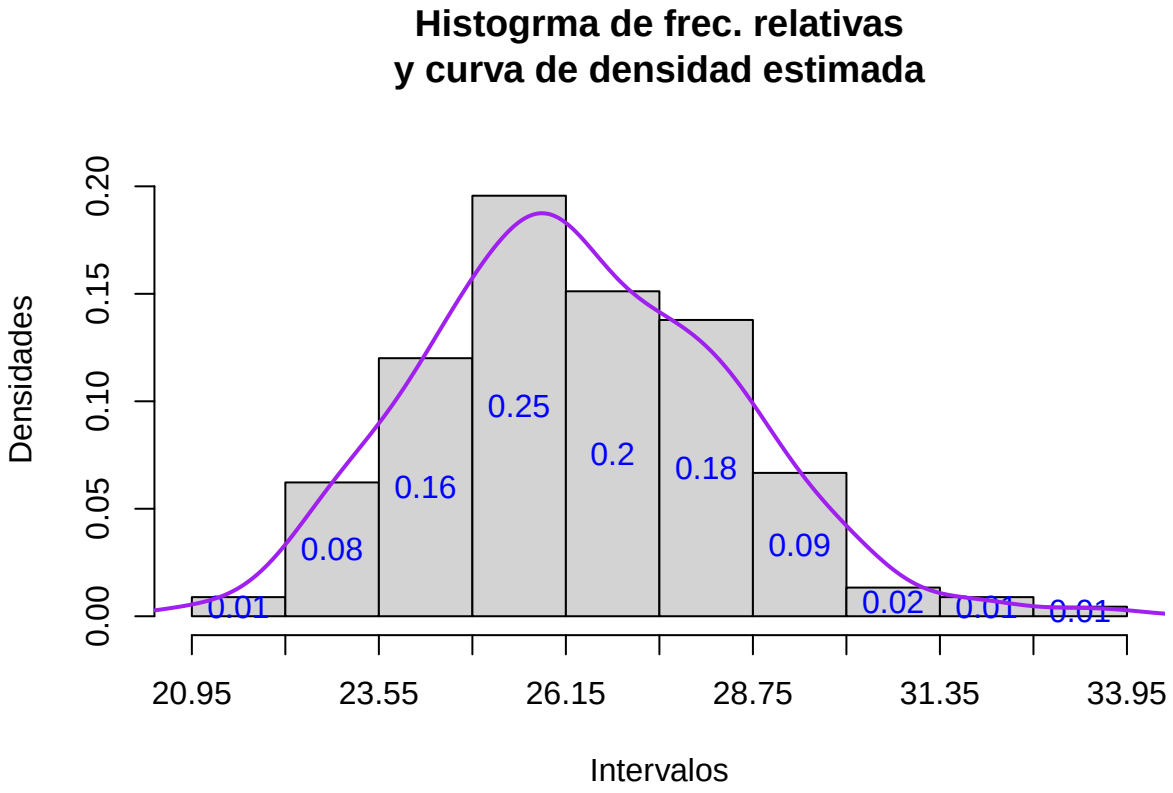


A continuación calculamos la densidad de `cw` y la representamos con `histRel`

```
str(density(cw))
```

```
## List of 7
##  $ x        : num [1:512] 19 19 19.1 19.1 19.1 ...
##  $ y        : num [1:512] 3.90e-05 4.50e-05 5.17e-05 5.94e-05 6.82e-05 ...
##  $ bw       : num 0.671
##  $ n        : int 173
##  $ call     : language density.default(x = cw)
##  $ data.name: chr "cw"
##  $ has.na   : logi FALSE
##  - attr(*, "class")= chr "density"
```

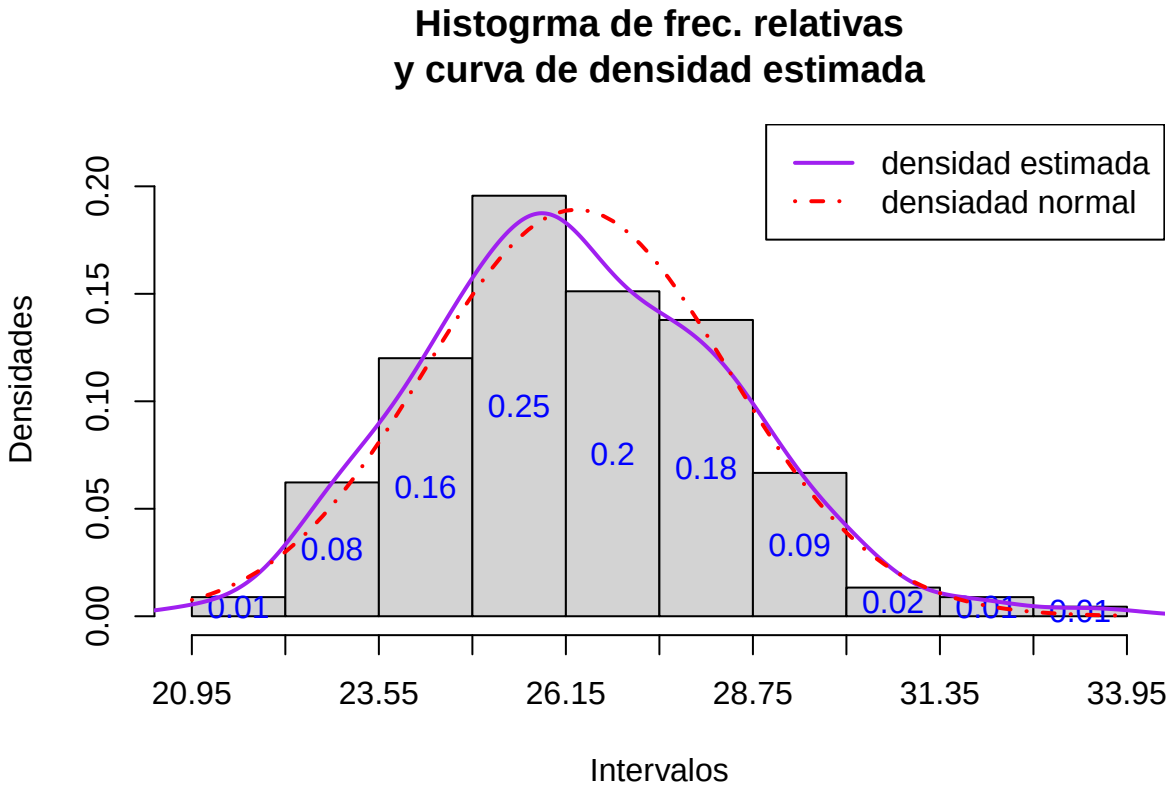
```
histRel(cw,L)
```



La curva de densidad que hemos obtenido en este gráfico tiene una forma de campana que nos recuerda la campana de Gauss. Para explorar este parecido, vamos a añadir al histograma la gráfica de la función de densidad de una distribución normal de media y desviación típica las del conjunto de datos original.

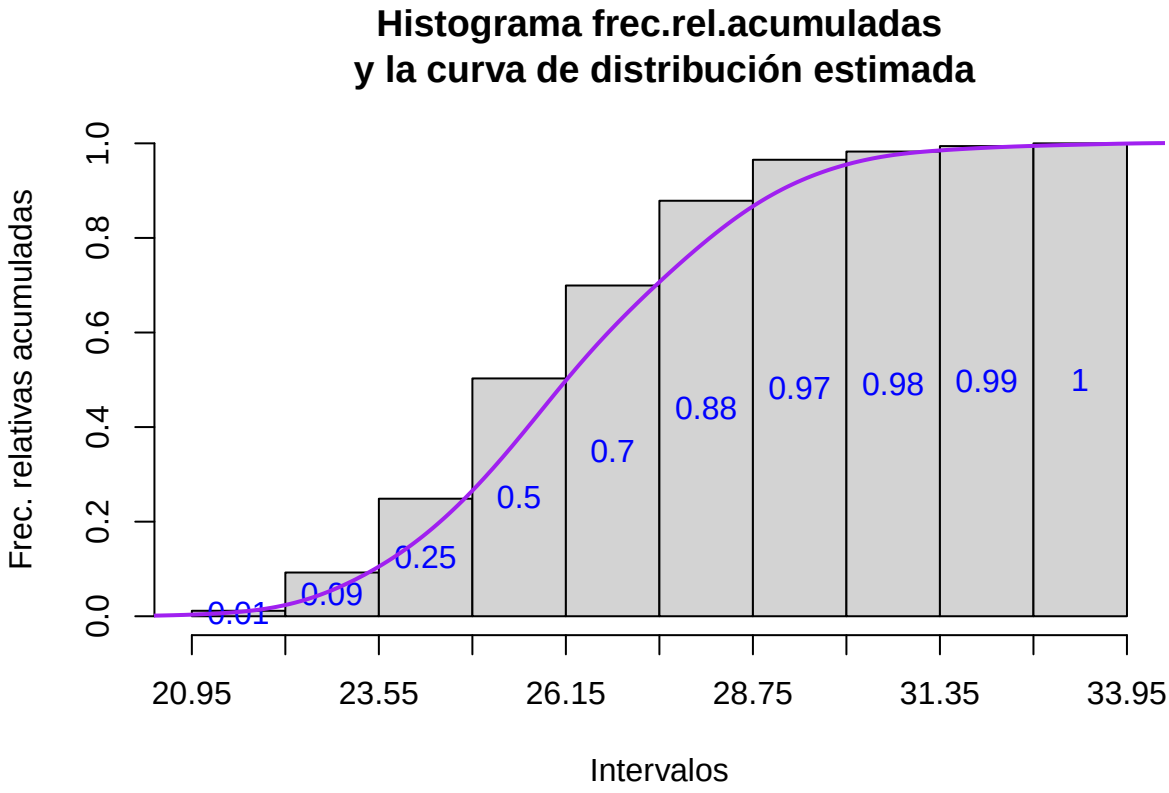
Así, aplicando las instrucciones siguientes, acabamos obteniendo.

```
histRel(cw,L)
curve(dnorm(x, mean(cw), sd(cw)), col="red", lty=4, lwd=2,
      add = TRUE)
legend("topright", lwd=c(2,2), lty=c(1,4), col = c("purple","red"),
       legend = c("densidad estimada","densiadad normal"))
```



Dibujamos el histograma con `histRelCum`

```
histRelCum(cw,L)
```


Histograma frec.rel.acumuladas
y la curva de distribución estimada
Frec. relativas acumuladas
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.09
0.25
0.5
0.7
0.88
0.97
0.98
0.99
1
20.95
23.55
26.15
28.75
31.35
33.95
Intervalos