Fundamentos de gráficos en R

Diego Andres Benitez

30/6/2020

Contents

1 Gráficos en R

1.1 Tallos y hojas

1.2 Histogramas

El histograma es un gráfico estadístico sobre un conjunto de datos de variable alaeatoria continua, en el cual a partir de barras se representan las frecuencias con las que aparecen mediciones agrupadas en rangos o intervalos. Para la construcción de un histograma se puede dividir el eje X en intervalos, que pueden ser de igual longitud, y luego contar el número de observaciones para cada intervalo. El número de barras o intervalos para la construcción de un histograma está dado por algunas reglas matemáticas, tales como la Regla de Sturges o la Regla de Scott.

La Regla de Sturges, considera un histograma que consta de k-intervalos, donde el i-ésimo intervalo contiene un determinado número de elementos ($i = 0, 1, \cdots, k-1$), representado como:

$$C_{k-1,i} = C_i^{k-1} = \binom{k-1}{i}$$

El número de elementos está determinado por el coeficiente binomial, expresado como:

$$C_{k-1,i} = \frac{(k-1)!}{i!(k-1)!}$$

$$n = \sum_{i=1}^{k-1} \binom{k-1}{i} = 1 + 1^{k-1} = 2^{k-1}$$

Para simplificar la ecuación, se pueden usar las propiedades de los logaritmos a ambos lados de la ecuación, teniendo así que:

$$log_2(n) = k - 1$$

Por lo cual, el número óptimo de intervalos de un histograma está dado por la expresión:

$$k = 1 + log_2(n)$$

Este método también puede emplearse con otros logaritmos, donde es necesario obtener una constante c que determine la equivalencia tal como se muestra a continuación con el logaritmo base 10

$$k = 1 + log_2(n)$$

Teniendo en cuenta las propiedades de los logaritmos se tiene que:

$$k = 1 + \frac{log_{10}(n)}{log_{10}(2)}$$

$$k = 1 + \frac{1}{log_{10}(2)} * log_{10}(n)$$

$$k = 1 + 3.321928 * log_{10}(n)$$

En las anteriores expresiones se tiene que:

- k es el número de clases
- n es el número de observaciones

La amplitud es la longitud de cada intervalo basado en la distribución de frecuencias de los datos; está se puede expresar como

$$a = \frac{Lim_{Sup} - Lim_{Inf}}{k}$$

Teniendo en cuenta lo anterior, los pasos para la construcción del histograma son:

A. Definir intervalos de igual longitud. B. Determinar la cantidad de observaciones por cada intervalo. C. Segmentar el eje x de acuerdo con los limites de cada intervalo. D. Graficar los rectangulos.

El lenguaje de programación estadística R permite realizar este tipo de gráficos de manera muy sencilla.

1.2.1 Elementos del histograma

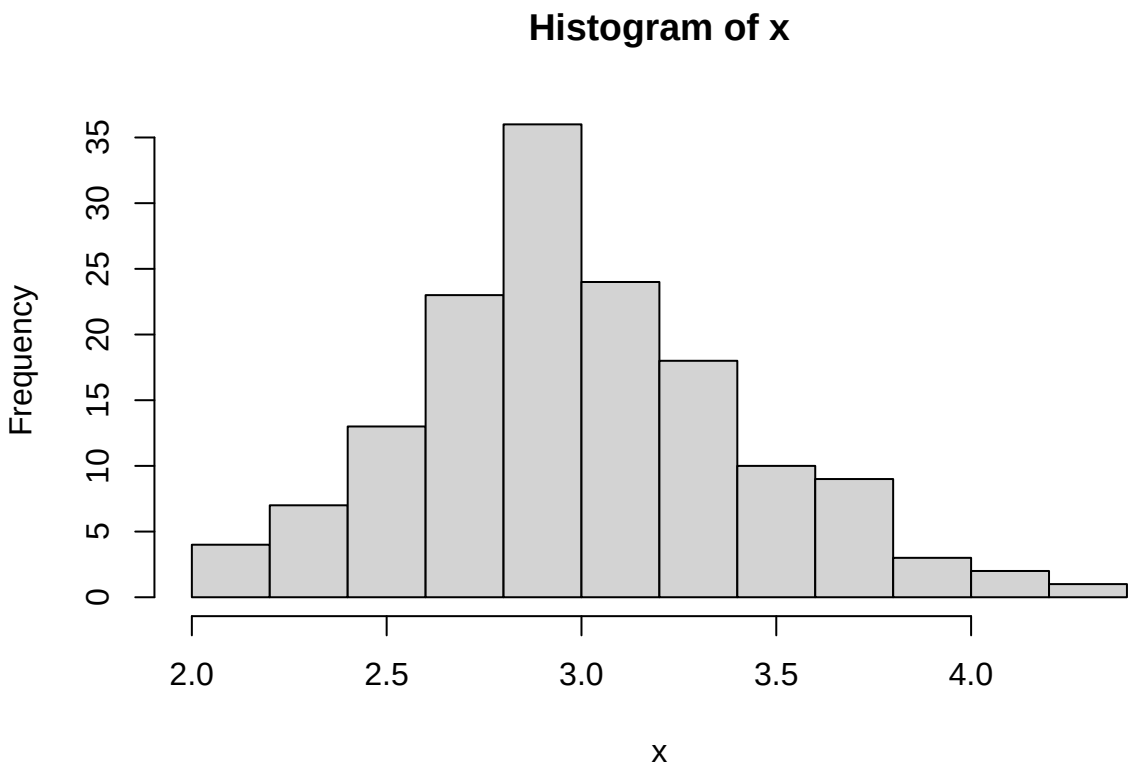
La función *hist()* es la encargada de gráficar los datos de acuerdo a su distribución de frecuencias. Los elementos de está función se listaran más adelante.

Para el trabajo gráfico con histogramas, se empleará la base de datos *iris*, la cual consta de mediciones de largo y ancho de sépalo y pétalo de flores de tres especies del género Iris. La variable a trabajar en el ejercicio gráfico será el ancho de sépalo de las flores del género mencionado.

```
x <- iris$Sepal.Width
```

El histograma del ancho de sépalo de las flores del género Iris sería:

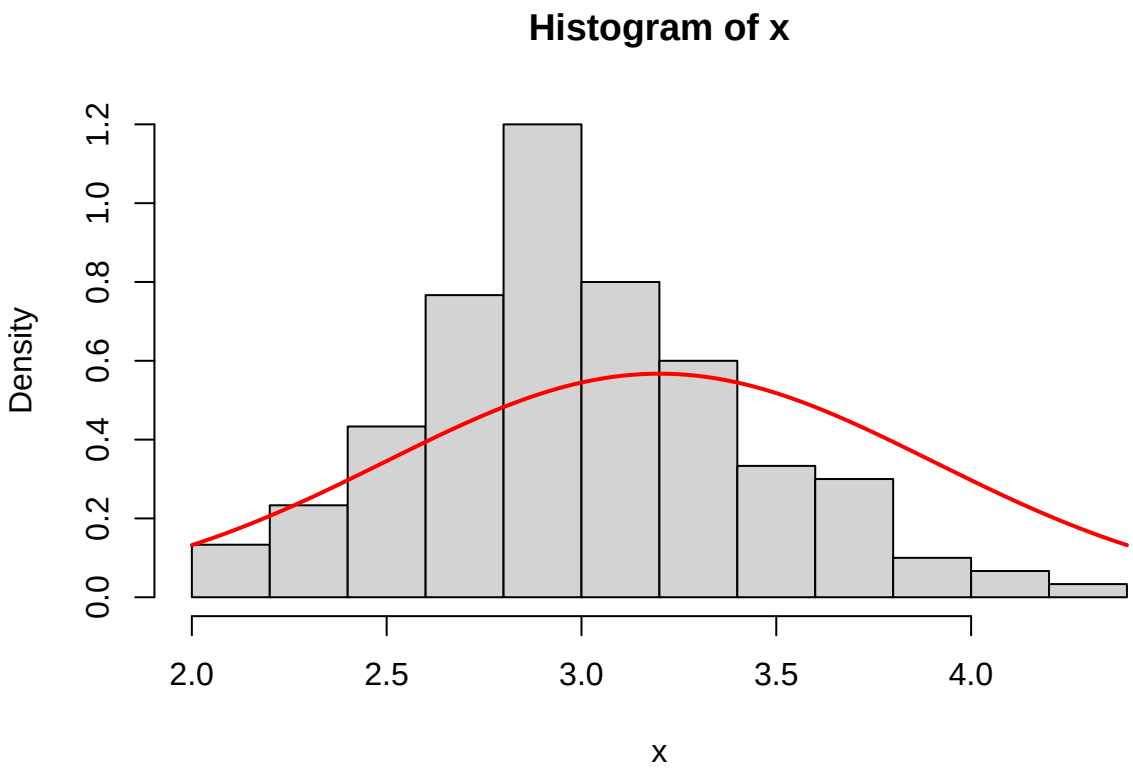
```
hist(x)
```



La función *hist()* solo requiere un vector de datos de entrada, generando una salida gráfica sencilla con todos los parámetros de la función por defecto. Como se puede notar en el anterior gráfico, los ejes no tienen un ajuste adecuado, los títulos por defecto se encuentran en inglés y el gráfico se encuentra en blanco y negro. Además, el nombre del eje x, así como el título, corresponden a el nombre de la variable, que para nuestro caso era x.

El gráfico por defecto se realiza sobre frecuencias absolutas, pero se puede realizar sobre frecuencias relativas de la siguiente manera:

```
hist(x, freq = F)
curve(dnorm(x, mean = mean(x), sd = sd(x)), add = T, col = "red", lwd = 2)
```



Notese que la escala del eje y cambia, cambiando la frecuencia por la densidad, es decir, cambiando la frecuencia absoluta por la frecuencia relativa, la cual se puede calcular de la siguiente manera

$$f_r = \frac{f_A}{n}$$

Y la altura del rectangulo corresponde a

$$h_\square = \frac{f_r}{a}$$

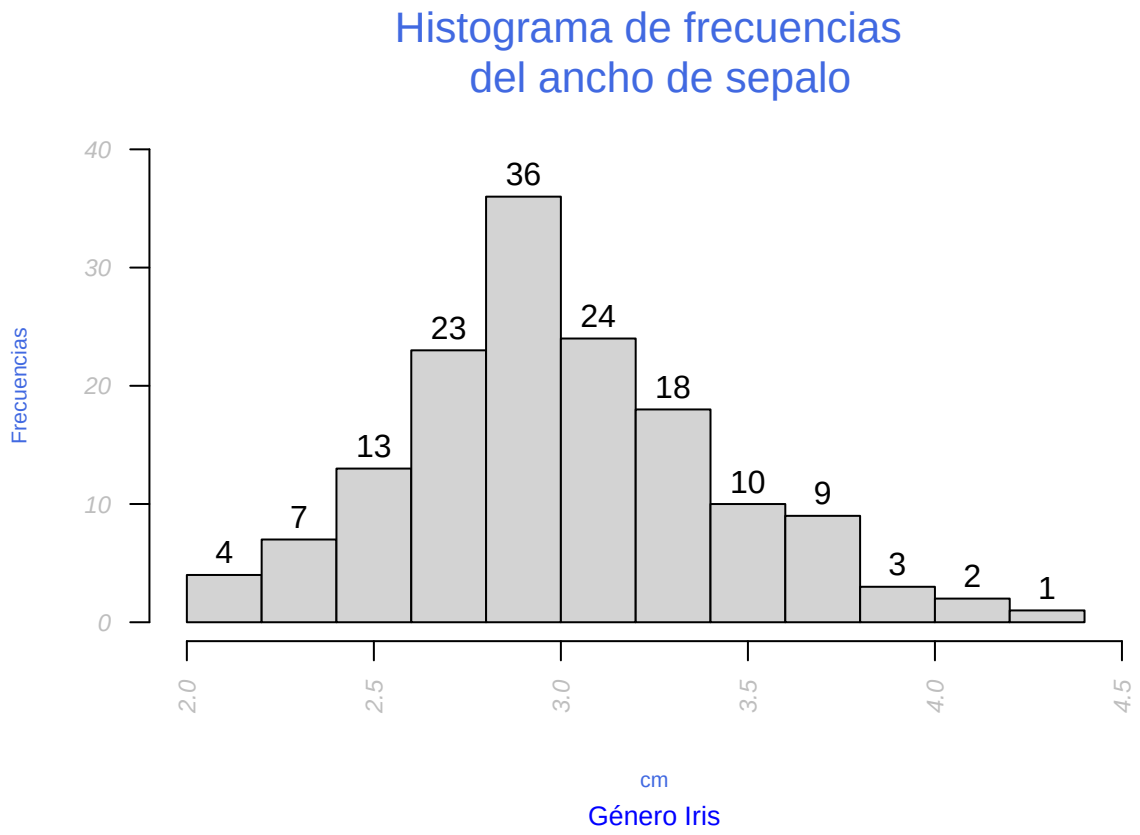
La selección entre frecuencias relativas y frecuencias absolutas dentro de la función *hist()* se puede realizar con el parámetro:

- **freq:** parámetro booleano que permite definir el tipo de frecuencias que se va a usar en el histograma, siendo **TRUE** la opción por defecto, es decir, gráficar frecuencias absolutas.

1.2.2 Títulos y etiquetas

Los títulos y etiquetas hacen referencia a los elementos de texto que acompañan el gráfico y proporcionan información del mismo.

```
#Títulos
hist(x, main = "Histograma de frecuencias \n del ancho de sepalo", xlab = "cm", ylab = "Frecuencias", su
```

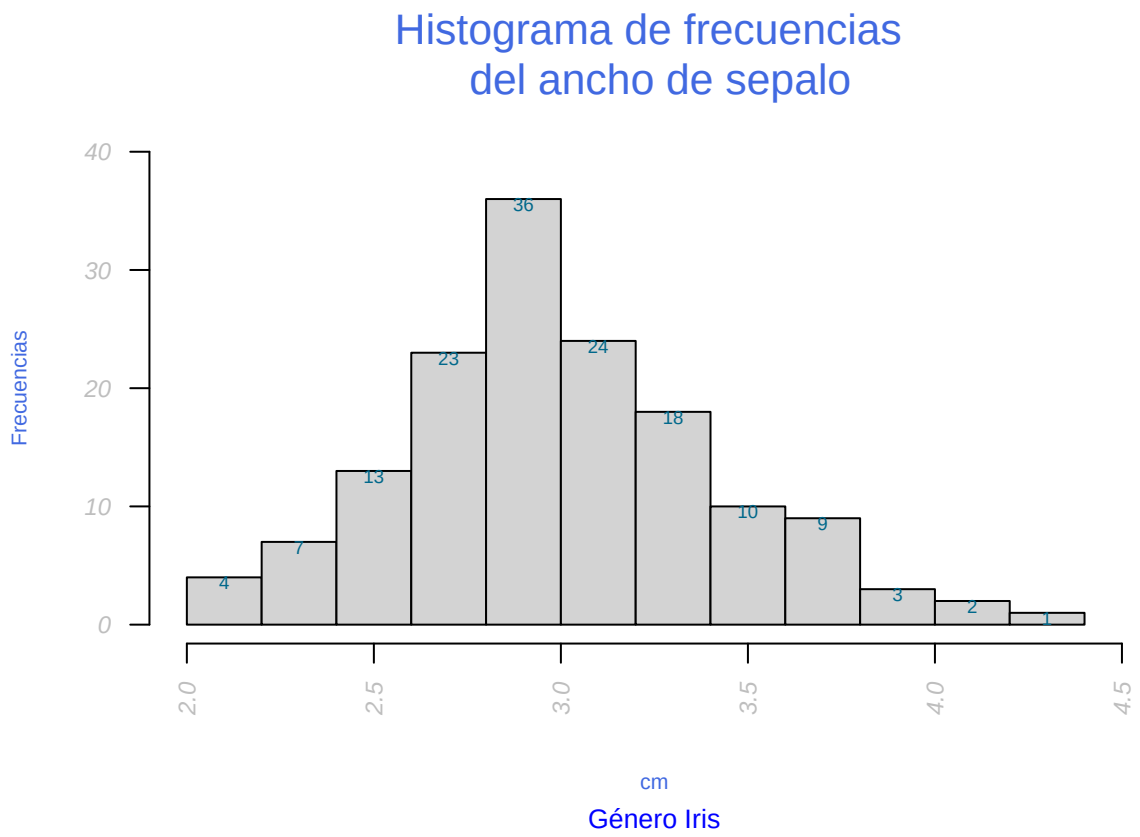


Los títulos en la función *hist()* constan de los siguientes parámetros:

- **main:** permite asignar el título al gráfico
- **sub:** permite asignar un subtítulo al gráfico
- **xlab:** permite asignar el rótulo de la variable del eje x
- **ylab:** permite asignar el rótulo de la variable del eje y
- **col:** permite dar color a cada uno de los títulos o rótulos, siendo *col.main*, *col.sub*, *col.lab* y *col.axis*. Los colores pueden definirse por nombre o por número, para indicar el color por su nombre se debe poner este entre comillas.
- **font:** permite asignar la fuente de los títulos o rótulos, siendo *font.main*, *font.sub*, *font.lab* y *font.axis*
- **cex:** permite asignar el tamaño de la fuente de los títulos o rótulos, siendo *cex.main*, *cex.sub*, *cex.lab* y *cex.axis*, estando en valores de 0 a ∞, donde 1 es el indicativo del 100% del tamaño
- **labels:** permite añadir la etiqueta de número de elementos por cada clase
- **las:** permite dar la orientación de los valores indicadores en cada eje, tomando valores de 0 a 3:
  - *0:* Es la salida por defecto y siempre situa los valores paralelos al gráfico
  - *1:* En esta salida se sitúan los valores horizontales al gráfico
  - *2:* En esta salida se sitúan los valores perpendiculares al gráfico
  - *3:* En esta salida se situan los valores verticales al gráfico

Otra forma de trabajar las etiquetas de tipo **labels** es asignando el gráfico a una nueva variable y aplicar a esta la función text, donde se definirán las propiedades de la etiqueta

```r
y <- hist(x, main = "Histograma de frecuencias \n del ancho de sepalo", xlab = "cm", ylab = "Frecuencias
text(y$mids, y$counts, labels = y$counts, adj = c(0.5, 0.9), cex = 0.6, col = "deepskyblue4")
```



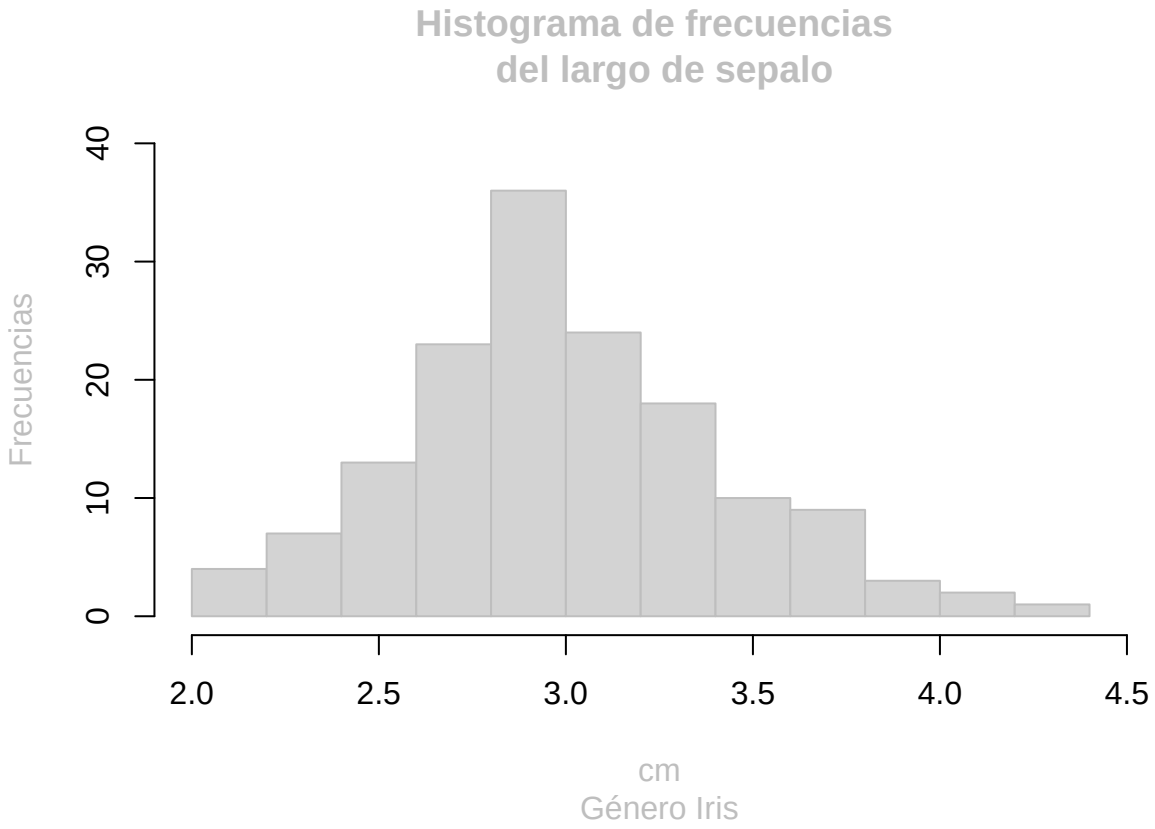
La función *text()* permite escribir sobre la gráfica realizada, añadiendo puntos de texto en las coordenadas definidas por el usuario. Los parámetros de la función usados son:

- **x:** Corresponde a las coordenadas en el eje x para cada una de las etiquetas. Para esté caso, se define como x el vector correspondiente a *y$mids* que contiene las marcas de clase del histograma.
- **y:** Corresponde a las coordenadas en el eje y para cada una de las etiquetas. Para esté caso, se define como y el vector correspondiente a *y$counts* que contiene la cantidad de elementos para cada clase.
- **adj:** Corresponde a la posición de la etiqueta alrededor de los puntos cartesianos ya fijados, donde 0.5 indica el punto de origen.

1.2.3 Manejo de ejes

Los ejes hacen referencia a las dimensiones del plano, donde estás se encuentran representadas de acuerdo con la escala de los datos.

```r
hist(x, main = "Histograma de frecuencias \n del largo de sepalo", xlab = "cm", ylab = "Frecuencias", su
```

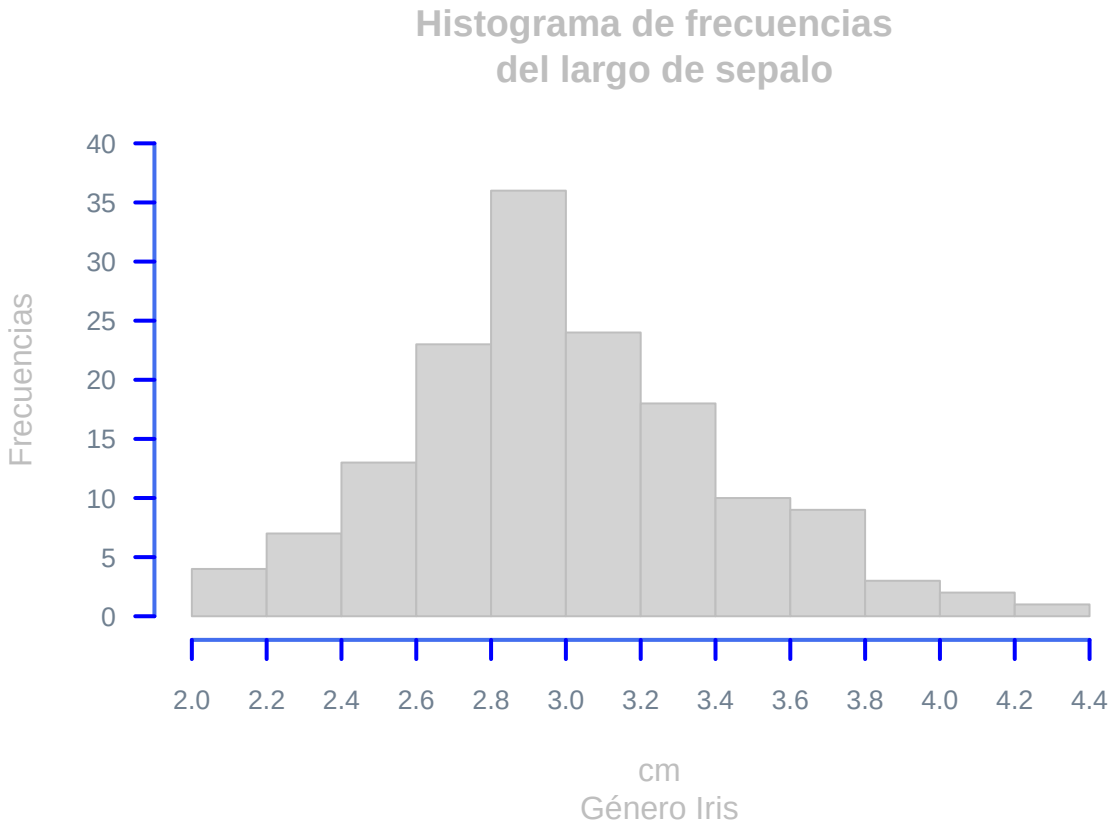


Los ejes en la función *hist()* constan de los siguientes parámetros:

- **xlim:** Establece la escala de valores en el gráfico a lo largo del eje x.
- **ylim:** Establece la escala de valores en el gráfico a lo largo del eje y.
- **axes:** Parámetro booleano que permite decidir si los ejes se muestran o no en el gráfico.

Otra forma de manipular los ejes consiste en usar la función auxiliar *axis()*, que dibuja estos en el gráfico, para lo cual es obligatorio eliminar los ejes por defecto en el gráfico usando el parámetro **axes = FALSE**.

```
hist(x,  main = "Histograma de frecuencias \n del largo de sepalo", xlab = "cm", ylab = "Frecuencias", s
axis(side = 1, at = seq(2.0, 4.4, 0.2), pos = -2, lwd = 2, lwd.ticks = 2, cex.axis = 0.8, tick = T, lty
axis(side = 2, at = seq(0, 40, 5), pos = 1.9 , lwd = 2, lwd.ticks = 2, cex.axis = 0.8, las = 1, tick = T
```



La función *axis()* constan de los siguientes parámetros:

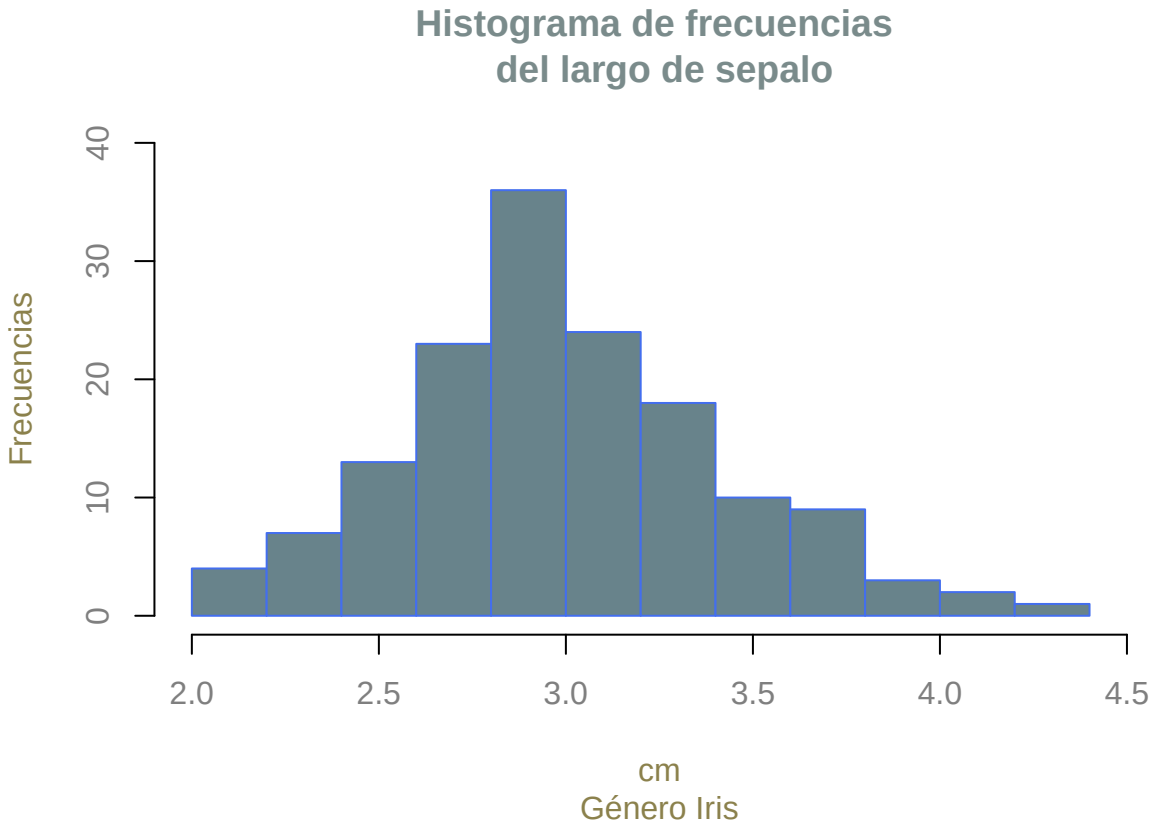
- **side:** Este parametro determina la ubicación del eje que se esta dibujando, permitiendo que ubicarlo en la parte inferior, la superior, a la izquierda o a la derecha del gráfico. Para esta labor, el parámetro puede tomar los valores:
  - **1:** El eje se ubica en la parte inferior.
  - **2:** El eje se ubica en la parte izquierda.
  - **3:** El eje se ubica en la parte derecha.
  - **4:** El eje se ubica en la parte superior.

  Es de aclarar que es necesario hacer un *axis()* por cada eje que se requiera, esto es debido a que solo se puede definir una posición en este parámetro.
- **at:** Define los valores en los cuales se va a añadir una marca en el eje, para lo cual se puede emplear la función *seq()* dentro del parametro para establecer la secuencia de los puntos de corte.
- **pos:** Sitúa al eje con base en una determinada posición cartesiana. Sin embargo, es importante tener en cuenta que la posición debe estar acorde a la escala de los datos.
- **lwd:** Permite definir el ancho de la linea del eje.
- **tick:** Es un parámetro de tipo booleano que permite decidir si se muestran o no la linea del eje en el gráfico.
- **col.ticks:** Permite definir el color de las marcas de separación del eje.
- **lwt.ticks:** Define el grosor de las marcas de separación del eje.

1.2.4 Colores y fondos

Los colores y fondos hacen referencia al manejo de la coloración de los elementos del gráfico, donde se pueden modificar los colores de cada elemento.

```r
hist(x, main = "Histograma de frecuencias \n del largo de sepalo", xlab = "cm", ylab = "Frecuencias", su
```



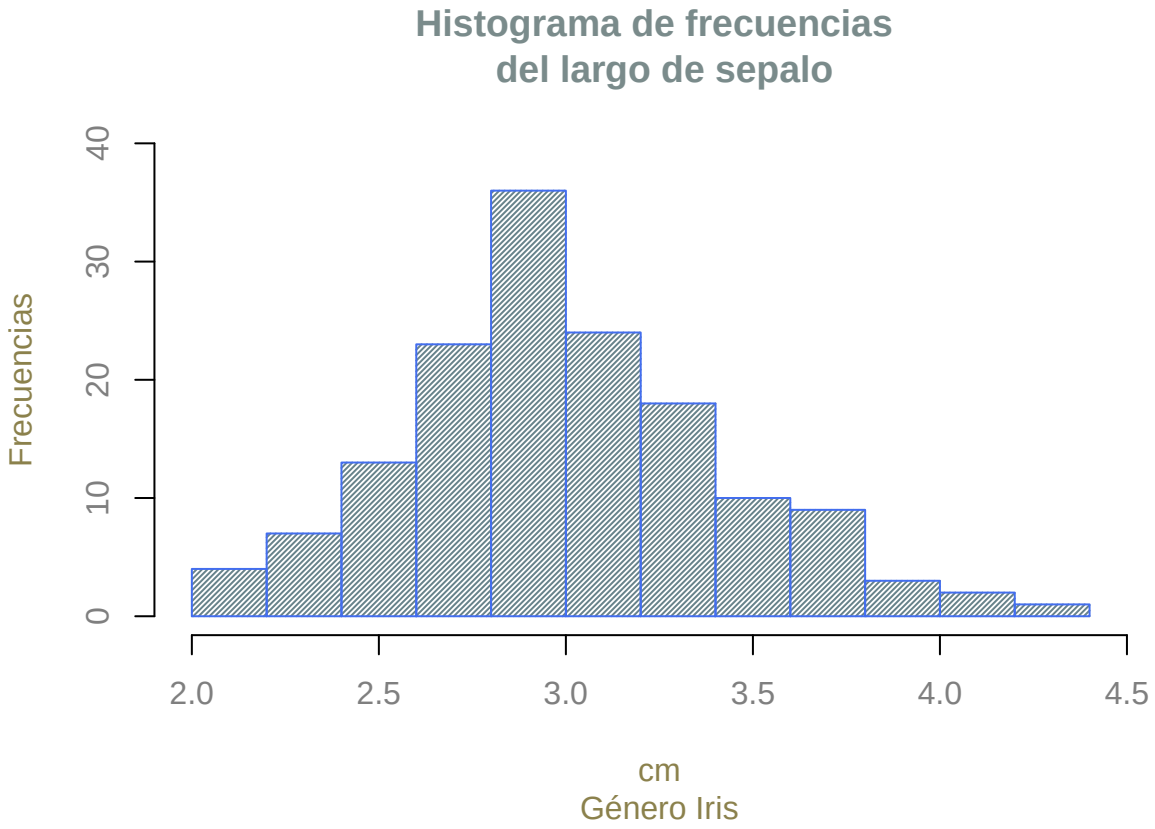
Como se abordo en la sección de títulos y etiquetas, la coloración se puede fijar directamente a partir de subparámetros de **col**. No obstante, hace falta abordar la función del parámetro **col** sin el uso de partículas.

- **col:** permite dar color a las secciones del gráfico, a modo de llenado. Los colores pueden definirse por nombre o por número, para indicar el color por su nombre se debe poner este entre comillas.
- **border:** permite dar color a los bordes de cada una de las secciones en el gráfico.

R cuenta con una amplia paleta de colores, la cual se puede consultar directamente usando el comando *colors()* o en el documento "Colors en R".

La opción de coloración anteriormente usada permite generar colores sólidos, sin embargo, se puede hacer un llenado con lineas usando una opción de densidad de llenado, tal como se puede observar a continuación.

```r
hist(x, main = "Histograma de frecuencias \n del largo de sepalo", xlab = "cm", ylab = "Frecuencias", su
```



Los parámetros de densidad dentro de la función *hist()* son:

- **density:** este parámetro permite definir la cantidad cantidad de lineas en el área de llenado del gráfico.
- **angle:** este parámetro permite definir la orientación de las lineas de llenado, percibiendo entradas númericas de 0 a 360, representando los grados.

1.2.5 Consideraciones finales

1.3 Boxplot

1.4 Densidad

1.5 Violines

1.6 Diagramas de barras

2 Paquetes para mejoras gráficas

2.1 extrafont

Solo corre para exportar las imagenes en formato en pdf o como parte de la capa theme() del paquete ggplot2 empleando el parámetro text = element_text(family = fuente, size = tamaño)

2.2 ggplot2

2.3 ggmap

2.4 lattice