Estadística Descriptiva con Datos Cualitativos

Ramon Ceballos

26/1/2021

REPRESENTACIÓN GRÁFICA DE ANÁLISIS ESTADÍSTICO CUALITATIVO

1. Diagrama de Barras

El tipo de gráfico más usado para representar variables cualitativas son los **diagramas de barras** (`bar plots`). Como su nombre indica, un diagrama de barras contiene, para cada nivel de la variable cualitativa, una barra de altura su frecuencia.

Por tanto, dado un Dat Frame, las columnas tendrá que ser un factor (variable cualitativa) dentro una tabla de frecuencias.

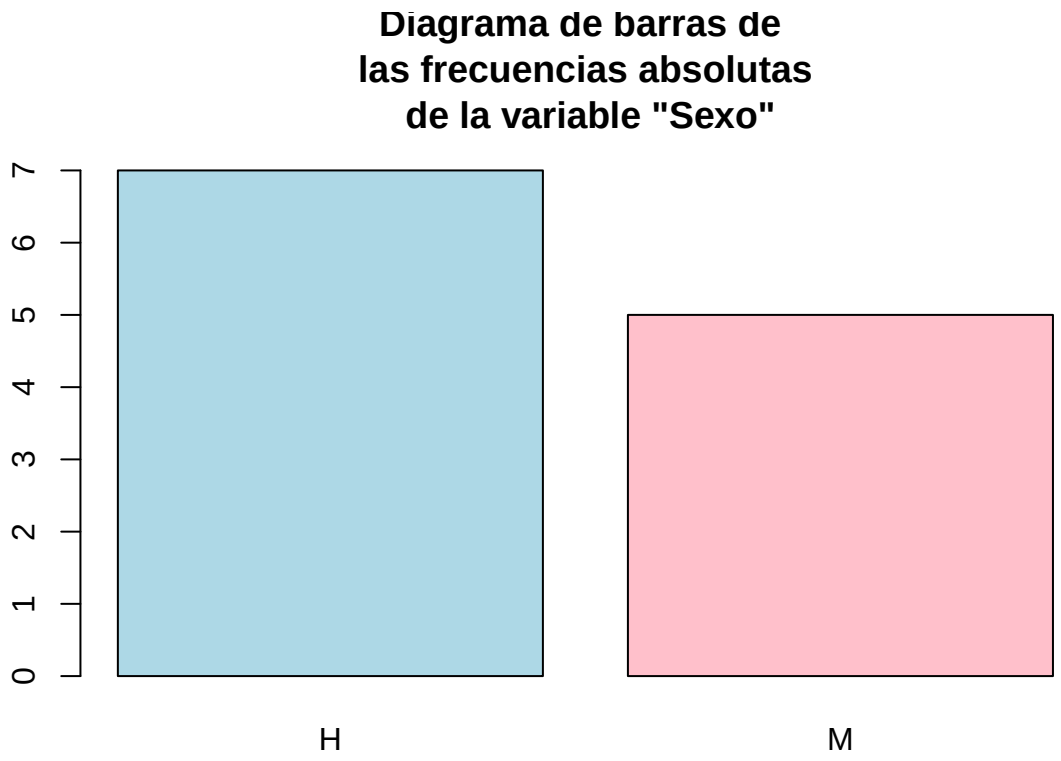
La manera más sencilla de dibujar un diagrama de barras de las frecuencias absolutas o relativas de una variable cualitativa es usando la instrucción **barplot()** aplicada a la tabla correspondiente.

¡Atención! Como pasaba con *prop.table()*, el argumento de *barplot* ha de ser una tabla, y, por consiguiente, se ha de aplicar al resultado de *table()* o de *prop.table()*, nunca al vector de datos original.

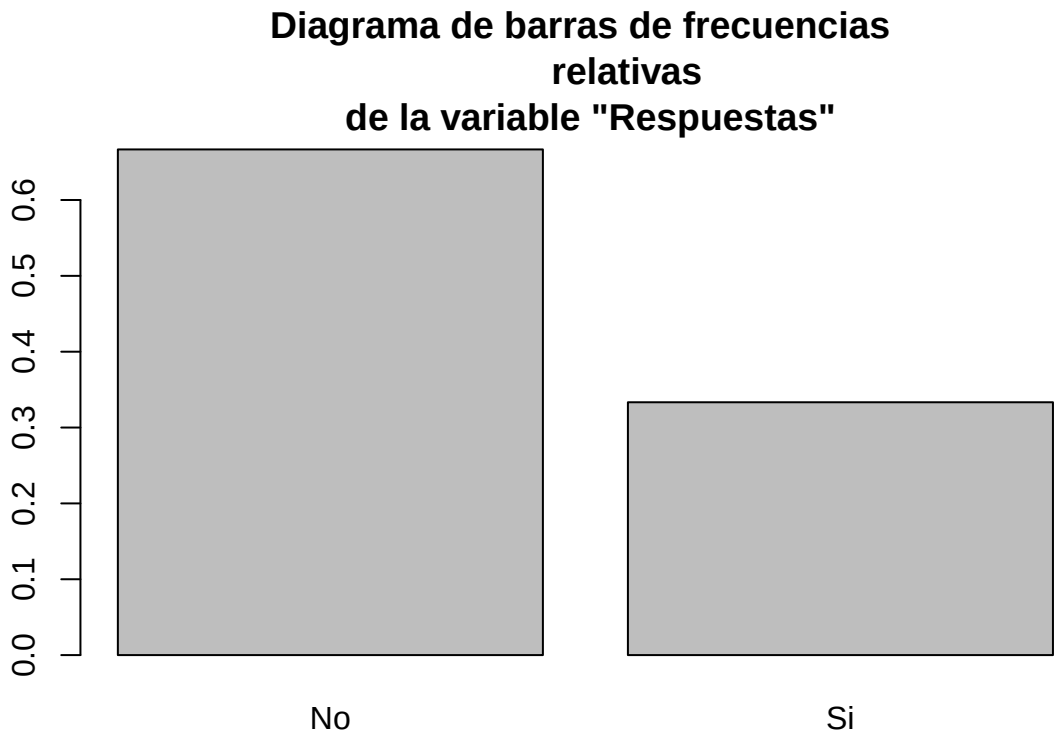
```
Sexo= sample(c("H", "M"), size = 12, replace = T) #H = hombre, M = mujer
Sexo
```

```
##  [1] "M" "H" "H" "H" "M" "M" "H" "H" "H" "M" "M" "H"
```

```
#main permite otorgar el título
#col permite da el color a las barras en orden de factores
barplot(table(Sexo),
col=c("lightblue","pink"), main="Diagrama de barras de
las frecuencias absolutas\n de la variable \"Sexo\"\n")
```




```r
Respuestas = sample(c("Si", "No"), size = 12, replace = T)

#bar plot de la frecuecnia relativa absoluta para respuestas
#sin color asignado
barplot(prop.table(table(Respuestas)), main="Diagrama de barras de frecuencias
        relativas\n de la variable \"Respuestas\"")
```



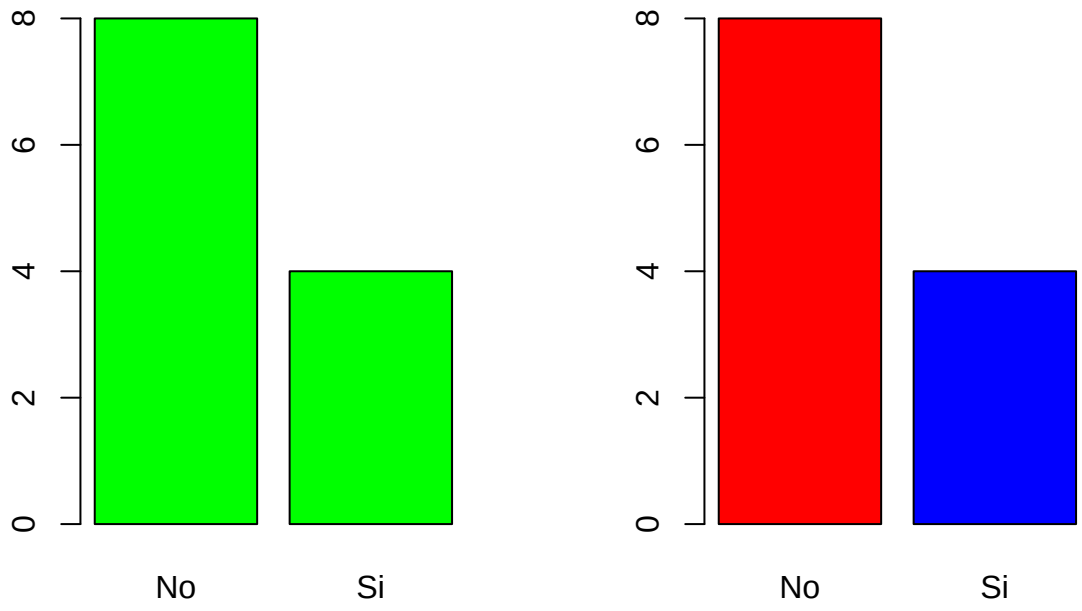
1.1. Parámetros para la función barplot()

Habréis observado que en las funciones `barplot()` anteriores hemos usado el parámetro `main` para poner título a los diagramas; en general, la función `barplot()` admite los parámetros de `plot` que tienen sentido en el contexto de los diagramas de barras: `xlab`, `ylab`, `main`, etc. Los parámetros disponibles se pueden consultar en `help(barplot)`. Aquí sólo vamos a comentar algunos.

Se pueden especificar los colores de las barras usando el parámetro `col`. Si se iguala a un solo color, todas las barras serán de este color, pero también se puede especificar un color para cada barra, igualando `col` a un vector de colores. Paquete Rcolorblue para más colores.

Cuando se da un vector de menos colores que niveles de variables representados, estos se repiten en bucle.

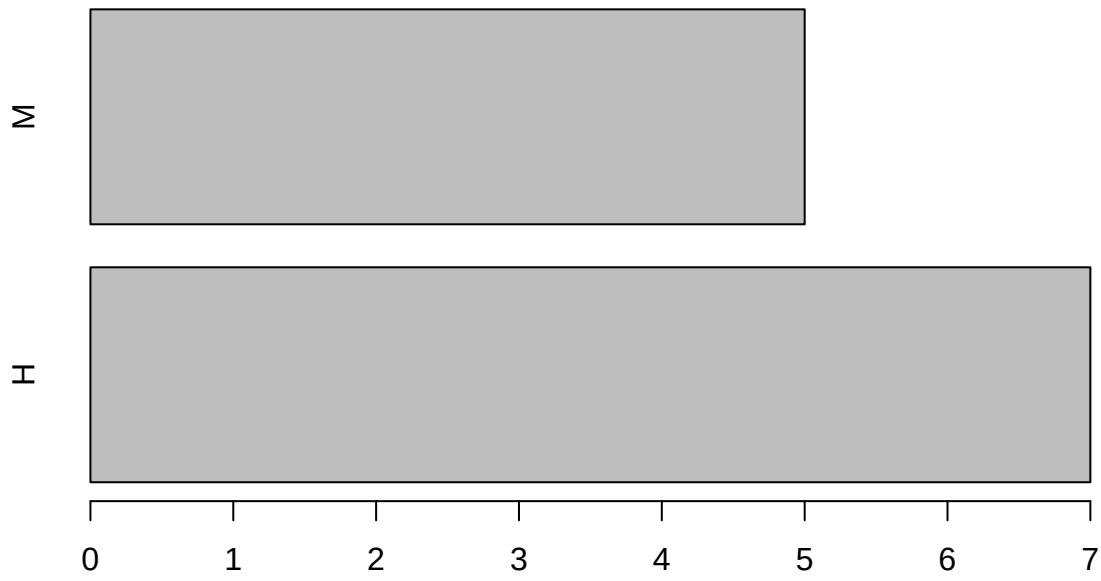
```
#Para definir que en una figura aparecerán un gráfico al lado del otro
par(mfrow=c(1,2))

barplot(table(Respuestas), col=c("green"))

barplot(table(Respuestas), col=c("red","blue"))
```




```
#Finaliza la representación par
par(mfrow=c(1,1))
```

Una opción interesante es dibujar las **barras horizontales** en vez de verticales: para hacerlo, se tiene que añadir el parámetro `horiz=TRUE`.

```
barplot(table(Sexo), horiz = TRUE)
```

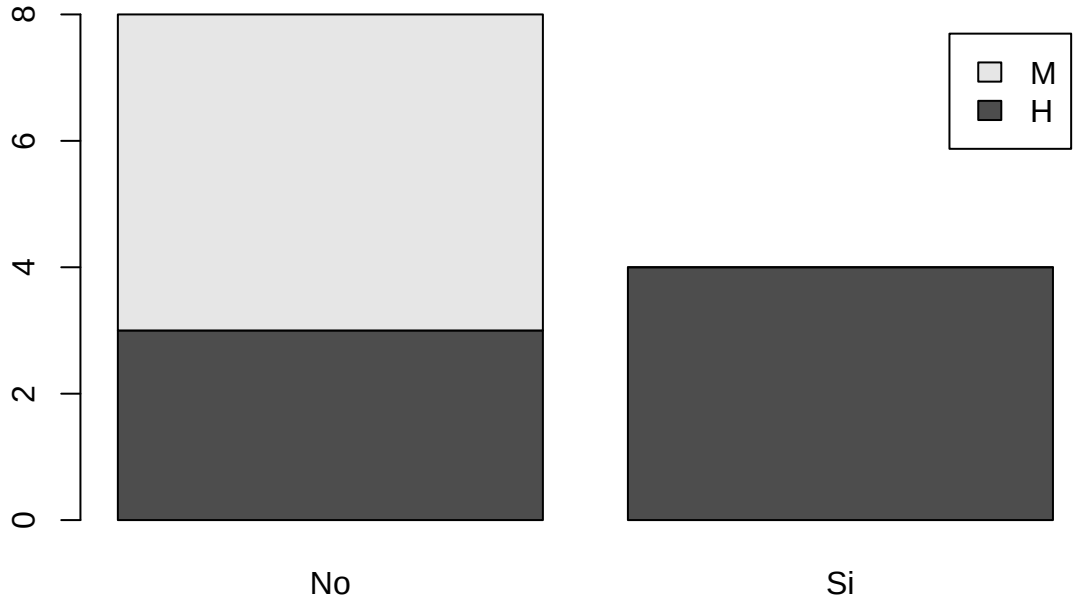


1.2. Tablas bidimensionales con la función barplot()

Se utiliza **barplot(1º variable, 2º variable)** para representarla. Por defecto. Por defecto dibuja la frecuecnia de la segunda variable cortada por la frecuencia de la primera variable. Se conoce como *diagrama de barras apiladas*.

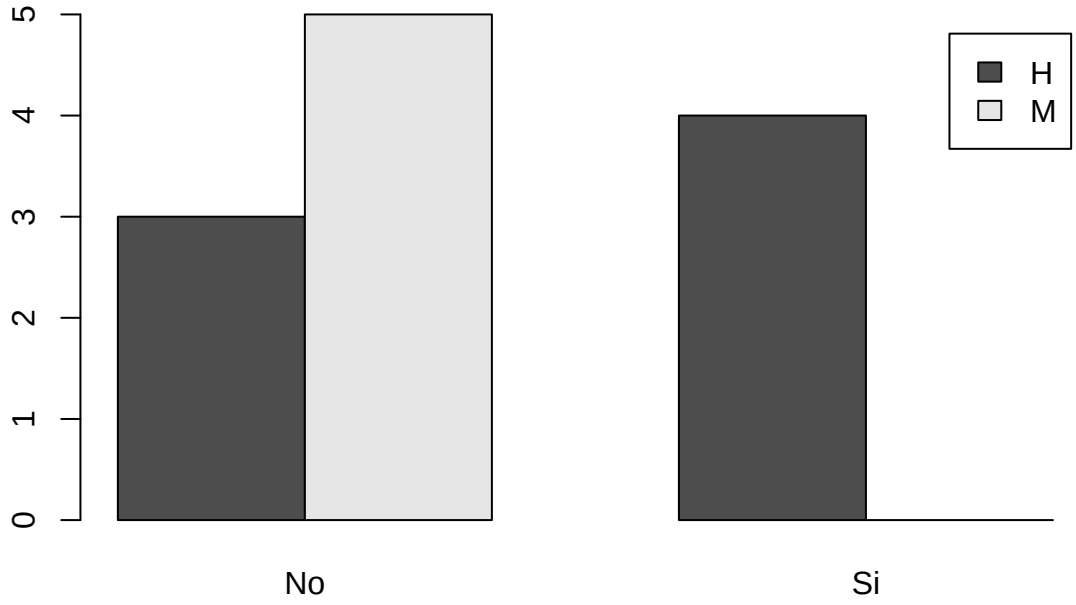
Las barras globales corresponden a los niveles de la variable que define la segunda columna del table(); mientras que cada una de estas barras globales se divide en sectores (en orden ascendente de niveles) que representan los niveles de la otra variable (1a variable) del table().

```
#Las respuestas aparecen divididas en dos (total de no y si)
#Luego aparecen los hombres que dicen si y no; y las mujeres que dicen si y no
barplot(table(Sexo,Respuestas), legend.text = TRUE)
```



Hay un parámetro llamado **beside=TRUE** que permite que las barras no aparezcan apiladas. Se obtiene un **diagrama de barras por bloques**.

```
barplot(table(Sexo,Respuestas), beside=TRUE, legend.text=TRUE)
```

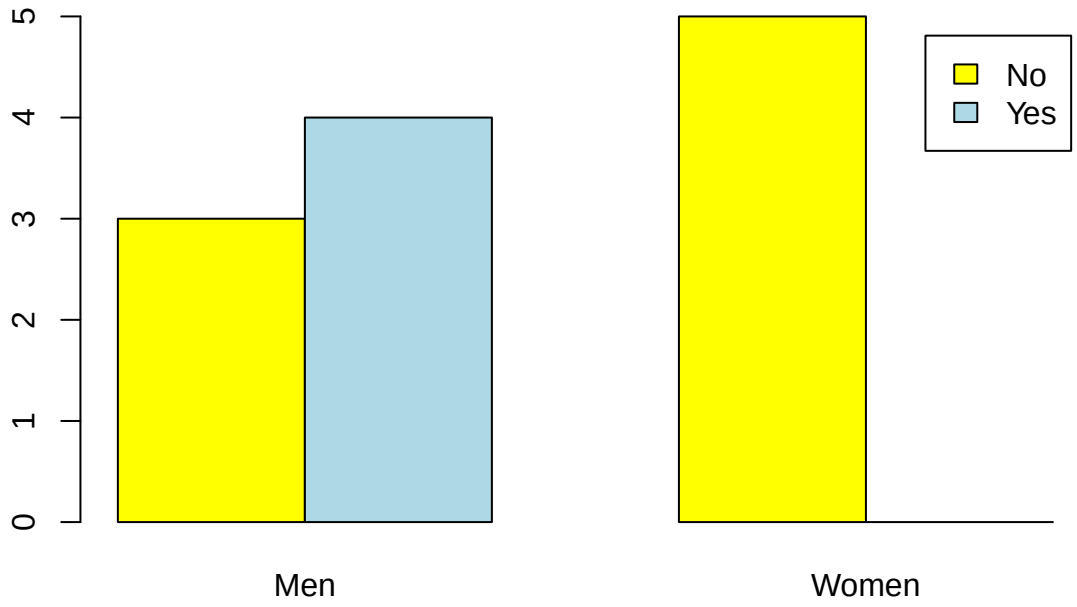


El objetivo final de esta representación, es presentar la información de una forma sencilla y adecuada. Esto dependerá del punto de vista buscado. es recomendable probar con diferentes distribuciones,

Es muy importante añadir a un diagrama de barras con dos variables cualitativas la legenda. Para adicionarla por defecto se usa **legend.text = TRUE**.

Si se quiere modificar el nombre de la leyenda se puede ajustar, tal como aparecen abajo.

```
barplot(table(Respuestas,Sexo), beside=TRUE, names=c("Men", "Women"),
        col=c("yellow","lightblue"), legend.text=c("No","Yes"))
```



2. Diagramas Circulares (pastel o tarta)

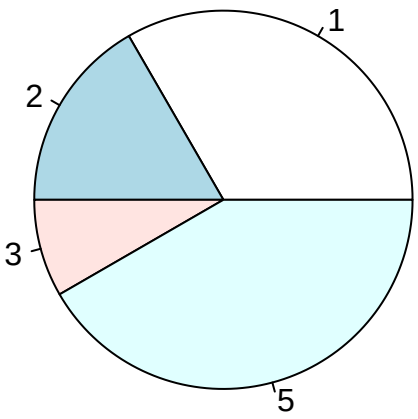
Un tipo muy popular de representación gráfica de variables cualitativas son los **diagramas circulares**. En un diagrama circular (`pie chart`) se representan los niveles de una variable cualitativa como sectores circulares de un círculo, de manera que el ángulo (o equivalentemente, el área) de cada sector sea proporcional a la frecuencia del nivel al que corresponde.

Con R, este tipo de diagramas se producen con la instrucción **pie ()**, de nuevo aplicada a una *tabla de frecuencias* y no al vector original.

2.1. Parámetros función pie()

La función `pie` admite muchos parámetros para modificar el resultado: se pueden cambiar los colores con `col`, se pueden cambiar los nombres de los niveles con `names`, se puede poner un título con `main`, etc.; podéis consultar la lista completa de parámetros en `help(pie)`.

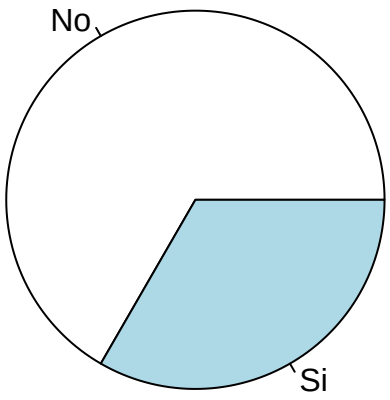
```
x =c(2, 5, 1, 5, 5, 5, 1, 1, 5, 1, 3, 2)
#Genera diagrama circular
pie(table(x), main="Diagrama circular de la variable x")
```




```
#Para respuestas
Respuestas
```

```
##  [1] "No" "Si" "Si" "Si" "No" "No" "No" "No" "Si" "No" "No" "No"
```

```
pie(table(Respuestas), main="Diagrama circular de la variable Respuestas")
```



2.2. Complejidad de visualización

Pese a su popularidad, es poco recomendable usar diagramas circulares porque a veces es difícil, a simple vista, comprender las relaciones entre las frecuencias que representan.

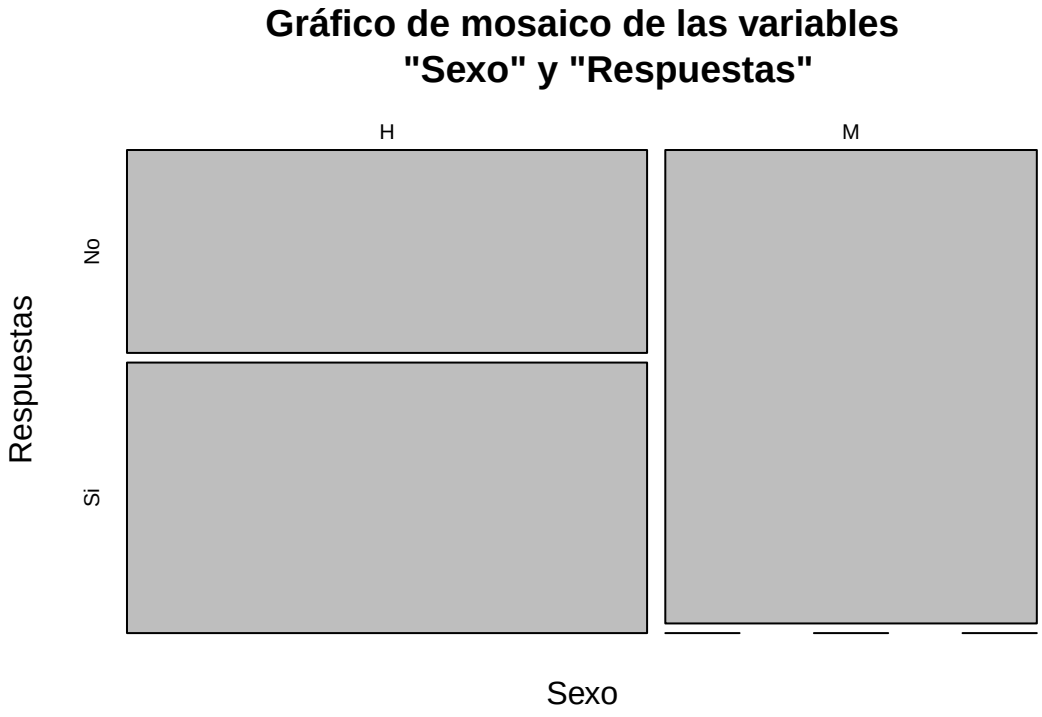
3. Diagrama o Gráfico de Mosaicos

Otra representación de las tablas multidimensionales de frecuencias son los **gráficos de mosaico**. Estos gráficos se obtienen sustituyendo cada entrada de la tabla de frecuencias por una región rectangular de área proporcional a su valor.

En concreto, para obtener el gráfico de mosaico de una tabla bidimensional, se parte de un cuadrado de lado 1, primero se divide en barras verticales de amplitudes iguales a las frecuencias relativas de una variable, y luego cada barra se divide, a lo alto, en regiones de alturas proporcionales a las frecuencias relativas marginales de cada nivel de la otra variable, dentro del nivel correspondiente de la primera variable.

Un gráfico de mosaico de una tabla se obtiene con R aplicando la función **plot()** a la tabla, o también la función **mosaicplot()**. Esta última también se puede aplicar a matrices.

```
#Ejemplo tabla bidimensional
plot(table(Sexo,Respuestas), main="Gráfico de mosaico de las variables
     \"Sexo\" y \"Respuestas\"")
```



Cuidado con este gráfico porque es complicado de leer.

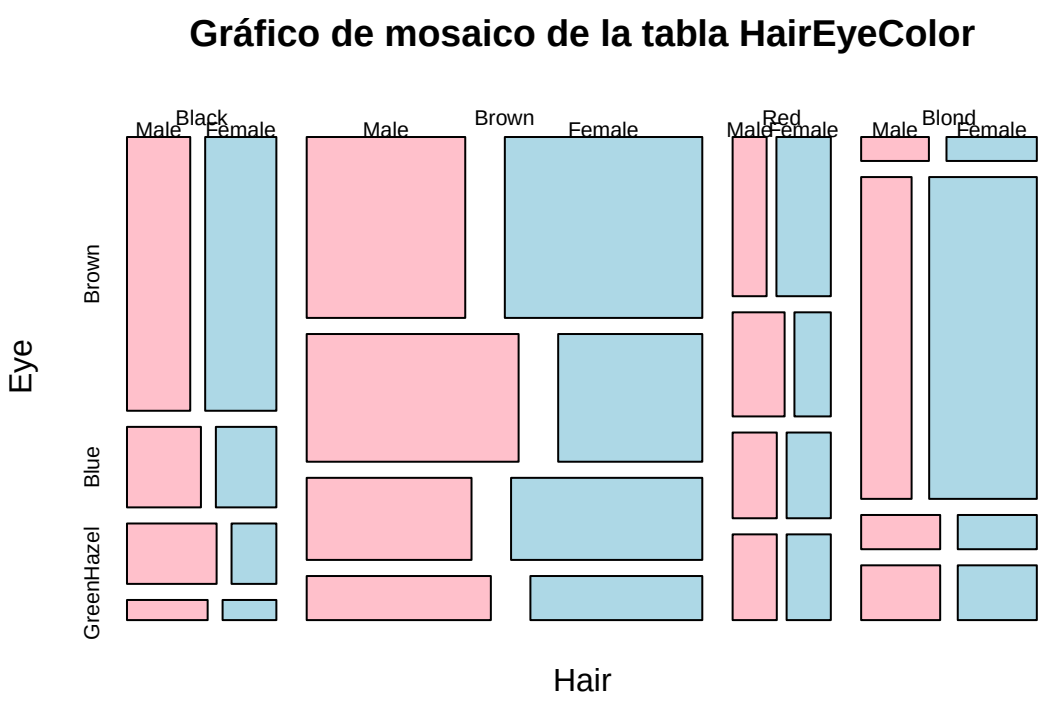
3.1. Gráfico de mosaicos para tablas tridimensionales

En el gráfico de mosaico de una tabla tridimensional, primero se divide el cuadrado en barras verticales de amplitudes iguales a las frecuencias relativas de una variable.

Luego cada barra se divide, a lo alto, en regiones de alturas proporcionales a las frecuencias relativas marginales de cada nivel de una segunda variable, dentro del nivel correspondiente de la primera variable.

Finalmente, cada sector rectangular se vuelve a dividir a lo ancho en regiones de amplitudes proporcionales a las frecuencias relativas marginales de cada nivel de la tercera variable dentro de la combinación correspondiente de niveles de las otras dos.

```
#Ejemplo tabla tridimensional
plot(HairEyeColor, main="Gráfico de mosaico de la tabla HairEyeColor",
     col=c("pink","lightblue"))
```



En la práctica, no se suele utilizar demasiado el diagrama de mosaicos.

4. Muchos más gráficos

Además de sus parámetros usuales, la función **plot** admite algunos parámetros específicos cuando se usa para producir el gráfico de mosaico de una tabla. Estos parámetros se pueden consultar en **help(mosaicplot)**.

Los paquetes **vcd** y **vcdExtra** incluyen otras funciones que producen representaciones gráficas interesantes de tablas tridimensionales.

- La función **cotabplot** de **vcd** produce un diagrama de mosaico para cada nivel de la tercera variable.
- La función **mosaic3d** de **vcdExtra** produce un diagrama de mosaico tridimensional en una ventana de una aplicación para gráficos 3D interactivos.

Estos paquetes son un poco antiguos, hoy día se utiliza ggplot con tidyverse.

```
#library(vcd)
#País = sample(c("Francia","Alemania","España"), size = length(Sexo), replace = T)
#cotabplot(table(Sexo,Respuestas,País))
```

```
#library(vcdExtra)
#mosaic3d(HairEyeColor, type="expected", box=TRUE,
#col=c("pink","lightblue"))
```