R básico

10-2022

Sección 1

Análisis de datos

Principales indicadores descriptivos de series de datos

Cuando tenemos una serie de datos que describen algunos aspectos de un conjunto de individuos queremos llevar a cabo un análisis estadístico. Estos análisis estadísticos se clasifican en:

- Análisis exploratorio, o descriptivo, o Key Perfomance indicators si nuestro objetivo es resumir, representar y explicar los datos concretos de los que disponemos. La estadística descriptiva/ análisis de datos es el conjunto de técnicas que se usan con este fin.
- Análisis inferencial, si nuestro objetivo es deducir (inferir), a partir de estos datos, información significativa sobre el total de la población o las poblaciones de interés. Las técnicas que se usan en este caso forman la >estadística inferencial.

Análisis estadístico de los datos

Existe relación entre ambos. Cualquier análisis inferencial se suele empezar explorando los datos que se usarán así cómo también muchas técnicas descriptivas permiten estimar propiedades de la población de la que se ha extraído la muestra.

Ejemplo

La media aritmética de las alturas de una muestra de individuos nos da un valor representativo de esta muestra, pero también estima la media de las alturas del total de la población

Análisis estadístico de los datos

Nos centraremos en entender algunas técnicas básicas de la estadística descriptiva orientadas al análisis de datos.

Estas consistirán en una serie de medidas, gráficos y modelos descriptivos que nos permitirán resumir y explorar un conjunto de datos.

Objetivo final: entender los datos lo mejor posible.

Tipos de datos

Trabajamos con datos multidimensionales: observamos varias características de una serie de individuos.

Se registran en un archivo de ordenador con un formato preestablecido. Por ejemplo texto simple (codificado en diferentes formatos: ASCII, isolatin...), hojas de cálculo (archivos de Open Office o Excel), bases de datos, etc.

Tipos de datos

Una de las maneras básicas de almacenar datos es en forma de tablas de datos. En R hacemos uso de data frames.

En una tabla de datos cada columna expresa una variable, mientras que cada fila corresponde a las observaciones de estas variables para un individuo concreto.

- Los datos de una misma columna tienen que ser del mismo tipo, porque corresponden a observaciones de una misma propiedad.
- Las filas en principio son de naturaleza heterogénea, porque pueden contener datos de diferentes tipos.

Tipos de datos

Los tipos de datos que consideramos son los siguientes:

- Datos de tipo atributo, o cualitativos: Expresan una cualidad del individuo. En R guardaremos las listas de datos cualitativos en vectores (habitualmente, de palabras), o en factores si vamos a usarlos para clasificar individuos.
- Datos ordinales: Similares a los cualitativos, con la única diferencia de que se pueden ordenar de manera natural. Por ejemplo, las calificaciones en un control (suspenso, aprobado, notable, sobresaliente). En R guardaremos las listas de datos ordinales en factores ordenados.
- Datos cuantitativos: Se refieren a medidas, tales como edades, longitudes, etc. En R guardaremos las listas de datos cuantitativos en vectores numéricos.

Tipos de datos

```
head(iris ,5)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
```

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Sección 2

Descripción de datos cualitativos

¿Qué son los datos cualitativos?

Los datos cualitativos corresponden a observaciones sobre cualidades de un objeto o individuo.

Suelen codificarse por medio de palabras, pero también se pueden usar números que jueguen el papel de etiquetas.

Ejemplo

Es habitual representar No (o Falso, Fracaso, Ausente...) con un 0, y Sí (o Verdadero, Éxito, Presente...) con un 1

¿Qué son los datos cualitativos?

Los datos cualitativos son aquellos que pueden ser iguales o diferentes, pero que no admiten ningún otro tipo de comparación significativa.

Es decir, que no tenga ningún sentido preguntarse si uno es más grande que otro, ni efectuar operaciones aritméticas con ellos, aunque estén representados por números.

¿Qué son los datos cualitativos?

Por lo tanto, un mismo conjunto de datos puede ser cualitativo o de otro tipo, según el análisis que vayamos a hacer de él.

Ejemplo

Si hemos anotado durante unos años los días de la semana en los que ha llovido y queremos contar cuántas veces ha ocurrido en lunes, cuántas en martes, etc., esta lista de nombres (o números) serán datos cualitativos. Si, en cambio, queremos estudiar cómo se comportan los días de lluvia según avanza la semana, y por lo tanto el orden de los días es relevante, serán datos ordinales

¿Qué son los datos cualitativos?

Variable cualitativa: lista de observaciones de un tipo de datos cualitativos sobre un conjunto concreto de objetos.

Niveles: diferentes valores que pueden tomar estos datos. Por ejemplo, los dos niveles de una variable Sexo serían `M` (Macho) y `H` (Hembra), o sinónimos.

Con R, usaremos vectores y factores para representar variables cualitativas. Los factores nos servirán para agrupar las observaciones según los niveles de la variable. De esta manera podremos segmentar la población que representa la variable en grupos o subpoblaciones, asignando un grupo a cada nivel, y podremos comparar el comportamiento de otras variables sobre estos grupos.

Estudio de Frecuencias

Dada una variable cualitativa, para cada uno de sus niveles podemos contar cuántos datos hay en ese nivel (frecuencia absoluta) y qué fracción del total representan (frecuencia relativa).

Estudio de Frecuencias

Ejemplo

Supongamos que tenemos un tipo de datos cualitativos con niveles

$$l_1, l_2, \cdots, l_k$$

Efectuamos n observaciones de este tipo de datos, y denotamos por

$$x_1, x_2, \cdots, x_n$$

los resultados que obtenemos con

$$x_j \in \{l_1, l_2, \cdots, l_k\}$$

Estas observaciones forman una variable cualitativa

Estudio de Frecuencias

Con estas notaciones:

La frecuencia absoluta, n_j, del nivel l_j en esta variable cualitativa es el número de observaciones en las que x_i toma el valor l_j.

La frecuencia relativa del nivel l_j en esta variable cualitativa es la fracción

$$f_j = \frac{n_j}{n}$$

Es decir, la frecuencia relativa del nivel l_j es la fracción (en tanto por uno) de observaciones que corresponden a este nivel.

La moda de esta variable cualitativa es su nivel, o niveles, de mayor frecuencia (absoluta o relativa).

Estudio de Frecuencias

Ejemplo

Supongamos que se ha realizado un seguimiento a 20 personas asistentes a un congreso. Uno de los datos que se han recogido sobre estas personas ha sido su sexo. El resultado ha sido una variable cualitativa formada por las 20 observaciones siguientes:

`Mujer, Mujer, Hombre, Mujer, Mujer, Mujer, Mujer, Mujer, Hombre, Mujer, Hombre, Hombre, Mujer, Mujer, Hombre, Mujer, Mujer, Mujer, Mujer, Hombre`

Sus dos niveles son `Hombre` y `Mujer`. En esta variable hay 14 mujeres y 6 hombres. Éstas son las frecuencias absolutas de estos niveles.

Estudio de Frecuencias

Puesto que en total hay 20 individuos, sus frecuencias relativas son

$$\text{Hombre} = \frac{6}{20} = 0.3, \qquad \text{Mujer} = \frac{14}{20} = 0.7$$

En este caso $l_1 = \text{Hombre}$ y $l_2 = \text{Mujer}$, $n = 20$ (el número de observaciones efectuadas), y $x_1, \cdots, x_{20}$ formarían la muestra de sexos

Estudio de Frecuencias

Ejemplo

La tabla siguiente resume las frecuencias absolutas y relativas de la variable cualitativa del ejemplo anterior, con las notaciones que acabamos de introducir.

Sexo	n_i	f_i	%
Hombre	6	0.3	30%
Mujer	14	0.7	70%
Total	20	1	100%

Su moda es el nivel `Mujer`

Tablas de frecuencias unidimensionales

Supongamos que tenemos una variable cualitativa guardada en un vector o un factor como la siguiente:

```
x = sample(1:5, size = 12, replace = TRUE)
x
```

```
##  [1] 1 3 5 5 4 3 4 5 5 1 5 4
```

```
Respuestas=factor(sample(c("Si", "No"), size = 12, replace = TRUE))
Respuestas
```

```
##  [1] No Si No Si Si Si Si Si Si Si No Si
## Levels: No Si
```

Tablas de frecuencias unidimensionales

Con R, la tabla de frecuencias absolutas de un vector que representa una variable cualitativa se calcula con la función `table()`.

```
table(x)
```

```
## x
## 1 3 4 5
## 2 2 3 5
```

```
table(Respuestas)
```

```
## Respuestas
## No Si
##  3  9
```

Tablas de frecuencias unidimensionales

El resultado de una función `table()` es un objeto de datos de un tipo nuevo: una tabla de contingencia, una `table` en el argot de R.

Al aplicar `table()` a un vector obtenemos una tabla unidimensional formada por una fila con los niveles de la variable y una segunda fila donde, debajo de cada nivel, aparece su frecuencia absoluta en el vector.

Tablas de frecuencias unidimensionales

Los nombres de las columnas de una tabla unidimensional se obtienen con la función `names()`.

```
names(table(x))
```

```
## [1] "1" "3" "4" "5"
```

```
names(table(Respuestas))
```

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

Tablas de frecuencias unidimensionales

En la `table` de un vector sólo aparecen los nombres de los niveles presentes en el vector. Si el tipo de datos cualitativos usado tenía más niveles y queremos que aparezcan explícitamente en la tabla (con frecuencia 0), hay que transformar el vector en un factor con los niveles deseados.

```
z=factor(x, levels=1:7) #Los niveles serán 1,2,3,4,5,6,7
z
```

```
##  [1] 1 3 5 5 4 3 4 5 5 1 5 4
## Levels: 1 2 3 4 5 6 7
```

```
table(z)
```

```
## z
## 1 2 3 4 5 6 7
## 2 0 2 3 5 0 0
```

Tablas de frecuencias unidimensionales

Podemos pensar que una tabla unidimensional es como un vector de números donde cada entrada está identificada por un nombre: el de su columna. Para referirnos a una entrada de una tabla unidimensional, podemos usar tanto su posición como su nombre (entre comillas, aunque sea un número).

```
table(x)[3] #La tercera columna de table(x)
```

```
## 4
## 3
```

```
table(x)["7"] #¿La columna de table(x) con nombre 7?
```

```
## <NA>
##   NA
```

Tablas de frecuencias unidimensionales

```
table(x)["5"] #La columna de table(x) con nombre 5
```

```
## 5
## 5
```

```
3*table(x)[2] #El triple de la segunda columna de table(x)
```

```
## 3
## 6
```

Tablas de frecuencias unidimensionales

Las tablas de contingencia aceptan la mayoría de las funciones que ya hemos utilizado para vectores.

```
sum(table(x)) #Suma de las entradas de table(x)
```

```
## [1] 12
```

```
sqrt(table(Respuestas))
```

```
## Respuestas
##       No       Si
## 1.732051 3.000000
```

```
# Raíces cuadradas de las entradas de table(Respuestas)
```

Tablas de frecuencias unidimensionales

La tabla de frecuencias relativas de un vector se puede calcular aplicando la función `prop.table()` a su `table`. El resultado vuelve a ser una tabla de contingencia unidimensional.

```
prop.table(table(x))
```

```
## x
##         1         3         4         5
## 0.1666667 0.1666667 0.2500000 0.4166667
```

```
prop.table(table(Respuestas))
```

```
## Respuestas
##   No   Si
## 0.25 0.75
```

Tablas de frecuencias unidimensionales

¡CUIDADO! La función `prop.table()` se tiene que aplicar al resultado de `table`, no al vector original. Si aplicamos `prop.table()` a un vector de palabras o a un factor, dará un error, pero si la aplicamos a un vector de números, nos dará una tabla.

Esta tabla no es la tabla de frecuencias relativas de la variable cualitativa representada por el vector, sino la tabla de frecuencias relativas de una variable que tuviera como tabla de frecuencias absolutas este vector de números, entendiendo que cada entrada del vector representa la frecuencia de un nivel diferente.

```
prop.table(x)
```

```
##  [1] 0.02222222 0.06666667 0.11111111 0.11111111 0.08888889 0.06666667
##  [7] 0.08888889 0.11111111 0.11111111 0.02222222 0.11111111 0.08888889
```

Tablas de frecuencias unidimensionales

```
X=c(1,1,1)
prop.table(table(X))
```

```
## X
## 1
## 1
```

```
prop.table(X)
```

```
## [1] 0.3333333 0.3333333 0.3333333
```

Tablas de frecuencias unidimensionales

También podemos calcular la tabla de frecuencias relativas de un vector dividiendo el resultado de `table` por el número de observaciones.

```
table(x)/length(x)
```

```
## x
##         1         3         4         5
## 0.1666667 0.1666667 0.2500000 0.4166667
```

Tablas de frecuencias unidimensionales

Dados un vector x y un número natural n, la instrucción

```
names(which(table(x)==n))
```

nos da los niveles que tienen frecuencia absoluta n en x.

```
table(x)
```

```
## x
## 1 3 4 5
## 2 2 3 5
```

```
names(which(table(x)==1))
```

```
## character(0)
```

Tablas de frecuencias unidimensionales

En particular, por lo tanto,

```
names(which(table(x)==max(table(x))))
```

nos da los niveles de frecuencia máxima en x: su moda.

```
names(which(table(x)==max(table(x))))
```

```
## [1] "5"
```

```
names(which(table(Respuestas)==max(table(Respuestas))))
```

```
## [1] "Si"
```

Tablas de frecuencias unidimensionales

Ejercicio

Recuperad el ejemplo de los 6 hombres y las 14 mujeres anterior y utilizando R, calculad su tabla de frecuencias absolutas, su tabla de frecuencias relativas y la moda.

Pista: usad la función `rep()` para no tener que escribir los datos a mano.

Tablas de frecuencias unidimensionales

```
Sexo_Ger=c("Mujer","Mujer","Hombre","Mujer","Mujer","Mujer",
           "Mujer","Mujer","Hombre","Mujer","Hombre","Hombre",
           "Mujer", "Mujer","Hombre","Mujer","Mujer","Mujer",
           "Mujer","Hombre")
t0=table(Sexo_Ger)
prop.table(t0)
```

```
## Sexo_Ger
## Hombre  Mujer
##    0.3    0.7
```

```
names(which(t0==max(t0)))
```

```
## [1] "Mujer"
```

Tablas de frecuencias bidimensionales

La función `table()` también permite construir tablas de frecuencias conjuntas de dos o más variables.

Supongamos que el vector `Respuestas` anterior contiene las respuestas a una pregunta dadas por unos individuos cuyos sexos tenemos almacenados en un vector `Sexo`, en el mismo orden que sus respuestas. En este caso, podemos construir una tabla que nos diga cuántas personas de cada sexo han dado cada respuesta.

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

```
##           Sexo
## Respuestas H M
##         No 0 3
##         Si 6 3
```

Tablas de frecuencias bidimensionales

Ejercicio

- Comprobad qué ocurre si cambiamos el orden de las columnas en la función `table()`
- Usad la función `t()` para transponer ambas tablas y comprobad el resultado

Tablas de frecuencias bidimensionales

Para referirnos a una entrada de una tabla bidimensional podemos usar el sufijo `[ , ]` como si estuviéramos en una matriz o un data frame. Dentro de los corchetes, tanto podemos usar los índices como los nombres (entre comillas) de los niveles.

```
table(Respuestas ,Sexo)[1,2]
```

```
## [1] 3
```

```
table(Respuestas ,Sexo)["No","M"]
```

```
## [1] 3
```

Tablas de frecuencias bidimensionales

Como en el caso unidimensional, la función `prop.table()` sirve para calcular tablas bidimensionales de frecuencias relativas conjuntas de pares de variables. Pero en el caso bidimensional tenemos dos tipos de frecuencias relativas:

Frecuencias relativas globales: para cada par de niveles, uno de cada variable, la fracción de individuos que pertenecen a ambos niveles respecto del total de la muestra.

Frecuencias relativas marginales: dentro de cada nivel de una variable y para cada nivel de la otra, la fracción de individuos que pertenecen al segundo nivel respecto del total de la subpoblación definida por el primer nivel.

Tablas de frecuencias bidimensionales

Dadas dos variables, se pueden calcular dos familias de frecuencias relativas marginales, según cuál sea la variable que defina las subpoblaciones en las que calculemos las frecuencias relativas de los niveles de la otra variable; no es lo mismo la fracción de mujeres que han contestado que sí respecto del total de mujeres, que la fracción de mujeres que han contestado que sí respecto del total de personas que han dado esta misma respuesta.

Tablas de frecuencias bidimensionales

La tabla de frecuencias relativas globales se calcula aplicando sin más la función `prop.table()` a la `table`.

```
prop.table(table(Sexo,Respuestas)) #Global
```

```
##     Respuestas
## Sexo   No   Si
##    H 0.00 0.50
##    M 0.25 0.25
```

De este modo, la tabla `prop.table(table(Sexo,Respuestas))` nos da la fracción del total que representa cada pareja (sexo, respuesta).

Tablas de frecuencias bidimensionales

Para obtener las marginales, debemos usar el parámetro `margin` al aplicar la función `prop.table()` a la `table`. Con `margin=1` obtenemos las frecuencias relativas de las filas y con `margin=2`, de las columnas.

```
prop.table(table(Sexo,Respuestas), margin=1) #Por sexo
```

```
##     Respuestas
## Sexo  No  Si
##    H 0.0 1.0
##    M 0.5 0.5
```

```
prop.table(table(Sexo,Respuestas), margin=2) #Por respuesta
```

```
##     Respuestas
## Sexo        No        Si
##    H 0.0000000 0.6666667
##    M 1.0000000 0.3333333
```

Tablas de frecuencias bidimensionales

La función `CrossTable()` del paquete `gmodels` permite producir (especificando el parámetro `prop.chisq=FALSE`) un resumen de la tabla de frecuencias absolutas y las tres tablas de frecuencias relativas de dos variables en un formato adecuado para su visualización.

La leyenda *Cell Contents* explica los contenidos de cada celda de la tabla: la frecuencia absoluta, la frecuencia relativa por filas, la frecuencia relativa por columnas, y la frecuencia relativa global. Esta función dispone de muchos parámetros que permiten modificar el contenido de las celdas, y que podéis consultar en `help(CrossTable)`.

Tablas de frecuencias bidimensionales

Una tabla de contingencia bidimensional es, básicamente, una matriz con algunos atributos extra. En particular, podemos usar sobre estas tablas la mayoría de las funciones para matrices que tengan sentido para tablas:

- `rowSums()` y `colSums()` se pueden aplicar a una tabla y suman sus filas y sus columnas, respectivamente.
- También podemos usar sobre una tabla bidimensional (o, en general, multidimensional) la función `apply()` con la misma sintaxis que para matrices.

```
table(Sexo,Respuestas)
```

```
##     Respuestas
## Sexo No Si
##    H  0  6
##    M  3  3
```

Tablas de frecuencias bidimensionales

```
colSums(table(Sexo,Respuestas))
```

```
## No Si
##  3  9
```

```
rowSums(table(Sexo,Respuestas))
```

```
## H M
## 6 6
```

Tablas de frecuencias bidimensionales

```
colSums(prop.table(table(Sexo,Respuestas)))
```

```
##   No   Si
## 0.25 0.75
```

```
rowSums(prop.table(table(Sexo,Respuestas)))
```

```
##   H   M
## 0.5 0.5
```

Tablas a partir de data frames de variables cualitativas

Como ya hemos comentado en varias ocasiones, la manera natural de organizar datos multidimensionales en R es en forma de data frame.

En esta sección explicaremos algunas instrucciones para calcular tablas de frecuencias absolutas a partir de un data frame de variables cualitativas.

Tablas a partir de data frames de variables cualitativas

Para ilustrarla, usaremos el fichero que se encuentra en el la carpeta de datos: `"data/EnergyDrink"`

Este fichero consiste en una tabla de datos con la siguiente información sobre 122 estudiantes de una Universidad de España: su sexo (variable `sexo`), el estudio en el que están matriculados (variable `estudio`) y si consumen habitualmente bebidas energéticas para estudiar (variable `bebe`).

```
Beb_Energ=read.table("../data/EnergyDrink",header=TRUE)
```

Tablas a partir de data frames de variables cualitativas

```
str(Beb_Energ)
```

```
## 'data.frame':    122 obs. of  3 variables:
##  $ estudio: chr  "Informatica" "Mates" "Industriales" "Informatica" ...
##  $ bebe   : chr  "No" "No" "Si" "Si" ...
##  $ sexo   : chr  "Mujer" "Hombre" "Mujer" "Hombre" ...
```

```
head(Beb_Energ,4)
```

```
##        estudio bebe   sexo
## 1  Informatica   No  Mujer
## 2        Mates   No Hombre
## 3 Industriales   Si  Mujer
## 4  Informatica   Si Hombre
```

Tablas a partir de data frames de variables cualitativas

Aplicando la función `summary()` a un data frame de variables cualitativas, obtenemos, a modo de resumen, una tabla con las frecuencias absolutas de cada variable.

```
summary(Beb_Energ)
```

```
##    estudio              bebe              sexo
##  Length:122         Length:122         Length:122
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

Tablas a partir de data frames de variables cualitativas

Esta tabla sólo sirve para ver la información, porque sus entradas son palabras.

```
summary(Beb_Energ)[,2]
```

```
##
## "Length:122        " "Class :character  " "Mode  :character  "
```

Para calcular en un solo paso la table de cada variable, podemos usar la función `apply()` de la manera siguiente:

Tablas a partir de data frames de variables cualitativas

```
apply(Beb_Energ, MARGIN=2, FUN=table)
```

```
## $estudio
##
## Industriales  Informatica        Mates   Telematica
##           37           53           16           16
##
## $bebe
##
## No Si
## 97 25
##
## $sexo
##
## Hombre  Mujer
##     83     39
```

Tablas a partir de data frames de variables cualitativas

De esta manera, obtenemos una `list` cuyas componentes son las tablas que queríamos.

```
apply(Beb_Energ,MARGIN=2,FUN=table)$sexo
```

```
##
## Hombre  Mujer
##     83     39
```

```
table(Beb_Energ$sexo)
```

```
##
## Hombre  Mujer
##     83     39
```

Tablas a partir de data frames de variables cualitativas

Si aplicamos la función `table()` a un data frame de variables cualitativas, obtenemos su tabla de frecuencias absolutas, con las variables ordenadas tal y como aparecen en el data frame.

```
table(Beb_Energ)
```

```
## , , sexo = Hombre
##
##               bebe
## estudio        No Si
##   Industriales 19  6
##   Informatica  30  7
##   Mates         8  1
##   Telematica   10  2
##
## , , sexo = Mujer
##
##               bebe
## estudio        No Si
##   Industriales 10  2
##   Informatica  11  5
##   Mates         6  1
##   Telematica    3  1
```

Tablas a partir de data frames de variables cualitativas

O también podemos hacer...

```
table(Beb_Energ[c(1,3)])
```

```
##              sexo
## estudio       Hombre Mujer
##   Industriales     25    12
##   Informatica      37    16
##   Mates             9     7
##   Telematica       12     4
```

Tablas a partir de data frames de variables cualitativas

Una tercera opción es usar la función `ftable()`, que produce la misma tabla de frecuencias pero en formato plano.

```
ftable(Beb_Energ)
```

```
##                   sexo Hombre Mujer
## estudio      bebe
## Industriales No            19    10
##              Si             6     2
## Informatica  No            30    11
##              Si             7     5
## Mates        No             8     6
##              Si             1     1
## Telematica   No            10     3
##              Si             2     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.

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.

Diagrama de barras

```
barplot(table(Sexo), col=c("lightblue","pink"),
main="Diagrama de barras de
las frecuencias absolutas\n de la variable \"Sexo\"")
```

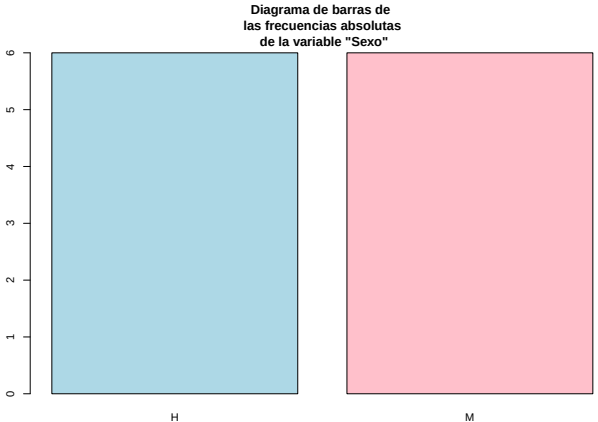


Diagrama de barras

```
barplot(prop.table(table(Respuestas)),
main="Diagrama de barras de frecuencias
      relativas\n de la variable \"Respuestas\"")
```

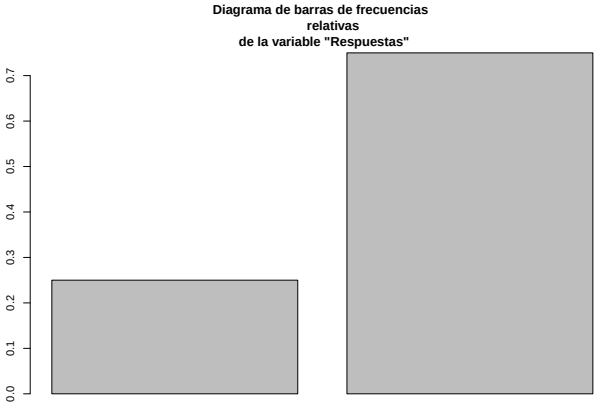


Diagrama de barras - Parámetros

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.

Diagrama de barras - Colores

```
par(mfrow=c(1,2))
barplot(table(Respuestas), col=c("green"))
barplot(table(Respuestas), col=c("red","blue"))
```

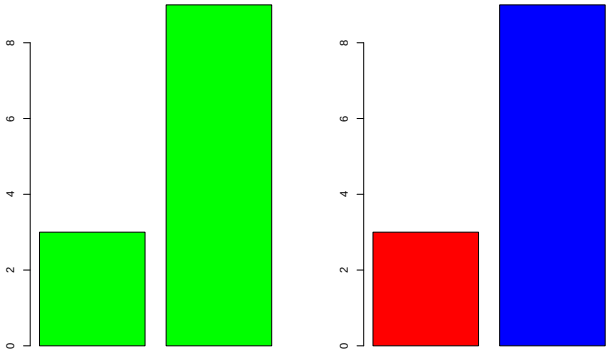


Diagrama de barras - Colores

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

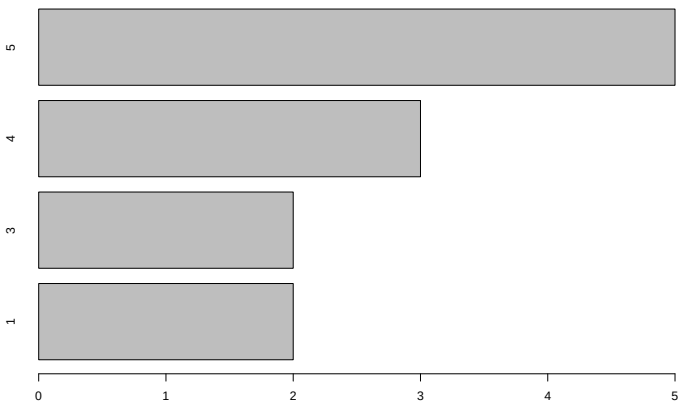


Diagrama de barras - Tabla bidimensional

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

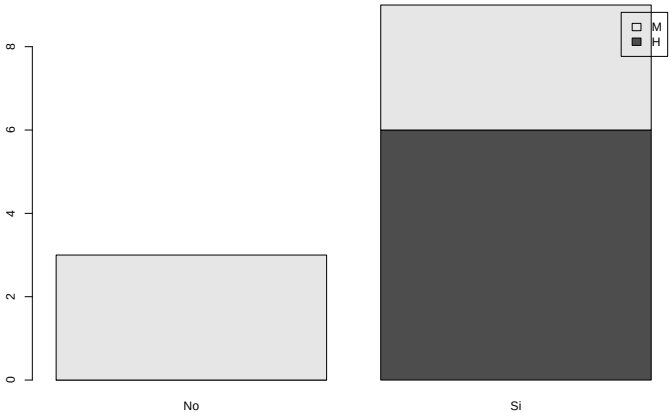


Diagrama de barras - Tabla bidimensional

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

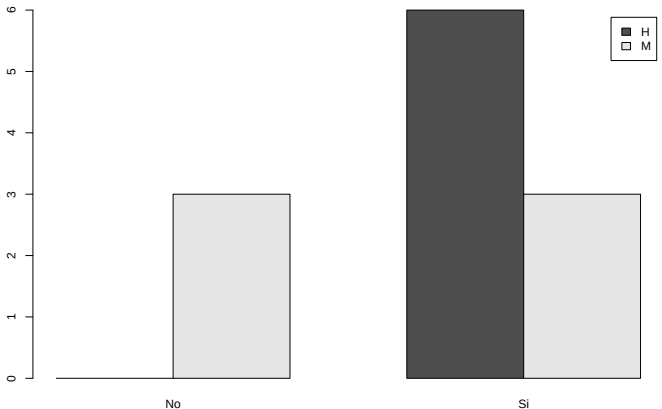


Diagrama de barras - Parámetros de las leyendas

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

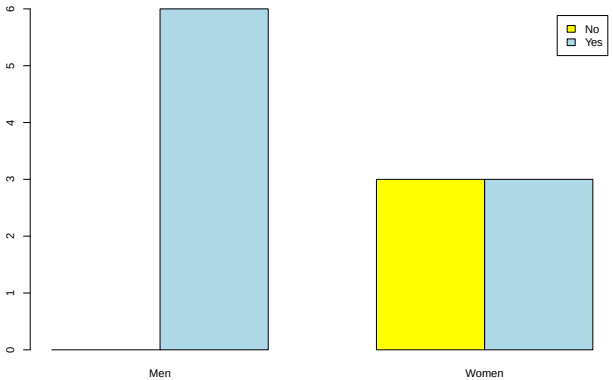


Diagrama circular

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.

Diagrama circular - Parámetros

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)`.

Diagrama circular

```
x
```

```
##  [1] 1 3 5 5 4 3 4 5 5 1 5 4
```

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

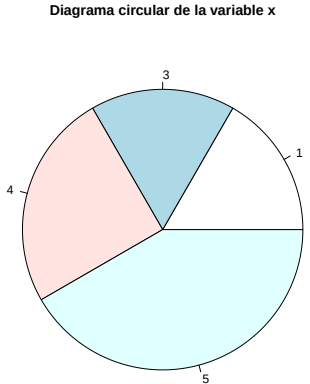


Diagrama circular

```
Respuestas
```

```
##  [1] No Si No Si Si Si Si Si Si Si No Si
## Levels: No Si
```

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

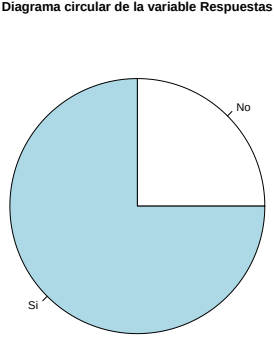


Diagrama circular

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.

Gráficos de mosaico

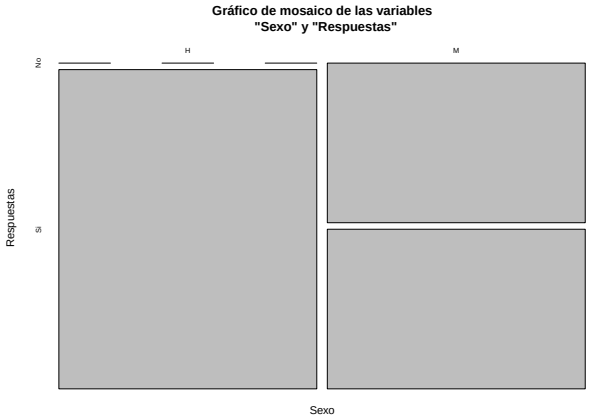
Otra representación de las tablas multidimensionales de frecuencias son los gráficos de mosaico. Estos gráficos se obtienen sustituyendo cada entrada de la tabla de frecuencias por una región rectangular de área proporcional a su valor.

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.

Gráficos de mosaico

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



Gráficos de mosaico

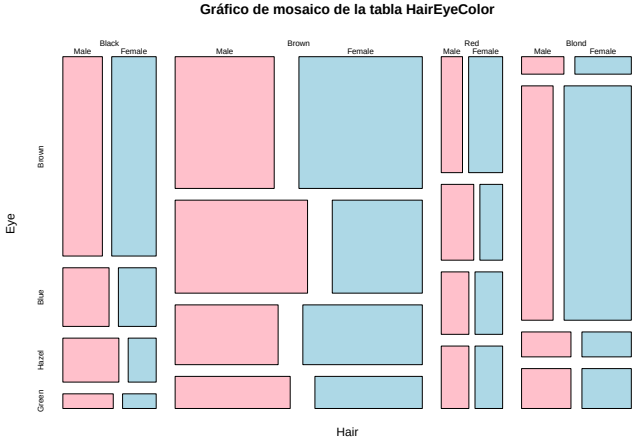
En el gráfico de mosaico de una tabla tridimensional, primero se divide el cuadrado en barras verticales de amplitudes iguales a las frecuencias relativas de una variable.

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.

Gráficos de mosaico

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



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.

Sección 3

Ejemplo final

Un ejemplo final

Vamos a llevar a cabo un análisis completo de un ejemplo con lo que hemos aprendido en esta lección y aprovecharemos para aprender algo nuevo.

El objeto de datos `HairEyeColor` que lleva predefinido R es una tabla de frecuencias absolutas de tres variables cualitativas: color de cabello (`Hair`), color de los ojos (`Eye`) y sexo (`Sex`).

Vamos a extraer de esta tabla una tabla bidimensional de frecuencias absolutas de las variables `Eye` y `Hair`, sin distinguir según el sexo. La manera más sencilla de obtener esta tabla es sumando las subtablas de frecuencias para hombres y mujeres, y aplicando `as.table()` al resultado para transformarlo en una `table` por si no lo es.

Un ejemplo final

Vamos a traducir al castellano los nombres de las variables de esta tabla y de sus niveles. Esto lo podemos llevar a cabo en un solo paso con la función `dimnames()` que ya usamos sobre data frames. El resultado de aplicar esta función a una table es una `list` cuyas componentes son los niveles de cada variable.

```
dimnames(HEC)
```

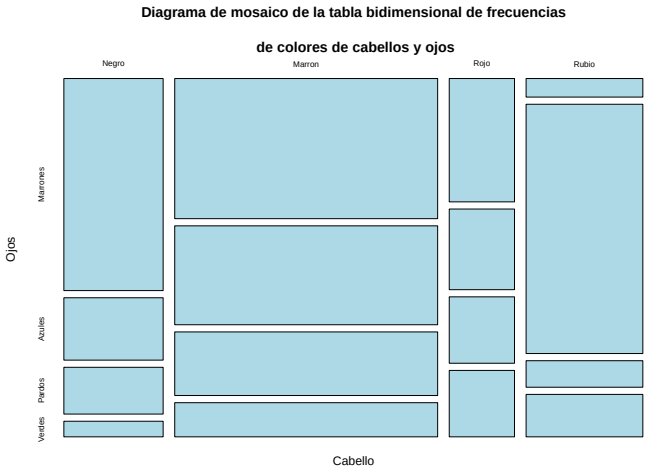
```
## $Hair
## [1] "Black" "Brown" "Red"   "Blond"
##
## $Eye
## [1] "Brown" "Blue"  "Hazel" "Green"
```

Ejercicio.

Redefinid dicha `list` para tener los niveles de los factores en castellano

Un ejemplo final

Vamos a dibujar un diagrama de mosaico de esta tabla, para visualizar gráficamente sus entradas.



Un ejemplo final

A continuación, vamos a calcular el número total de individuos representados en esta tabla:

```
## [1] 592
```

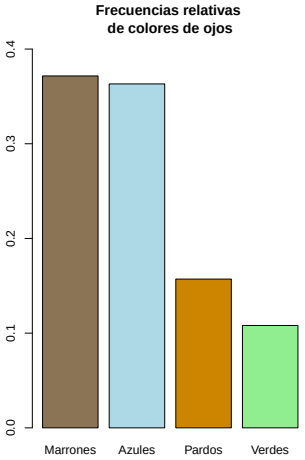
Un ejemplo final

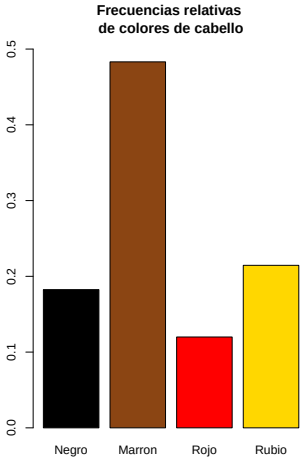
Las tablas de frecuencias absolutas y relativas de cada variable,

```
## Marrones   Azules   Pardos   Verdes
##      220      215       93       64

##  Negro Marron   Rojo  Rubio
##    108    286     71    127

## Marrones   Azules   Pardos   Verdes
##    0.372    0.363    0.157    0.108

##  Negro Marron   Rojo  Rubio
##  0.182  0.483  0.120  0.215
```

Un ejemplo final

Representaremos estas últimas en sendos diagramas de barras.





Un ejemplo final

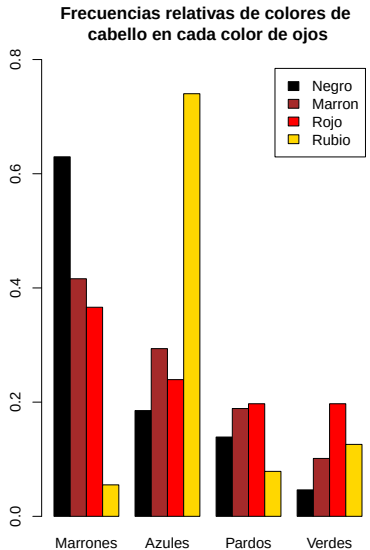
En el diagrama anterior vemos que el color dominante de cabellos es el castaño, mientras que en el color de ojos el marrón y el azul están prácticamente empatados. Pasamos ahora a calcular las tablas de frecuencias relativas y dibujar los dos diagramas de barras de las frecuencias relativas marginales.

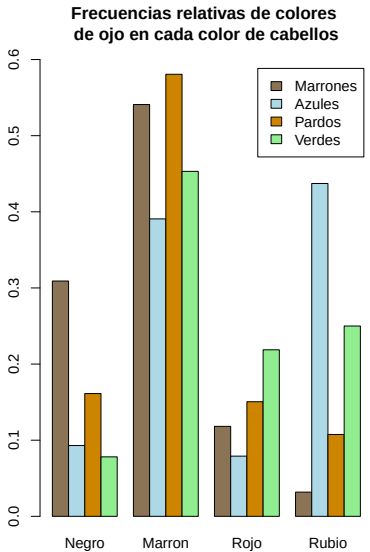
```
##         Ojos
## Cabello  Marrones Azules Pardos Verdes
##   Negro     0.115  0.034  0.025  0.008
##   Marron    0.201  0.142  0.091  0.049
##   Rojo      0.044  0.029  0.024  0.024
##   Rubio     0.012  0.159  0.017  0.027
```

Un ejemplo final

```
##         Ojos
## Cabello  Marrones Azules Pardos Verdes
##   Negro     0.630  0.185  0.139  0.046
##   Marron    0.416  0.294  0.189  0.101
##   Rojo      0.366  0.239  0.197  0.197
##   Rubio     0.055  0.740  0.079  0.126

##         Ojos
## Cabello  Marrones Azules Pardos Verdes
##   Negro     0.309  0.093  0.161  0.078
##   Marron    0.541  0.391  0.581  0.453
##   Rojo      0.118  0.079  0.151  0.219
##   Rubio     0.032  0.437  0.108  0.250
```

Un ejemplo final





Un ejercicio para vosotros

Ejercicio

Instalad y cargad el paquete `MASS`. Encontraréis una tabla de datos llamada `birthwt` sobre factores que pueden incidir en el peso de los niños al nacer. Con `str()` y `head()`, explorad la estructura, y con `help()`, mirad el significado de cada variable.

- Calculad una tabla de frecuencias relativas marginales de los pares (raza de la madre, peso inferior a 2.5 kg o no) que permita ver si la raza de la madre influye en el peso del bebé. Dibujad un diagrama de mosaico de esta tabla.
- Dibujad un diagrama bidimensional de barras, con las barras organizadas en bloques, que permita visualizar esta información. Poned nombres adecuados a los bloques, colores a las barras, y añadid una leyenda que explique qué representa cada barra. ¿Se puede obtener alguna conclusión de esta tabla y de este diagrama de barras?

Un ejercicio para vosotros

- Repetid los dos puntos anteriores para los pares (madre fumadora o no, peso inferior a 2.5 kg o no) y para los pares (madre hipertensa o no, peso inferior a 2.5 kg o no).
- Calculad una tabla de frecuencias relativas marginales de las ternas (raza de la madre, madre fumadora o no, peso inferior a 2.5 kg o no) que permita ver si la raza de la madre y su condición de fumadora o no fumadora influyen en el peso del bebé. Dibujad un diagrama de mosaico de esta tabla.