Estadística Descriptiva Con Datos Cualitativos

Oscar Gerardo Hernández Martínez

26/8/2019

Introducción a la estadística descriptiva

Análisis estadístico de los 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*, si nuestro objetivo es resumir, representar y explicar los datos concretos de los que disponemos. La *estadística descriptiva* 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*.

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

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.

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.

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 de un examen (reprobado, 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.

```
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 ...
```

```
#Observamos que consiste de 4 datos cuantitativos y 1 dato cualitativo.
```

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

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.

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. > (La cualidad es llover o no llover en un día de la semana). > Desde el momento en que me interesa el orden, se vuelve en ordenado.

- *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**).

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

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).

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.

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

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)
#Generamos 12 números entre 1 y 5 aleatoriamente.
#Que se puedan repetir, (de ahí replace = TRUE)
x
```

```
##  [1] 3 3 4 2 3 1 3 5 4 3 4 1
```

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

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

```
#Se convierten a factor para poder trabajar con variables cualitativas.
```

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

```
table(Respuestas)
```

```
## Respuestas
## No Sí
##  3  9
```

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.

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

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

```
## [1] "1" "2" "3" "4" "5"
```

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

```
## [1] "No" "Sí"
```

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.

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

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

```r
table(z)
```

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

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).

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

```
## 3
## 5
```

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

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

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

```
## 5
## 1
```

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

```
## 2
## 3
```

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

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

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

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

```
## Respuestas
##       No       Sí
## 1.732051 3.000000
```

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.

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

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

```
## 0.16666667 0.08333333 0.41666667 0.25000000 0.08333333
```

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

```
## Respuestas
##   No   Sí
## 0.25 0.75
```

¡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] 0.08333333 0.08333333 0.11111111 0.05555556 0.08333333 0.02777778
##  [7] 0.08333333 0.13888889 0.11111111 0.08333333 0.11111111 0.02777778
```

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

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

```
prop.table(X) #R entiende que de los tres niveles
```

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

```
#que tiene mi vector, el 1 sale en la misma proporción
#todo el tiempo; de ahí que el resultado sea 0.33333
#en cada entrada.
```

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          2          3          4          5
## 0.16666667 0.08333333 0.41666667 0.25000000 0.08333333
```

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

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

```
## [1] "2" "5"
```

```
#Esta instrucción nos regresa las etiquetas
#de los elementos que tienen frecuencia
#absoluta 1
```

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] "3"
```

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

```
## [1] "Sí"
```

Ejercicio

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

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

```
HyM = c(rep("H", 6), rep("M", 14))
table(HyM)
```

```
## HyM
##  H  M
##  6 14
```

```
prop.table(table(HyM))
```

```
## HyM
##   H   M
## 0.3 0.7
```

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

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

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
##         Sí 6 3
```

Ejercicio

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

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

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

```r
t(table(Sexo, Respuestas))
```

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

```r
t(table(Respuestas, Sexo))
```

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

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.

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

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

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

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

```r
#Ambas formas son equivalentes
```

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.

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.

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

```r
prop.table(table(Sexo, Respuestas)) #GLOBAL
```

```
##     Respuestas
## Sexo   No   Sí
##    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).

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.

```
#POR SEXO
prop.table(table(Sexo, Respuestas), margin = 1) #Por filas, debe sumar 1
```

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

```
#POR RESPUESTA
prop.table(table(Sexo, Respuestas), margin = 2) #Por columnas, debe sumar 1
```

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

Crosstable

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)`.

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 Sí
##    H  0  6
##    M  3  3
```

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

```
## No Sí
##  3  9
```

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

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

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

```
##   No   Sí
## 0.25 0.75
```

```
#Qué porcentaje respondió que "No" y qué
#porcentaje respondió que "Sí"
rowSums(prop.table(table(Sexo, Respuestas)))
```

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

```
#Qué porcentaje de los encuestados eran
#Hombres y qué porcentaje eran
#Mujeres
```

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.

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)
str(Beb_Energ)
```

```
## 'data.frame':    122 obs. of  3 variables:
##  $ estudio: Factor w/ 4 levels "Industriales",..: 2 3 1 2 1 3 2 1 2 2 ...
##  $ bebe   : Factor w/ 2 levels "No","Si": 1 1 2 2 1 1 2 1 1 1 ...
##  $ sexo   : Factor w/ 2 levels "Hombre","Mujer": 2 1 2 1 2 2 1 1 1 1 ...
```

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

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

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
##  Industriales:37   No:97   Hombre:83
##  Informatica :53   Si:25   Mujer :39
##  Mates       :16
##  Telematica  :16
```

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

```
summary(Beb_Energ)[,2] #Este comando, a pesar de arrojar
```

```
##
## "No:97  " "Si:25  "        NA        NA
```

```
#Algo, no tiene mucho sentido.
```

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

```
apply(Beb_Energ, MARGIN = 2, FUN = table) #Por columnas
```

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

```
#Aplicamos las tres posibles frecuencias absolutas del DF
#Por separado
```

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
```

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
```

O también podemos hacer...

```
table(Beb_Energ[c(1,3)]) #Para primera y tercera columma
```

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

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
```

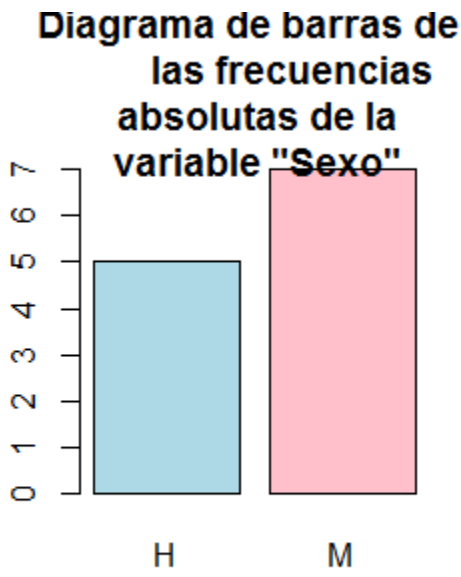
Diagrama de barras

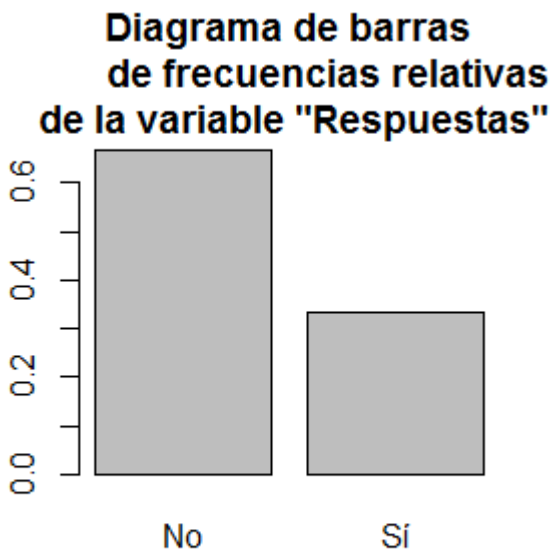
El tipo de gráfico más usado para representar variables cualitativas son los *diagramas de barras (`bar plots`)*. Como su nombre indica, un diagrama de barras contiene, para cada nivel de la variable cualitativa, una barra de altura su frecuencia.

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.

> barplot(table(Sexo), col=c("lightblue","pink"), main="Diagrama de barras de las frecuencias absolutas de la variable"Sexo"")
>
> barplot(prop.table(table(Respuestas)), main="Diagrama de barras de frecuencias relativas de la variable"Respuestas"")
>
> Se inserta una imagen de un posible resultado debido a las complicaciones generadas con el código. (Podemos hacer un salto de renglón dando un intro o con la instrucción "\n")





Parámetros

Observamos 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.

Colores

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.

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.

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)`.

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.

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.

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.

Un ejercicio para practicar

Ejercicio

Instala y carga el paquete `MASS`. Encontrarás 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.

- Calcula 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.
- Dibuja un diagrama bidimensional de barras, con las barras organizadas en bloques, que permita visualizar esta información. Pon nombres adecuados a los bloques, colores a las barras, y añade una leyenda que explique qué representa cada barra. ¿Se puede obtener alguna conclusión de esta tabla y de este diagrama de barras?
- Repite 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).
- Calcula 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. Dibuja un diagrama de mosaico de esta tabla.