Datos cuantitativos agrupados

Ramon Ceballos

31/1/2021

AGRUPACIÓN DE DATOS CUANTITATIVOS (TEORÍA)

1. Introducción

Antes de estudiar unos datos agrupados, hay que, obviamente, agruparlos. Este proceso consta de 4 pasos:

1. Decidir el número de intervalos que vamos a utilizar.
2. Decidir la amplitud de estos intervalos.
3. Acumular los extremos de los intervalos.
4. Calcular el valor representativo de cada intervalo, su **marca de clase**.

No hay una forma de agrupar datos mejor que otra. Eso sí, cada uno de los diferentes agrupamientos para un conjunto de datos podría sacar a la luz características diferentes del conjunto.

2. La función hist()

La función de R por excelencia para estudiar datos agrupados es `hist`. Dicha función implementa los 4 pasos del proceso de agrupación.

Si le indicamos como argumentos el vector de datos y el número de intervalos que deseamos, o bien el método o algoritmo para determinarlo (cosa que veremos a continuación), la función agrupará los datos en el número de clases que le hemos introducido, más o menos. Eso sí, sin control de ningún tipo por nuestra parte sobre los intervalos que produce.

Esto puede venirnos bien en algunos casos, pero no en otros.

3. Pasos del proceso de agrupación

3.1. Establecer el número de clases (k) de la agrupación

En este tema explicaremos una receta para agrupar datos. Lo dicho, ni mejor ni peor que el resto.

Lo primero es establecer el número k de clases en las que vamos a dividir nuestros datos. Podemos decidir en función de nuestros intereses o podemos hacer uso de alguna de las reglas existentes. Destacaremos las más populares. Sea n el número total de datos de la muestra, tenemos:

- **Regla de la raíz cuadrada**: $k = \lceil \sqrt{n} \ \rceil$, del valor de la raíz cuandrada de n toma la parte entera superior.
- **Regla de Sturges**: $k = \lceil 1 + \log_2(n) \rceil$, del valor obtenido toma la parte entera superior.

- **Regla de Scott**: Se determina primero la **amplitud teórica**, A_S de las clases:

$$A_S = 3.5 \cdot \tilde{s} \cdot n^{-\frac{1}{3}}$$

$\tilde{s}$ es la desviación típica muestral. Luego se toma:

$$k = \left\lceil \frac{\max(x) - \min(x)}{A_S} \right\rceil$$

- **Regla de Freedman-Diaconis**: Se determina primero la **amplitud teórica**, A_{FD} de las clases:

$$A_{FD} = 2 \cdot (Q_{0.75} - Q_{0.25}) \cdot n^{-\frac{1}{3}}$$

(donde, recordemos, $Q_{0.75} - Q_{0.25}$, es el rango intercuantílico) y entonces:

$$k = \left\lceil \frac{\max(x) - \min(x)}{A_{FD}} \right\rceil$$

Si os fijáis, las dos primeras reglas solo dependen de n, mientras que las dos últimas reglas también tienen en cuenta, de formas diferentes, la dispersión de los datos (amplitud teórica). De nuevo, no hay ninguna mejor que las demás. Pero sí puede ocurrir que métodos diferentes den lugar a la observación de características diferentes en los datos.

Establecer el número de clases (k) con R

Las instrucciones para llevar a cabo las 3 últimas reglas con R son, respectivamente,

- `nclass.Sturges`
- `nclass.scott`
- `nclass.FD`

Puede ocurrir que las diferentes reglas en R den valores diferentes a aplicar la regla tal cual. Depende del software empleado para el cálculo.

3.2. Decidir la amplitud (A) del agrupamiento

Una vez determinado k, hay que decidir su amplitud.

La forma más fácil y la que nosotros utilizaremos por defecto es que la amplitud de todos los intervalos sea la misma, A. Esta forma no es la única, ya que se podría definir que en los extremos la amplitud fuera mayor respecto a los valores centrales y así.

Para calcular A, lo que haremos será dividir el rango de los datos entre k, el número de clases, y redondearemos por exceso a un valor de la precisión de la medida.

Si se da el improbable caso en que el cociente de exacto, tomaremos como A ese cociente más una unidad de precisión.

3.3. Extremos de los intervalos (Li)

Es la hora de calcular los extremos de los intervalos. Nosotros tomaremos estos intervalos siempre cerrados por su izquierda y abiertos por la derecha, debido a que esta es la forma en que R los construye y porque es así como se utilizan en *Teoría de Probabilidades* al definir la distribución de una variable aleatoria discreta y también en otras muchas situaciones cotidianas.

Utilizaremos la siguiente notación:

$$[L_1, L_2), [L_2, L_3), \ldots, [L_k, L_{k+1})$$

L_i denotan los extremos de los intervalos. Estos se calculan de la siguiente forma:

$$L_1 = \min(x) - \frac{1}{2} \cdot \text{precisión}$$

A partir de L_1, el resto de intervalos se obtiene de forma recursiva:

$$L_2 = L_1 + A$$

$$L_3 = L_2 + A$$

$$\vdots$$

$$L_{k+1} = L_k + A$$

Si nos fijamos bien, los extremos forman una progresión aritmética de salto A:

$$L_i = L_1 + (i-1)A, \qquad i = 2, \ldots, k+1$$

De esta forma garantizamos que los extremos de los intervalos nunca coincidan con valores del conjunto de datos, puesto que tinen una precisión mayor.

3.4. Calcular la marca de clase (Xi)

Solo nos queda determinar la **marca de clase**, X_i, de cada intervalo $[L_i, L_{i+1})$.

Este no es más que un valor del intervalo que utilizaremos para identificar la clase y para calcular algunos estadísticos.

Genralmente,

$$X_i = \frac{L_i + L_{i+1}}{2}$$

es decir, X_i será el punto medio del intervalo, para así garantizar que el error máximo cometido al describir cualquier elemento del intervalo por medio de su marca de clase sea mínimo o igual a la mitad de la amplitud del respectivo intervalo.

Es sencillo concluir que, al tener todos los intervalos amplitud A, la distancia entre X_i **y** X_{i+1} tambien será A. Por consiguiente,

$$X_i = X_1 + (i-1)A, \qquad i = 2, \ldots, k$$

donde $X_1 = \frac{L_1+L_2}{2}$

AGRUPACIÓN DE DATOS CUANTITATIVOS (PRÁCTICA)

Vamos a considerar el conjunto de datos de `datacrab`. Para nuestro estudio, trabajaremos únicamente con la variable `width`.

Llevaremos a cabo los 4 pasos explicados con anterioridad: cálculo del número de intervalos (k), determinación de la amplitud (A), cálculo de los extremos y las marcas de clase(Xi).

En primer lugar, cargamos los datos en un data frame:

```r
crabs = read.table("../../../data/datacrab.txt", header = TRUE)
str(crabs)
```

```
## 'data.frame':    173 obs. of  6 variables:
##  $ input : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ color : int  3 4 2 4 4 3 2 4 3 4 ...
##  $ spine : int  3 3 1 3 3 3 1 2 1 3 ...
##  $ width : num  28.3 22.5 26 24.8 26 23.8 26.5 24.7 23.7 25.6 ...
##  $ satell: int  8 0 9 0 4 0 0 0 0 0 ...
##  $ weight: int  3050 1550 2300 2100 2600 2100 2350 1900 1950 2150 ...
```

A continuación, definimos la variable **cw** que contiene los datos de la variable **width**.

```r
cw = crabs$width
```

Calculemos el **número de clases** según las diferentes reglas que hemos visto:

- *Regla de la raíz cuadrada*

```r
#longitud de la variable cw
n = length(cw)

#ceiling() permite escoger la parte entera superior
k1 = ceiling(sqrt(n))
k1
```

```
## [1] 14
```

- *Regla de Sturges*

```r
k2 = ceiling(1+log(n,2))
k2
```

```
## [1] 9
```

- *Regla de Scott*

```r
#Amplitud teórica
As = 3.5*sd(cw)*n^(-1/3)

#determino rango (max-min) y empleo la amplitud teorica
k3 = ceiling(diff(range(cw))/As)
k3
```

```
## [1] 10
```

- *Regla de Freedman-Diaconis*

```r
#Amplitud teórica
#names=FALSE para que solo de el valor
Afd = 2*(quantile(cw,0.75, names = FALSE)-quantile(cw,0.25,names = FALSE))*n^(-1/3)

#determino rango (max-min) y empleo la amplitud teorica
k4 = ceiling(diff(range(cw))/Afd)
k4
```

```
## [1] 13
```

Podemos comprobar nuestros 3 últimos resultados con R directamente:

```r
#Regla de Sturges
nclass.Sturges(cw)
```

```
## [1] 9
```

```r
#Regla de Scott
nclass.scott(cw)
```

```
## [1] 10
```

```r
#Regla de Freedman-Diaconis
nclass.FD(cw)
```

```
## [1] 13
```

De momento, **vamos a seguir la Regla de Scott**. Es decir, vamos a considerar 10 intervalos.

Regla de Scott

A continuación, debemos elegir la **amplitud de los intervalos**.

```r
A = diff(range(cw)) / 10
A
```

```
## [1] 1.25
```

Como nuestros datos están expresados en mm con una precisión de una cifra decimal, debemos redondear por exceso a un cifra decimal el resultado obtenido. Por lo tanto, nuestra amplitud será de:

```r
A = 1.3
```

Recordad que si el cociente nos hubiera dado un valor exacto con respecto a la precisión, tendríamos que haberle sumado una unidad de precisión.

Esto sirve para que ningún dato de la variable se encuentre en el extremo del intervalo.

Ahora nos toca **calcular los extremos** $L_1, \ldots, L_{11}$ de los intervalos.

Recordad que nuestros intervalos tendrán la siguiente forma (necesitamos 11 valores para 10 intervalos):

$$[L_1, L_2),\ \ldots,\ [L_{10}, L_{11})$$

Calculamos el primer extremo:

```
#la precisión en este caso es 0.1
#Primer extremo del intervalo
L1 = min(cw)-1/2*0.1
L1
```

```
## [1] 20.95
```

El valor 0.1 es nuestra precisión (décimas de unidad, en este caso).

El resto de extremos se calculan del siguiente modo:

```
L2 = L1 + A
L3 = L2 + A
L4 = L3 + A
L5 = L4 + A
L6 = L5 + A
L7 = L6 + A
L8 = L7 + A
L9 = L8 + A
L10 = L9 + A
L11 = L10 + A

#Vector de los extremos de los intervalos
L = c(L1,L2,L3,L4,L5,L6,L7,L8,L9,L10,L11)
L
```

```
##  [1] 20.95 22.25 23.55 24.85 26.15 27.45 28.75 30.05 31.35 32.65 33.95
```

O bien, si queremos facilitarnos el trabajo, también los podemos calcular mucho más rápido del siguiente modo:

```
#Multiplicada la amplitud A por cada uno de los valores de 0 a 10
L = L1 + A*(0:10)

#Vector de los extremos de los intervalos
L
```

```
##  [1] 20.95 22.25 23.55 24.85 26.15 27.45 28.75 30.05 31.35 32.65 33.95
```

Así, nuestros intervalos serán los siguientes:

$$[20.95, 22.25),\ [22.25, 23.55),\ [23.55, 24.85),\ [24.85, 26.15),\ [26.15, 27.45),$$
$$[27.45, 28.75),\ [28.75, 30.05),\ [30.05, 31.35),\ [31.35, 32.65),\ [32.65, 33.95)$$

Y hemos llegado al úlitmo paso: **calcular las marcas de clase**.

Recordemos que $X_i = \frac{L_i + L_{i+1}}{2} \quad \forall i = 1, \ldots, 10.$

Empecemos calculando X_1

```
X1 = (L[1]+L[2])/2
X1
```

```
## [1] 21.6
```

Y, el resto de marcas de clase se calculan del siguiente modo:

```
X2 = X1 + A
X3 = X2 + A
X4 = X3 + A
X5 = X4 + A
X6 = X5 + A
X7 = X6 + A
X8 = X7 + A
X9 = X8 + A
X10 = X9 + A

#Vector de las marcas de clase
X = c(X1,X2,X3,X4,X5,X6,X7,X8,X9,X10)
X
```

```
##  [1] 21.6 22.9 24.2 25.5 26.8 28.1 29.4 30.7 32.0 33.3
```

O bien, si queremos facilitarnos el trabajo, también los podemos calcular mucho más rápido como sucesión:

```
X = X1 + A*(0:9)
X
```

```
##  [1] 21.6 22.9 24.2 25.5 26.8 28.1 29.4 30.7 32.0 33.3
```

O también, como punto medio del intervalo:

```
X = (L[1:length(L)-1]+L[2:length(L)])/2
X
```

```
##  [1] 21.6 22.9 24.2 25.5 26.8 28.1 29.4 30.7 32.0 33.3
```

EJERCICIO

Repetir este proceso para el número de clases obtenido con

- la regla de la raíz
- la regla de Sturges
- la regla de Freedman-Diaconis