Datos Cuantitativos Agrupados

Santiago Pérez Moncada

29/6/2020

Introduccion

Aunque no seamos completamente consientes de ello, tendemos a agrupar datos cuantitativos constantemente.

Sin ir más lejos, calificamos de excelente a todas las notas que están sobre 9. También decimos que una persona tiene 20 años cuando se encuentra en el intervalo [20,21). Es decir, cuando ha cumplido los 20 pero aún no tiene los 21.

En estadística, existen innumerables motivos por los cuales nos interesa agrupar los datos cuando son cuantitativos. Uno de estos motivos puede ser perfectamente que los datos sean muy heterogéneos. En este caso, nos encontrariamos con que las frecuencias de los valores individuales serían todas muy similares, lo que daría lugar a un diagrama de barras muy difícil de interpretar, tal y como mostramos en el siguiente ejemplo.

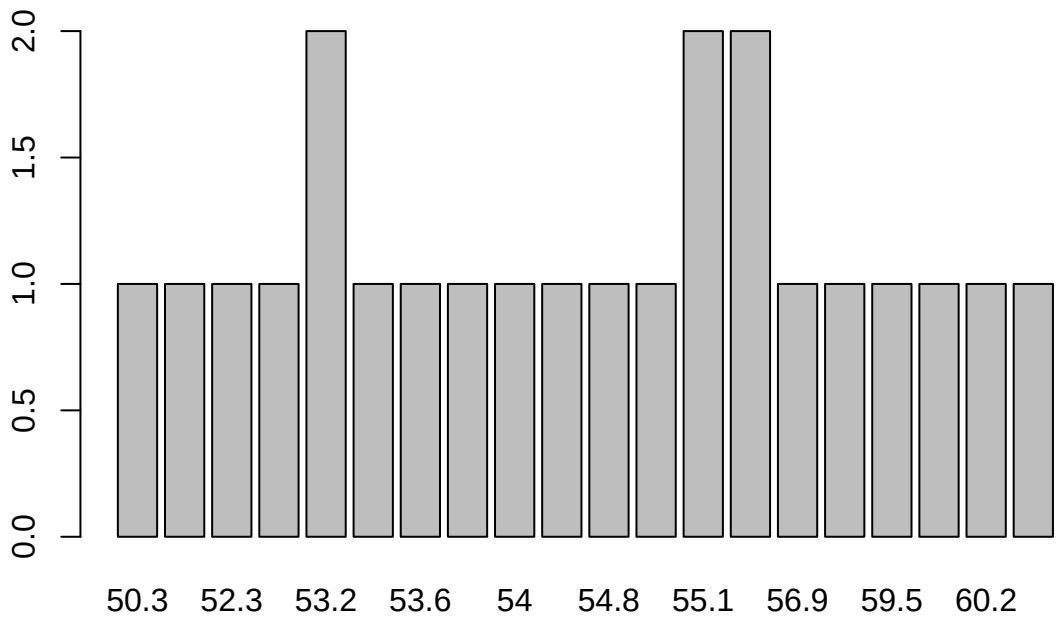
Ejemplo 1

Consideremos la siguiente muestra de 24 pesos de estudiantes:

```
pesos = c(55.2,54.0,55.2,53.7,60.2,53.2,54.6,55.1,53.2,
          54.8,52.3,56.9,57.0,55.0,53.5,50.9,
          55.1,53.6,61.2,59.5,50.3,52.7,60.0)
```

El diagrama de barra de sus frecuencias absolutas, tomando como posibles niveles todos los pesos entre su mínimo y máximo.

```r
barplot(table(pesos))
```



Como vemos, todas las frecuencias se encuentran entre 0,2 cosa que no nos da mucha información.

En cambio, si dividimos todos esos posible valores que puede tomar la variable cuantitativa en intervalos y tomamos como sus frecuencias todos los valores que caen en dicho intervalo, la cosa cambia.

En este caso , seria mucho mas fácil interpretar los resultados, ya que estos daran mucha más información. Más adelante veremos como crear estos intervalos.

Otro de los motivos por el que necesitamos muchas veces agrupar los datos cuantitativos es por que, como ya dijimos en temas anteriores, la precisión infinita no existe.Por tanto, esta imposibilidad de medir de manera exacta muchas de las magnitudes continuas(tiempo,peso,altura...) nos obliga a trabajar con aproximaciones o redondeos de valores reales y cada uno de estos represente todo un intervalo de posibles valores.

Por lo general, existen 3 situaciones en las que conviene sin lugar a dudas agrupar datos cuantitativos en intervalos, tambien llamados clases**.**

- Cuando los datos son continuos, su redondeo ya define un agrupamiento debido a la inexistencia de presición infinita.
- Cuando los datos son discretos, pero con un número considerablemente grande de posibles valores.
- Cuando tenemos muchísimos datos y estamos interesados en estudiar las frecuencias de sus valores.

¿Cómo Agrupar los Datos?

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 los 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.

La Función `hist()`

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

Si le indicamos como argumentos el vector de datos y el número de intervalos que deseamos, o bien el método para determinarlo(cosa que veremos a continuación), la función agrupará los datos en el numero 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.

Estableciendo el número de clases

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 decir en función de nuestros intereses o podemos hacer uso de alguna de las reglas esxistentes. Destacaremos las más populares. Sea n el número total de datos de la muestra.

- **Regla de la raiz cuadrada:** $k = \lceil \sqrt{n} \rceil$
- **Regla de Sturges:** $k = \lceil 1 + \log_2(n) \rceil$
- **Regla de Scott:** Se determina primero la **Amplitud teórica**, A_S de las clases.

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

Donde $\tilde{s}$ 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 intercuartilico) y entonces.

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

Si nos fijamos, las dos primera solo dependen de n, miestras que las dos últimas también tienen en cuenta, de formas diferentes, la dispersión de los datos. 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.

Estableciendo el número de clases 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 den valores diferentes, o no.

Decidiendo la amplitud.

Una vez determinadas las clases 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.

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

$$A = \frac{\max(x) - \min(x)}{k}$$

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

Extremos de los intervalos

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

Utilizaremos la siguiente notación

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

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

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

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

$$\begin{aligned} L_2 &= L_1 + A \\ L_3 &= L_2 + A \\ &\vdots \\ L_{k+1} &= L_k + A \end{aligned}$$

Si nos fijamos bien, los extremos forman una progresión aritmetica 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 tienen una precisión mayor.

Marca de Clase

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 calcular algunos estadisticos.

Generalmente,

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

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

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

IMPLEMENTACIÓN DE LAS REGLAS DE AGRUPACIÓN

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

Llevaremos a cado los 4 pasos explicados con anterioridad:cálculo del número de intervalos, determinación de la amplitud, cálculo de los extremos y las marcas de clase.

Solución

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

```
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**.

```
cw = crabs$width
```

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

- **Regla de la raiz cuadrada**

```
n = length(cw)
k1 = ceiling(sqrt(n))
k1
```

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

- **Regla de Sturges**

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

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

- **Regla de Scott**

```
As = 3.5*sd(cw)*n^-(1/3)
k3 = ceiling((max(cw)-min(cw))/As) # ceiling(diff(range(cw))/As)
k3
```

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

- **Regla de Freedman-Diaconis**

```
Afd = 2*(quantile(cw,0.75,names = FALSE)-quantile(cw,0.25,name=FALSE))*n^(-1/3)
Afd = 2*IQR(cw)*n^(-1/3)
k4 = ceiling((max(cw)-min(cw))/Afd)
k4
```

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

Podemos comprobar nuestros 3 últimos resultados con R:

```
nclass.Sturges(cw)
```

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

```
nclass.scott(cw)
```

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

```
nclass.FD(cw)
```

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

USANDO REGLA DE SCOTT

De momento, vamos a elegir la regla de Scott.Es decir, vamos a considerar 10 intervalos.

A continuación, debemos elegir la amplitud de los intervalos.

```
K_Scott = k3 # 10
A = (max(cw)-min(cw))/K_Scott
A
```

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

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

```
A = 1.3
```

Recordemos que si el cociente nos hubiera dado exacto con respecto a la precisión, tendriamos que haberle sumado una unidad de precisión.

Ahora nos toca calcular los extremos $L_1, ..., L_{11}$ de los intervalos. Recordemos que nuestros intervalos tendrán la siguiente forma:

$$[L_1, L_2), [L_2, L_3), \ldots, [L_{10}, L_{11})$$

Calculamos el primer extremo:

```
L1 = min(cw)-(1/2)*0.1
L1
```

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

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

Y el resto de los 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

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, tambien podemos calcular mucho más rapido del siguiente modo.

```
L = L1 + A*(0:10)
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:

$$[20.95, 22.25), [22.25, 23.55), [23.55, 24.85), [24.85, 26.15), [26.15, 27.45),$$
$$[27.45, 28.75), [28.75, 30.05), [30.05, 31.35), [31.35, 32.65), [32.65, 33.9)$$

Y hemos llegado al último paso: Calcular las marcas de clase Recordemos que $X_i = \frac{L_i + L_{i+1}}{2} \qquad \forall i = 1, ..., 10$

Empecemos calculando X_i

```
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

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, tambien los podemos calcular mucho más rapido 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 tambien, 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
```

USANDO REGLA DE LA RAIZ

Primero determinamos el numero de clases.

```
n = length(cw)
k = ceiling(sqrt(n))
k
```

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

Ahora vamos a determinar la amplitud de cada clase

```
A = (max(cw)-min(cw))/k
A
```

```
## [1] 0.8928571
```

```
A = 0.9
precision = 0.1
```

Seguido de esto calculamos los extremos de los intervalos

```
L1 = min(cw) - (1/2)*precision
L = L1 + (0:15)*A
L
```

```
##  [1] 20.95 21.85 22.75 23.65 24.55 25.45 26.35 27.25 28.15 29.05 29.95 30.85
## [13] 31.75 32.65 33.55 34.45
```

Por ultimo calculamos las marcas de clase de los intervalos

```
X1 = (L[1]+L[2])/2
X = X1 + (0:14)*A
X
```

```
##  [1] 21.4 22.3 23.2 24.1 25.0 25.9 26.8 27.7 28.6 29.5 30.4 31.3 32.2 33.1 34.0
```

USANDO REGLA DE STURGES

Primero determinamos el numero de clases.

```
n = length(cw)
k = ceiling(1+log2(n))
k
```

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

Ahora vamos a determinar la amplitud de cada clase

```
A = (max(cw)-min(cw))/k
A
```

```
## [1] 1.388889
```

```
A = 1.4
precision = 0.1
```

Seguido de esto calculamos los extremos de los intervalos

```
L1 = min(cw) - (1/2)*precision
L = L1 + (0:10)*A
L
```

```
##  [1] 20.95 22.35 23.75 25.15 26.55 27.95 29.35 30.75 32.15 33.55 34.95
```

Por ultimo calculamos las marcas de clase de los intervalos

```
X1 = (L[1]+L[2])/2
X = X1 + (0:9)*A
X
```

```
##  [1] 21.65 23.05 24.45 25.85 27.25 28.65 30.05 31.45 32.85 34.25
```

USANDO REGLA DE FREEDMAN-DIACONIS

Primero determinamos el numero de clases.

```
n = length(cw)
Afd = 2*IQR(cw)*n^(-1/3)
k = ceiling((max(cw)-min(cw))/Afd)
k
```

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

Ahora vamos a determinar la amplitud de cada clase

```
A = (max(cw)-min(cw))/k
A
```

```
## [1] 0.9615385
```

```
A = 1.0
precision = 0.1
```

Seguido de esto calculamos los extremos de los intervalos

```
L1 = min(cw) - (1/2)*precision
L = L1 + (0:14)*A
L
```

```
##  [1] 20.95 21.95 22.95 23.95 24.95 25.95 26.95 27.95 28.95 29.95 30.95 31.95
## [13] 32.95 33.95 34.95
```

Por ultimo calculamos las marcas de clase de los intervalos

```
X1 = (L[1]+L[2])/2
X = X1 + (0:13)*A
X
```

```
##  [1] 21.45 22.45 23.45 24.45 25.45 26.45 27.45 28.45 29.45 30.45 31.45 32.45
## [13] 33.45 34.45
```