08-Agrupacion De Datos

Adrian

18/1/2022

Agrupacion de datos

Los 4 pasos

1. Decidir el numero 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.

Funcion hist()

Funcion por excelencia en R para estudiar datos agrupados. La funcion implementa los 4 pasos del proceso.

Reglas para establecer el numero de clases

Lo primero es establecer el numero k de clases en las que vamos a dividir nuestros datos. Podemos hacerlo en funcion de nuestros intereses o podemos hacer uso de alguna de las reglas existentes.

- Regla de la raiz cuadrada: $k = \lceil \sqrt{n} \rceil$ = Tomar la parte entera superior de la raiz cuadrada, si ha dado 7.4 cogemos 8
- Regla de Sturges: $k = \lceil 1 + \log_2(n) \rceil$ En R se utiliza –> nclass.Surges
- Regla de Scott: Se determina primero la amplitud teorica. $A_S = 3.5 \cdot \tilde{S} \cdot n^{-\frac{1}{3}}$ donde $\tilde{S}$ es la desviacion tipica muestral. Luego se toma $k = \lceil \frac{max(x) - min(x)}{A_S} \rceil$ En R se utiliza –> nclass.scott
- Regla de Freedman-Diaconis: Se determina primero la amplitud teorica. $A_{FD} = 2 \cdot (Q_{0.75} - Q_{0.25}) \cdot n^{-\frac{1}{3}}$ Donde $Q_{0.75} - Q_{0.25}$ es el rango intercuantilico y entonces $k = \lceil \frac{max(x) - min(x)}{A_{FD}} \rceil$ En R se utiliza –> nclass.FD

Extremos de los intervalos

Se utiliza la notacion $[L_1, L_2), [L_2, L_3)...$ Donde: $L_1 = min(x) - \frac{1}{2} \cdot precision$ El resto se obtiene de forma recursiva $L_2 = L_1 + A$

Los extremos forman una progresion aritmetica de salto A. $L_i = L_1 + (i-1) \cdot A \; i = 2, ..., k+1$

Marca de clase

Es un valor del intervalo que utilizaremos para identificar la clase y para calcular algunos estadisticos. $x_i = \frac{L_i + L_{i+1}}{2}$

Ejemplo

```
crabs = read.table("../../../data/datacrab.txt", header = T)
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 ...
```

```
# Obtener la columna de anchura
cw = crabs$width
# Obtener la longitud de cw
n = length(cw)

# Regla de la raiz cuadrada
k1 = ceiling(sqrt(n))
k1
```

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

```
# Regla de Sturges
k2 = ceiling(1+log(n,2))
k2
```

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

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

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

```
# Regla de Scott
## Amplitud teorica
AS = 3.5*sd(cw)*n^(-1/3)
k3 = ceiling(diff(range(cw))/AS)
k3
```

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

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

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

```r
# Regla de Freedman-Diaconis
## Amplitud teorica
Afd = 2*(quantile(cw,0.75, names = F)-quantile(cw,0.25, names = F))*n^(-1/3)
k4 = ceiling(diff(range(cw))/Afd)
k4
```

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

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

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

Segun la regla de Scott tendriamos que crear 10 intervalos.

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

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

```r
# El resultado es 1.25 pero todos nuestros valores tienen 1 solo decimal por lo que redondeamos A a la
A = 1.3
```

```r
# Calculamos los extremos, como necesitamos 10 intervalos necesitamos 11 extremos
L1 = min(cw)-1/2*0.1
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
```

```r
# Tambien lo podemos hacer de la siguiente forma
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
```

```r
# Marcas de clase: SON LOS PUNTOS MEDIOS ENTRE CADA INTERVALO
X1 = (L[1]+L[2])/2
X1
```

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

```r
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
```

```r
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
```