Agrupacion de datos

Victor Lopez

2023-01-25

¿Cuando es necesario agrupar los datos?

- Cuando los datos son heterogeneos. Debido a que nos encontrariamos con que las frecuencias de los valores serian muy similares, lo que daria un diagrama de barras dificil de interpretar
- Cuando los datos son continuos, su redondeo ya define un agrupamiendo debido a la inexistencia de la precision infinita
- Cuando los datos son discretos pero con un numero considerablemente grande de posibles valores
- Cuando tenemos muchisimos datos y queremos estudiar las frecuencias de sus valores

El proceso de agrupamiento de los datos

1. Decidir el numero de intervalos que vamos a utilizar

- Lo primero es establecer el número k de clases en las que vamos a dividir nuestros datos.
- Regla de la raíz 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 intercuantílico) y entonces

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

```
nclass.Sturges(vector)
nclass.scott(vector)
nclass.FD(vector)
# A veces no dan los mismos resultados que calcularlos manualmente
```

2. Decidir la amplitud de estos intervalos

- A = Amplitud
- A = Rango / k
- Redondearemos por exceso a un valor de la precision de la medida.
- El valor de precision de la medida se define por la cantidad de decimales, es decir, si los datos tienen un decimal y A = 1.25, se debe redondear a 1.3.
- En el caso improbable de que A nos salga directamente con el valor de precision de la medida, debemos sumarle una unidad de precision

3. Acumular los extremos de los intervalos

- 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 (para que los valores nunca toquen los extremos de las clases):

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

  Esto para que el nivel de precision en los intervalos sea mas alto, y que de esta manera los datos nunca toquen los extremos de los intervalos

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

$$L_2 = L_1 + A$$

$$L_3 = L_2 + A$$

$$\vdots$$

Nota: Si los datos tocan los extremos de los intervalos, se crea una ambigüedad en la interpretación de los datos, ya que no se sabe a qué intervalo pertenecen esos datos. Es por eso que se recomienda dejar un espacio vacío entre los datos y los extremos de los intervalos, conocido como "márgenes de clase" o "márgenes de intervalo". Esto permite una mayor claridad en la interpretación de los datos y una distribución más precisa de los datos dentro de cada intervalo.

4. Calcular la marca de clase, es decir el valor representativo

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.

Nota: no hay una forma de agrupar datos mejor que otra. Sin embargo una puede dar diferentes resultados a otra

Nota: 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 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.

Ejemplo en 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 ...
```

```
cw = crabs$width
```

- Regla de la raíz cuadrada:

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

- Regla de Sturges:

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

nclass.Sturges(cw)
```

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

- Regla de Scott:

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

nclass.scott(cw)
```

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

- Regla de Freedman-Diaconis:

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

nclass.FD(cw)
```

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

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

```r
A = diff(range(cw)) / 10 # 1.25
A = 1.3
```

```r
# Primer extremo:
L1 = min(cw)-1/2*0.1

# Los demas extremos
L = L1 + A*(0:10) # Para 10 intervalos necesitamos 11 extremos, por eso del 0 al 10
```

```r
# Marcas de clase
X1 = (L[1]+L[2])/2

X = X1 + A*(0:9) # Para 10 intervalos necesitamos 10 marcas de clase, por eso del 0 al 9

X = (L[1:length(L)-1]+L[2:length(L)])/2 # Forma alternativa. Cuestion de gustos.
```