Estadísticos para Datos Agrupados

Santiago Pérez Moncada

13/7/2020

Al tener una muestra de datos numéricos, conviene calcular los **estadísticos** antes de realizar los agrupamientos, puesto de los contrario podemos perder información.

No obstante, hay situaciones en que los datos los obtenemos ya agrupados.En estos casos, aún sigue siendo posible calcular estadísticos y utilizarlos como aproximaciones de los datos "reales", los cuales no conocemos.

Estadísticos para datos agrupados

La media $\bar{x}$, la varianza s^2, la varianza muestral $\tilde{s}^2$, la desviación típica s y la desviación típica muestral $\tilde{s}$ de un conjunto de datos agrupados se calculan mediante las mismas fórmulas que para los datos no agrupados la única diferencia es que sustituimos cada clase por su marca de clase y la contamos con su frecuencia.

Es decir si tenemos k clases, con sus respectivas marcas $X_1, ..., X_k$ con frecuencias absolutas $n_1, ..., n_k$ de forma que $n = \sum_{j=1}^{k} n_j$ entonces.

$$\bar{x} = \frac{\sum_{j=1}^{k} n_j X_j}{n}$$

$$s^2 = \frac{\sum_{j=1}^{k} n_j X_j^2}{n} - \bar{x}^2$$

$$\tilde{s}^2 = \frac{n}{n-1} \cdot s^2$$

$$s = \sqrt{s^2}, \qquad \tilde{s} = \sqrt{\tilde{s}^2}$$

Intervalo Modal

En lo referente a la moda, esta se sustituye por el **intervalo modal**, que es la clase con mayor frecuencia(absoluta o relativa).

En el caso en que un valor numérico fuera necesario, se tomaría su marca de clase.

Intervalo crítico para la mediana.

Se conoce como **intervalo crítico para la mediana**, $[L_c, L_{c+1})$ al primer intervalo donde la frecuencia relativa acumulada sea mayor o igual que 0.5

Denotemos por n_c su frecuencia absoluta, por $A_c = L_{c+1} - L_c$ su amplitud y por N_{c-1} la frecuencia acumulada del intervalo inmediatemente anterior(en caso de ser $[L_c, L_{c+1}) = [L_1, L_2)$, entonces $N_{c-1} = 0$). Entonces, M será una aproximacion para la mediana de los datos "reales" a partir de los datos agrupados.

$$M = L_c + A_c \cdot \frac{\frac{n}{2} - N_{c-1}}{n_c}$$

,

Aproximación de los Cunatiles

La fórmula anterior nos permite aproximar el cuantil Q_p de los datos reales apartir de los datos agrupados.

$$Q_p = L_p + A_p \cdot \frac{p \cdot n - N_{p-1}}{n_p}$$

donde el intervalo $[L_p, L_{p+1})$ denota el primer intervalo cuya frecuencia relativa acumulada es mayor o igual a p.

Ejemplo

```
TablaFrecs1 = function(x,k,A,p){
  L = min(x)-p/2+A*(0:k)
  x_cut = cut(x, breaks = L, right = FALSE)
  intervals = levels(x_cut)
  mc = (L[1]+L[2])/2 +A *(0:(k-1))
  Fr.abs = as.vector(table(x_cut))
  Fr.rel = round(Fr.abs/length(x),4)
  Fr.cum.abs = cumsum(Fr.abs)
  Fr.cum.rel = cumsum(Fr.rel)
  tabla = data.frame(intervals, mc, Fr.abs,Fr.cum.abs,Fr.rel,Fr.cum.rel)
  tabla
}


data = read.table("../data/datacrab.txt",header = TRUE)
CW = data$width
k = nclass.Sturges(CW)
A = (max(CW)-min(CW))/k
A = 1.5
p = 0.1


tabla = TablaFrecs1(CW,k,A,p)
tabla
```

```
##     intervals   mc Fr.abs Fr.cum.abs Fr.rel Fr.cum.rel
## 1 [20.9,22.4) 21.7      2          2 0.0116     0.0116
## 2 [22.4,23.9) 23.2     21         23 0.1214     0.1330
## 3 [23.9,25.4) 24.7     37         60 0.2139     0.3469
## 4 [25.4,26.9) 26.2     48        108 0.2775     0.6244
## 5 [26.9,28.4) 27.7     38        146 0.2197     0.8441
## 6 [28.4,29.9) 29.2     18        164 0.1040     0.9481
## 7 [29.9,31.4) 30.7      6        170 0.0347     0.9828
## 8   [31.4,33) 32.2      2        172 0.0116     0.9944
## 9   [33,34.5) 33.7      1        173 0.0058     1.0002
```

```r
TotalObs = tabla$Fr.cum.abs[9]
TotalObs
```

```
## [1] 173
```

```r
anchura.media = round(sum(tabla$Fr.abs * tabla$mc)/TotalObs ,3)
anchura.media
```

```
## [1] 26.373
```

```r
anchura.var = round(sum(tabla$Fr.abs*tabla$mc^2)/TotalObs - anchura.media^2,3)
anchura.var
```

```
## [1] 4.674
```

```r
anchura.std = round(sqrt(anchura.var),3)
anchura.std
```

```
## [1] 2.162
```

```r
I.modal = tabla$intervals[which(tabla$Fr.abs == max(tabla$Fr.abs))]
I.modal
```

```
## [1] "[25.4,26.9)"
```

Por lo tanto, con los datos de los que disponemos, podemos afirmar que la anchura media de los cangrejos de la muestra es de $\bar{x} = 26.373$mm , con una desviacion típica $s = 4.674$mm y que el grupo de anchuras más numeroso era el de $[25.4, 26.9)$.

Pasemos ahora a calcular el intervalo crítico para la mediana.

```r
I.critic = tabla$intervals[which(tabla$Fr.cum.rel >= 0.5)]
I.critic[1]
```

```
## [1] "[25.4,26.9)"
```