Datos cuantitativos agrupados

Ramon Ceballos

3/2/2021

PARTE TEÓRICA

1. Estadísticos para datos agrupados

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

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

1.1 Media, varianza y desviación típica

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 con la única diferencia de 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 $X_1, \ldots, X_k$ con frecuencias absolutas $n_1, \ldots, n_k$ de forma que $n = \sum_{j=1}^{k} n_j$. Entonces:

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

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

1.2 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, tanto da).

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

1.3 Intervalo crítico para la mediana

La mediana como tal no existe y en su lugar se utiliza el ntervalo 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 la frecuencia absoluta del intervalo crítico, por $A_c = L_{c+1} - L_c$ su amplitud y por N_{c-1} la frecuencia acumulada del intervalo inmediantamente anterior (en caso de ser $[L_c, L_{c+1}) = [L_1, L_2)$, entonces $N_{c-1} = 0$). Entonces, la mediana M será una aproximación para la mediana de los datos "reales" a partir de los agrupados:

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

1.4 Aproximación de los cuantiles

La fórmula anterior nos permite **aproximar el cuantil** Q_p de los datos "reales" a partir de los datos agrupados, empleando la siguiente ecuación:

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

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

PARTE PRÁCTICA

Ejemplo 1. Anchura de Cangrejos

Vamos a seguir trabajando con nuestra variable `cw` y, esta vez, lo que haremos será calcular los estadísticos de la variable con los datos agrupados y, para acabar, estimaremos la mediana y algunos cuantiles.

Recordemos todo lo que habíamos obtenido sobre nuestra variable `cw`:

```
#Variable cuantitativa de trabajo
crabs = read.table("../../../../data/datacrab.txt", header = TRUE)
cw = crabs$width
```

```
#Amplitud de cada intervalo
A = 1.3
```

```
#Extremo 1 de los intervalos
L1 = min(cw)-1/2*0.1
```

Y, 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
```

```r
#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
```

```r
#Guardo los intervalos en una variable
intervals = as.character(c("[20.95,22.25)","[22.25,23.55)","[23.55,24.85)","[24.85,26.15)","[26.15,27.45
#Defino una tabla con las frecuencias del agrupamiento de la variable
TF.L = function(x,L,V){
  x_cut = cut(x, breaks=L, right=FALSE, include.lowest=V)
  mc = (L[1:(length(L)-1)]+L[2:length(L)])/2
  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
  }
tabla = TF.L(cw,L,FALSE)
tabla
```

```
##        intervals   mc Fr.abs Fr.cum.abs Fr.rel Fr.cum.rel
## 1  [20.95,22.25) 21.6      2          2 0.0116     0.0116
## 2  [22.25,23.55) 22.9     14         16 0.0809     0.0925
## 3  [23.55,24.85) 24.2     27         43 0.1561     0.2486
## 4  [24.85,26.15) 25.5     44         87 0.2543     0.5029
## 5  [26.15,27.45) 26.8     34        121 0.1965     0.6994
## 6  [27.45,28.75) 28.1     31        152 0.1792     0.8786
## 7  [28.75,30.05) 29.4     15        167 0.0867     0.9653
## 8  [30.05,31.35) 30.7      3        170 0.0173     0.9826
## 9  [31.35,32.65) 32.0      2        172 0.0116     0.9942
## 10 [32.65,33.95) 33.3      1        173 0.0058     1.0000
```

Con toda la información anterior definida, podemos calcular los **estadísticos** del agrupamiento de la varibale cuantitativa.

```r
#Nº total de individuos estudiados en la variable cuantitativa de trabajo
TOT = tabla$Fr.cum.abs[10]
TOT
```

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

```r
#Media del agrupamiento
anchura.media = round(sum(tabla$Fr.abs*tabla$mc)/TOT,3)
anchura.media
```

```
## [1] 26.312
```

```r
#Varianza del agrupamiento
anchura.var = round(sum(tabla$Fr.abs*tabla$mc^2)/TOT-anchura.media^2,3)
anchura.var
```

```
## [1] 4.476
```

```r
#Desviación típica del agrupamiento
anchura.dt = round(sqrt(anchura.var),3)
anchura.dt
```

```
## [1] 2.116
```

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

#Da el intefvalo en que la Frec. Abs. es máxima
I.modal
```

```
## [1] [24.85,26.15)
## 10 Levels: [20.95,22.25) [22.25,23.55) [23.55,24.85) ... [32.65,33.95)
```

Por lo tanto, con los datos de los que disponemos, podemos afirmar que la anchura media de los cangrejos de la muestra es de 26.312 mm, con una desviación típica de unos 4.476 mm, y que el grupo de anchuras más numeroso era el de [24.85,26.15).

Pasemos ahora a calcular el **intervalo crítico para la mediana**.

```r
#Intervalo critico para la mediana
I.critic = tabla$intervals[which(tabla$Fr.cum.rel >= 0.5)]

#Da la primera posición mayor o iguar a 0.5 para Frec. rel.acumulada
#Es necesario porque "which" devuelve todos los datos
I.critic[1]
```

```
## [1] [24.85,26.15)
## 10 Levels: [20.95,22.25) [22.25,23.55) [23.55,24.85) ... [32.65,33.95)
```

Ahora, ya podemos calcular una **estimación de la mediana** de los datos "reales".

```r
#Nº total de individuos del estudio
n = TOT

#Cuarto extremo de los intervalos que se corresponde con
#el extremo inferior del intervalo crítico para la mediana
Lc = L[4]

#el extremo superior del intervalo crítico para la mediana
Lc.pos = L[5]

#Amplitud del intervalo crítico para la mediana
Ac = L[5]-L[4]
```

```r
#Frecuencia absoluta acumulada del intervalo anterior
Nc.ant = tabla$Fr.cum.abs[3]

#Frecuencia absoluta del intervalo actual conocido como
#intervalo crítico para la mediana
nc = tabla$Fr.abs[4]

#Aproximación de la mediana de los datos "reales"
M = Lc+Ac*((n/2)-Nc.ant)/nc
M
```

```
## [1] 26.13523
```

Si lo comparamos con la mediana real de los datos de la variable de estudio, observamos que se aproxima mucho.

```r
#Mediana de los datos "reales"
median(cw)
```

```
## [1] 26.1
```

También podemos hacer **aproximaciones de los cuantiles**. Hemos creado una función `aprox.quantile.p` para no tener que copiar la operación cada vez que queramos calcular un cuantil aproximado.

```r
# Minifunción para calculo de aproximaciones de los cuantiles
aprox.quantile.p = function(Lcrit,Acrit,n,p,Ncrit.ant,ncrit){
  round(Lcrit+Acrit*(p*n-Ncrit.ant)/ncrit,3)
}

#Primer cuartil
aprox.quantile.p(Lc,Ac,n,0.25,Nc.ant,nc)
```

```
## [1] 24.857
```

```r
#Tercer cuartil
aprox.quantile.p(Lc,Ac,n,0.75,Nc.ant,nc)
```

```
## [1] 27.413
```

Y ahora, calculemos los cuartiles de los datos “reales”

```r
#Primer cuartil
quantile(cw,0.25)
```

```
##  25%
## 24.9
```

```r
#Tercer cuartil
quantile(cw,0.75)
```

```
##  75%
## 27.7
```

Ejemplo 2. Notas de Bachillerato

```r
set.seed(144)
notas = sample(0:10,100, replace = TRUE)
set.seed(NULL)
```

```r
#Definimos vector de extremos de los intervalos
L = c(0,5,7,9,10)

#Definimos notas4 como el resultado de la codificación en intervalos
#utilizando como etiquetas Susp, Aprob, Not y Exc
notas4 = cut(
  notas,
  breaks = L,
  labels = c("Susp", "Aprob", "Not", "Exc"),
  right = FALSE,
  include.lowest = TRUE)
```

```r
notasHist = hist(
  notas,
  breaks = L,
  right = FALSE,
  include.lowest = TRUE,
  plot = FALSE)


FAbs = notasHist$count #Frecuencias absolutas

FRel = prop.table(FAbs)

FAbsCum = cumsum(FAbs)

FRelCum = cumsum(FRel)
```

```r
#Segunda función
TablaFrecs.L = function(x,L,V){
  x_cut = cut(x, breaks=L, right=FALSE, include.lowest=V)
  intervals = levels(x_cut)
  mc = (L[1:(length(L)-1)]+L[2:length(L)])/2
  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
  }
```

```
estudio = TablaFrecs.L(notas, L, TRUE)
estudio
```

```
##   intervals  mc Fr.abs Fr.cum.abs Fr.rel Fr.cum.rel
## 1     [0,5) 2.5     46         46   0.46       0.46
## 2     [5,7) 6.0     16         62   0.16       0.62
## 3     [7,9) 8.0     19         81   0.19       0.81
## 4    [9,10] 9.5     19        100   0.19       1.00
```

Con toda la información anterior definida, podemos calcular los **estadísticos** del agrupamiento de la varibale cuantitativa.

```
#Nº total de individuos estudiados en la variable cuantitativa de trabajo
n_total = estudio$Fr.cum.abs[4]

#Media del agrupamiento
notas.media = round(sum(estudio$Fr.abs*estudio$mc)/n_total,3)
notas.media
```

```
## [1] 5.435
```

```
#Varianza del agrupamiento
notas.var = round(sum(estudio$Fr.abs*estudio$mc^2)/n_total-notas.media^2,3)
notas.var
```

```
## [1] 8.403
```

```
#Desviación típica del agrupamiento
notas.dt = round(sqrt(notas.var),3)
notas.dt
```

```
## [1] 2.899
```

```
#Intervalo modal del agrupamiento
notas_I.modal = estudio$intervals[which (estudio$Fr.abs == max(estudio$Fr.abs))]

#Da el intefvalo en que la Frec. Abs. es máxima
notas_I.modal
```

```
## [1] [0,5)
## Levels: [0,5) [5,7) [7,9) [9,10]
```

Pasemos ahora a calcular el **intervalo crítico para la mediana**.

```
#Intervalo critico para la mediana
notas_I.critic = estudio$intervals[which(estudio$Fr.cum.rel >= 0.5)]

#Da la primera posición mayor o iguar a 0.5 para Frec. rel.acumulada
#Es necesario porque "which" devuelve todos los datos
notas_I.critic[1]
```

```
## [1] [5,7)
## Levels: [0,5) [5,7) [7,9) [9,10]
```

Ahora, ya podemos calcular una **estimación de la mediana** de los datos "reales".

```
#el extremo inferior del intervalo crítico para la mediana
Lc = L[2]


#el extremo superior del intervalo crítico para la mediana
Lc.pos = L[3]

#Amplitud del intervalo crítico para la mediana
Ac = L[3]-L[2]

#Frecuencia absoluta acumulada del intervalo anterior
Nc.ant = estudio$Fr.cum.abs[1]

#Frecuencia absoluta del intervalo actual conocido como
#intervalo crítico para la mediana
nc = estudio$Fr.abs[2]

#Aproximación de la mediana de los datos "reales"
M.notas = Lc+Ac*((n_total/2)-Nc.ant)/nc
M.notas
```

```
## [1] 5.5
```

```
median(notas)
```

```
## [1] 5
```