Estadística descriptiva con datos cuantitativos

Ramon Ceballos

30/1/2021

1. Medidas de posición

Las **medidas de posición** estiman qué valores dividen las observaciones o población en unas determinadas proporciones.

Los valores que determinan estas posiciones son conocidos como los **cuantiles**. Dependiendo de la posición del cuantil suelen recibir un nombre genérico para cada posición.

Pensándolo de este modo, la mediana puede interpretarse como una medida de posición, debido a que divide la variable cuantitativa en dos mitades.

Dada una proporción $p \in (0, 1)$ (mayor o estricto que 0 y menor o estricto que 1), el **cuantil de orden p* de una variable cuantitativa, Q_p, es el valor más pequeño tal que su frecuencia relativa acumulada es mayor o igual a p. Por tanto, antes de determinal un determinado cuantil se deben de determinar las frecuencias relativas de dicha variable cuantitativa.

Dicho de otro modo, si tenemos un conjunto de observaciones $x_1, \ldots, x_n$ y los ordenamos de menor a mayor, entonces Q_p será el número más pequeño que deja a su izquierda (incluyéndose a sí mismo) como mínimo a la fracción p de los datos. Es decir, $p \cdot n$ datos quedarían a la izquierda de dicho elemento.

Así, ahora es más claro ver que la mediana vendría a ser $Q_{0.5}$, el cuantil de orden 0.5.

Ejemplo 1

Consideremos un experimento en el que lanzamos 50 veces un dado de rol de 4 caras y obtenemos los siguientes resultados.

```
#definimos una semilla fija
set.seed(260798)

#determinamos la variable dado con sus 50 observaciones
dado = sample(1:4, 50, replace = TRUE)

#cerramos la semilla abierta
set.seed(NULL)

#nº de observaciones
length(dado)
```

```
## [1] 50
```

```r
#Los ordenamos de menor a mayor
dado = sort(dado)
dado
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 4 4
## [39] 4 4 4 4 4 4 4 4 4 4 4 4
```

Una vez obtenida la variable dado ordenada de menor a mayor en cuanto a sus observaciones, definimos un data frame que recoja las diversas tablas de frecuencias (absoluta, relativa, absoluta acumulada y relativa acumulada) para esta variable cuantitativa.

```r
df.dado = data.frame(Puntuacion = 1:4,
                     Fr.abs = as.vector(table(dado)),
                     Fr.rel = as.vector(round(prop.table(table(dado)),2)),
                     Fr.acu = as.vector(cumsum(table(dado))),
                     Fr.racu = as.vector(round(cumsum(prop.table(table(dado))),2)))
df.dado
```

```
##   Puntuacion Fr.abs Fr.rel Fr.acu Fr.racu
## 1          1     16   0.32     16    0.32
## 2          2     15   0.30     31    0.62
## 3          3      5   0.10     36    0.72
## 4          4     14   0.28     50    1.00
```

Si nos piden el cuantil $Q_{0.3}$, sabemos que este es el primer elemento de la lista cuya frecuencia relativa acumulada es mayor o igual a 0.3. Si observamos la tabla de frecuencias anterior (df.dado), este cuantil $Q_{0.3}$ se corresponde con la puntuación 1.

También podríamos hallarlo de otro modo: fijándonos en la lista ordenada de puntuaciones, el cuantil $Q_{0.3}$ sería el primer elemento de dicha lista tal que fuera mayor o igual que, como mínimo, el 30% de los datos. Si calculamos el 30% de 50, obtenemos que es 15. Esto lo que nos dice es que el cuantil que buscamos es el número que se encuentrae en la quinceava posición de la lista ordenada.

```r
#cuantil 30%
dado[15]
```

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

1.1 Cuantiles

Algunos cuantiles tienen nombre propio:

- Los **cuartiles** son los cuantiles $Q_{0.25}, Q_{0.5}$ y $Q_{0.75}$. Respectivamente, son llamados primer, segundo y tercer cuartil. El primer cuartil, $Q_{0.25}$, será el menor valor que es mayor o igual a una cuarta parte de las observaciones y $Q_{0.75}$, el menor valor que es mayor o igual a tres cuartas partes de los datos observados.
- El cuantil $Q_{0.5}$ es la **mediana**.
- Los **deciles** son los cuantiles Q_p con p un múltiplo de 0.1 ($Q_{0.1}, Q_{0.2}, Q_{0.3}, \ldots, Q_{0.9}$).
- Los **percentiles** son son los cuantiles Q_p con p un múltiplo de $0.01 (Q_{0.01}, Q_{0.02}, \ldots, Q_{0.98}, Q_{0.99})$. Los más utilizados son $Q_{0.05}$ y $Q_{0.095}$.

La definición de cuantil anteriormente dada es orientativa. La realidad es que, exceptuando el caso de la *mediana*, no hay consenso sobre cómo deben calcularse los cuantiles. En verdad, existen diferentes métodos que pueden dar lugar a soluciones distintas.

Al fin y al cabo, nuestro objetivo no es el de encontrar el primer valor de una muestra cuya frecuencia relativa acumulada en la variable sea mayor o igual a p, sino estimar el valor de esta cantidad para el total de la población.

Para calcular los cuantiles de orden p de una variable cualitativa x con R, se utiliza la instrucción **`quantile(x,p)`**, la cual dispone de *9 métodos diferentes* que se especifican con el parámetro **`type`**. El valor por defecto es **`type = 7`** y no hace falta especificarlo, como veremos en el siguiente ejemplo. Para más información sobre todos los valores posibles de este parámetro, haced click en el enlace a Wikipedia.

Ejemplo 2

```
#Defino una semilla
set.seed(0)

#creo la variable dado2
dados2 = sample(1:6,15, replace = TRUE)
dados2
```

```
##  [1] 6 1 4 1 2 5 3 6 2 3 3 1 5 5 2
```

```
#anulo la semilla
set.seed(NULL)

#Primer cuartil
quantile(dados2,0.25)
```

```
## 25%
##   2
```

```
#cuantil 0.8 de dados2
quantile(dados2,0.8)
```

```
## 80%
##   5
```