Definición de Cuantiles

Departamento de Posgrados - UDA

Introducción

Los cuantiles son métricas, que permiten saber la posición relativa de una observación en un conjunto de datos ordenado. Veamos un ejemplo cargando un conjunto de datos CSV.

```
data <- read.csv("cuantiles.csv")
head(data)
```

```
##   id        x
## 1  1 18.97416
## 2  2 18.29668
## 3  3 27.17307
## 4  4 14.78397
## 5  5 17.05436
## 6  6 15.71126
```

Este conjunto de datos posee $n = 24$ observaciones, las cuales pueden ser ordenadas con la función *order()*, a continuación extraemos el valor máximo y mínimo para calcular el rango.

```
n <- nrow(data)
data <- data[order(data$x),]
rango <- min(data) - max(data)
```

Medidas de posición

Imaginemos que al conjunto ordenado lo queremos partir en dos partes iguales, entonces determinaríamos qué valor usar para realizar la división de la manera más justa. La métrica que nos indica qué valor de x debemos usar para realizar la separación es la mediana que en R se obtiene con la función *median()*.

```
median(data$x)
```

```
## [1] 19.46458
```

De forma más general, los cuantiles son estos valores que nos permiten partir un conjunto de datos para formar intervalos conteniendo un porcentaje determinado de observaciones. En R los cuantiles se pueden encontrar con la función *quantile()*, brindándole el porcentaje (como un número entre 0 y 1) acumulado de observaciones. Por ejemplo para el 25% de observaciones usaríamos:

```
quantile(data$x, 0.25)
```

```
##      25%
## 17.36263
```

Esto quiere decir que alrededor del 25% de observaciones es menor a este valor y el 75% es mayor. Entonces otra forma de encontrar la mediana (sabiendo que acumula el 50% de observaciones) sería:

```
quantile(data$x, 0.5)
```

```
##      50%
## 19.46458
```

Cuartiles

Es conveniente dividir al conjunto de datos ordenado en n partes iguales. Si lo dividimos en 4 partes, los antes mencionados cuantiles que serían los umbrales entre cada grupo, toman el nombre de *cuartiles*. Entonces tenemos 3 cuartiles los cuales podrían encontrarse con la misma función, sabiendo que en cada grupo tenemos 25% de observaciones y que por lo tanto acumulan el 25%, 50% y 75% respectivamente.

```
quantile(data$x, c(0.25,0.5,0.75))
```

```
##      25%      50%      75%
## 17.36263 19.46458 22.89244
```

No existe uniformidad en la literatura acerca del cálculo de los mismos, puesto que muchas veces es necesario realizar interpolaciones para hallar los valores aproximados. R implementa 9 diferentes algortimos para encontrarlos los cuales se pueden especificar con el argumento *type*. Por ejemplo si se emplea el método clásico descrito que utiliza la siguiente fórmula para determinar la posición de un cuartil k:

$$\frac{k(n+1)}{4}$$

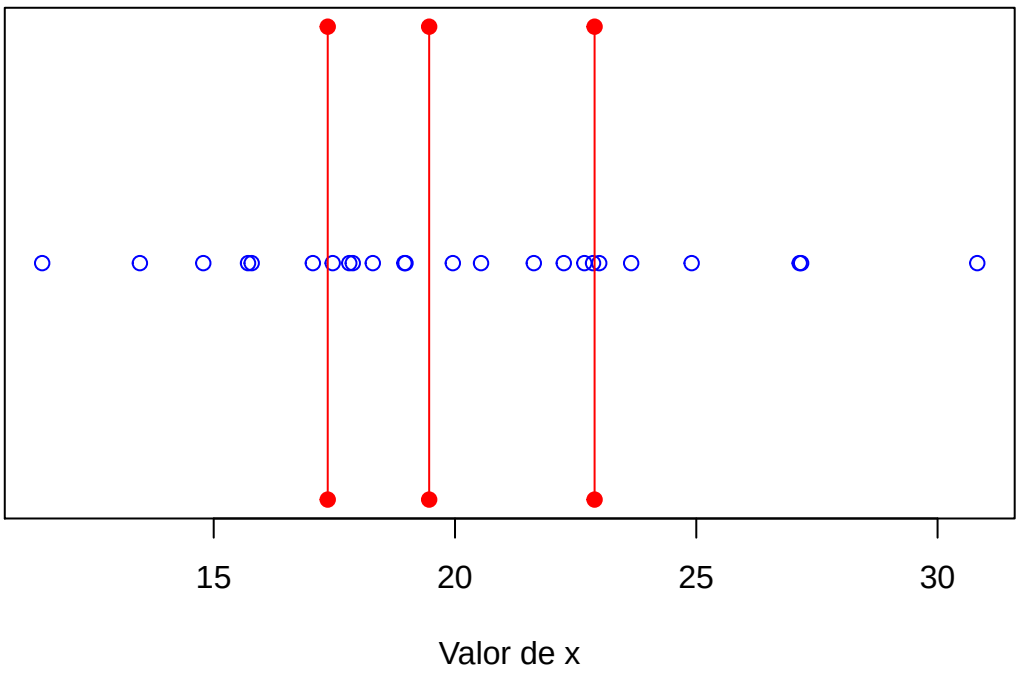
entonces se debe usar un *type=6* de la siguiente manera:

```
quantile(data$x, c(0.25,0.5,0.75),type = 6)
```

```
##      25%      50%      75%
## 17.15712 19.46458 22.95676
```

graficamente se pueden observar estas divisiones en una recta numérica:

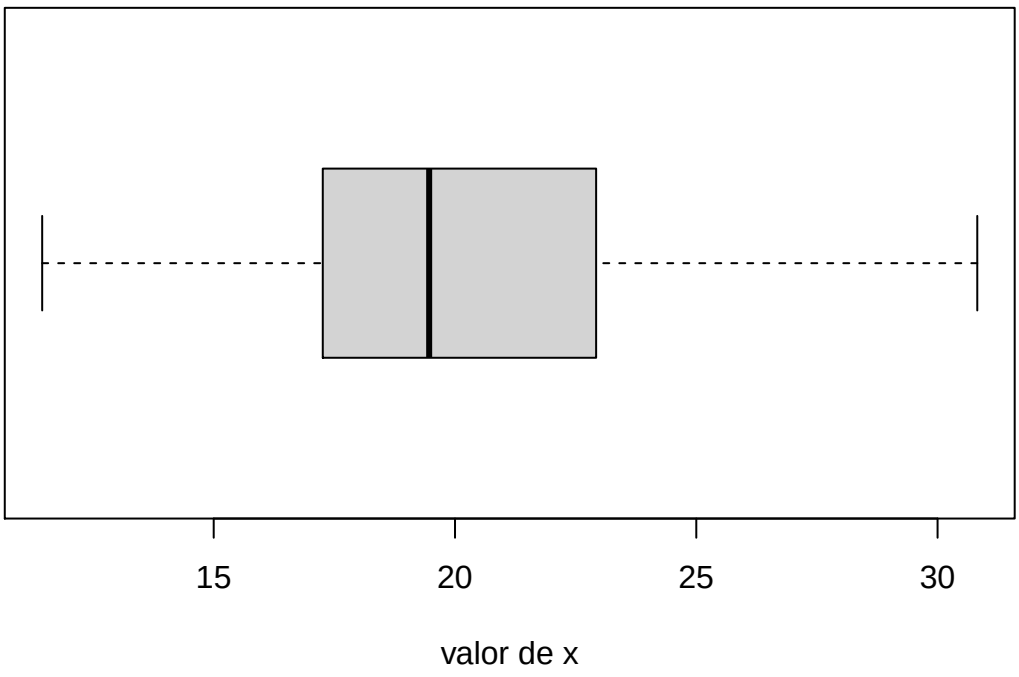
```
source("utils2.R")
plot_points(data$x, quartiles = T, xlab="Valor de x")
```




```
##      25%      50%      75%
## 17.36263 19.46458 22.89244
```

otra forma gráfica de apreciar los cuartiles es mediante un boxplot (diagrama de cajas), que en R se genera mediante la función *boxplot()*.

```
boxplot(data$x,horizontal = T, xlab="valor de x")
```



Percentiles

Si ahora lo dividimos en 100 partes, los cuantiles toman el nombre de *percentiles*. Entonces tenemos 99 percentiles los cuales podrían encontrarse con la misma función, especificando los porcentajes acumulados en el segundo argumento. Por ejemplo para encontrar los percentiles 80 y 90 usaríamos.

```
quantile(data$x, c(0.80, 0.90))
```

```
##      80%      90%
## 23.25368 26.47103
```

De igual manera, si se emplea el método clásico descrito que utiliza la siguiente fórmula para determinar la posición de un percentil i:

$$\frac{i(n+1)}{100}$$

entonces nuevamente se debe usar un *type=6*. Se de tomar en cuenta que el percentil 25 coincide con el cuartil 1, el percentil 50 con el cuartil 2, y el percentil 75 con el cuartil 3. Estas métricas se pueden consultar rápidamente mediante la función *summary()*.

```
summary(data$x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.45   17.36   19.46   20.18   22.89   30.82
```

Finalmente, el **rango intercuartil (RIC)** se define como la distancia entre el cuartil 1 y el cuartil 3, es decir:

```
q1 <- as.numeric(quantile(data$x, 0.25, type=6)) #cuartil 1
q3 <- as.numeric(quantile(data$x, 0.75, type=6)) #cuartil 3
ric <- q3 - q1
```

Esta métrica es útil para encontrar valores atípicos, es decir valores extremos aislados (de muy poca frecuencia). La regla de <Tukey's> define un intervalo fuera del cual se encuentra estos valores. El intervalo es:

$$(Q1 - 1.5RIC, Q3 + 1.5RIC)$$

en este ejemplo tendríamos:

```
c(q1 - 1.5*ric , q3 + 1.5*ric)
```

```
## [1]  8.45765 31.65622
```

debido a que los valores mínimo y máximo en el conjunto de datos: 11.4453349, 30.8230465 respectivamente están dentro de este intervalo, concluimos que no tenemos valores atípicos.

Universidad del Azuay (2021)

Departamento de Posgrados

imendoza@uazuay.edu.ec