Complemento unidad 2. Otras estadísticas descriptivas

Cuartiles

Definición

Se definen como cuartiles a las cantidades denotadas por Q_1, Q_2 y Q_3, que dividen al conjunto de datos en cuatro segmentos del mismo tamaño, cada uno con el 25 % de las observaciones. En particular

- Q_1 es el primer cuartil y cumple la siguiente propiedad: el 25 % de las observaciones son menores o iguales a Q_1.
- Q_2 es el segundo cuartil y cumple la propiedad: el 50 % de las observaciones son menores o iguales a Q_2. Este estadístico coincide con la mediana.
- Q_3 es el tercer cuartil y cumple la propiedad: el 75 % de las observaciones son menores o iguales a Q_3.

Definición

Se define el *rango intercuartil* como

$$IQR = Q_3 - Q_1$$

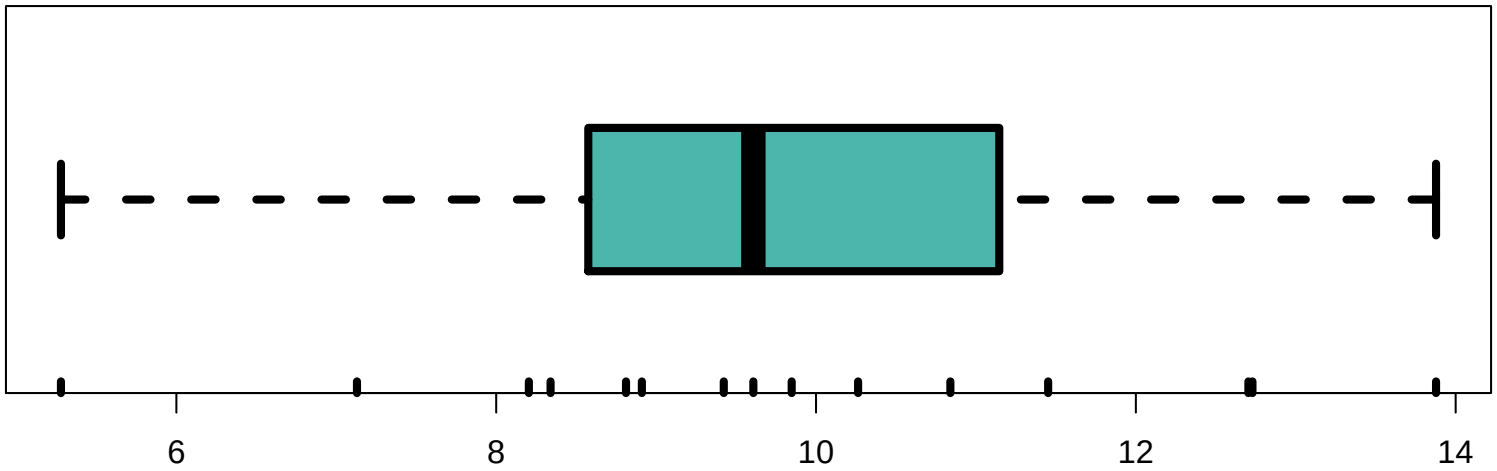
Boxplot o gráfico de caja (y bigotes)

Este gráfico permite explorar la distribución de los datos a partir de 4 estadísticos: Q_1, Q_2, Q_3 e IQR. La construcción de describe a continuación, nuestro interés es un *boxplot* horizontal.

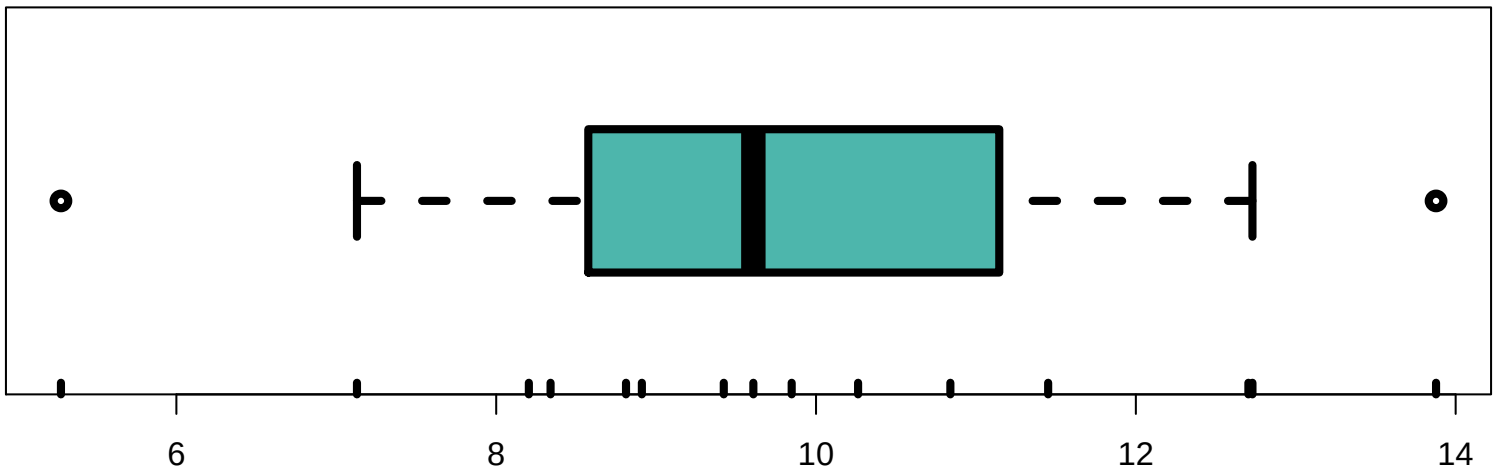
- Graficar un rectángulo delimitado por Q_1 y Q_3, de manera que la base del rectángulo es igual al IQR. La altura del rectángulo no es relevante.
- Al interior de la caja se grafica un segmento vertical ubicado en Q_2. Ahora la caja está dividida en dos partes, sus proporciones se utilizan para estudiar la simetría de la distribución de los datos.
- Se agregan dos *bigotes* horizontales a los costados de la caja. A *priori* las longitudes las rectas se fija como un múltiplo k del IQR, pero éstas se ajustan a las observaciones más extremas que distan menos de $kIQR$ de la caja.
- Finalmente se agregan como puntos para representar a las observaciones que están fuera del intervalo $(Q_1 - kIQR, Q_3 + kIQR)$, si las hay.

En `R` se utiliza el comando `boxplot` para hacer gráficos de caja y bigotes.

Ejemplo: boxplot



k = 1.5



k = 1

Cuantiles

Definición

Dado un conjunto de observaciones $x_1, \ldots, x_n$, se define el cuantil muestral α, con $\alpha \in (0, 1)$, denotado por q_α, como un número tal que el $100\alpha\,\%$ de las observaciones son menores o iguales a q_α.
Por ejemplo, $Q_1 = q_{0.25}$, la mediana es $q_{0.5}$ y $Q_3 = q_{0.75}$.

- Existen distintas formas de calcular los cuantiles muestrales, que se basan en distintos supuestos sobre la distribución subyacente que generó a las observaciones.
- Para calcularlos utilizaremos la función `quantile` de `R`. En el argumento `prob` se introduce el valor de α.
- Para tener una idea de la cantidad de *algoritmos* disponibles para calcular los cuantiles podemos introducir revisar la ayuda de la función `quantile` y revisar el apartado del argumento `type`.
- Algunos cuantiles tienen nombre (mediana, cuartiles, deciles, percentiles) según el número de partes iguales en que dividen a la población.

Distribución empírica

Definición

Dado un conjunto de observaciones $x_1, \ldots, x_n$, se define la función de distribución empírica, denotada por F_n, a la función $F_x : \mathbb{R} \to [0, 1]$ dada por

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, x)}(x_i), \qquad \forall x \in \mathbb{R}.$$

- La función de distribución empírica de n observaciones es una función escalonada, con saltos en $x_1, \ldots, x_n$ de tamaño $1/n$.
- Si se asume que los datos constituyen una muestra *aleatoria* de una distribución **teórica** F, entonces F_n *se parace a* F, y la similitud crece conforme $n \to \infty$.
- F_n se puede utilizar para estudiar la forma de la distribución, incluso es posible compararla con alguna distribución teórica y decidir si los datos provienen o no de tal distribución.
- En `R`, se utiliza el comando `ecdf` para calcular la fde de un conjunto de datos.

Ejemplo: distribución empírica