(1) Observatorio Astronómico de Córdoba, Argentina

1 Diagrama de caja (boxplot)

En esta sección detallaremos los fundamentos teóricos detrás de la principal herramienta estadística utilizada durante este trabajo: el diagrama de caja o boxplot.
En la estadística descriptiva los diagramas de caja o *boxplots* son presentaciones visuales que describe varias características importantes, al mismo tiempo, tales como la dispersión y la simetría.

1.1 Componentes de un diagrama de caja

- **Cuartil 1 - Q1**: Es el valor que contiene el 25% de los datos ordenados, es decir, es la mediana de la mitad menor de los datos.
- **Cuartil 2 - Q2**: Este valor indica el 50% de los datos ordenados y coincide con la mediana de todos los datos.
- **Caurtil 3 - Q3**: Representa el 75% de los datos ordenados y es la mediana de la mitad mayor de los datos.
- **Rango intercuartílico - IQR**: Es la diferencia entre el tercer y el primer cuartil de una distribución. Es una medida de la dispersión estadística.
- **Bigotes**: Son las líneas que se extienden desde el primer y el tercer cuartil, se extienden hasta los valores máximo y mínimo de los datos o hasta 1,5 veces el IQR.
- **Valores atípicos**: Cuando los datos se extienden más allá de los límites de los bigotes, significa que hay valores atípicos en la serie de datos. Por lo tanto, se consideran atípicos los valores inferiores a $Q1 - 1.5IQR$ o superiores a $Q3 + 1.5IQR$.
- **Muescas**: Muestran el intervalo de confianza alrededor de la mediana. Según *Graphical Methods for Data Analysis* (Chambers, 1983), aunque no es una prueba formal, si las muescas de dos boxplot no se superponen, existe una "evidencia sólida" (95% de confianza) de que sus medianas difieren.

La figura 1 es un diagrama que muestra los tres cuartiles, el bigote inferior y superior, el rango intercuartílico, el intervalo de confianza de la medina y los posibles datos atípicos o *outlier*.

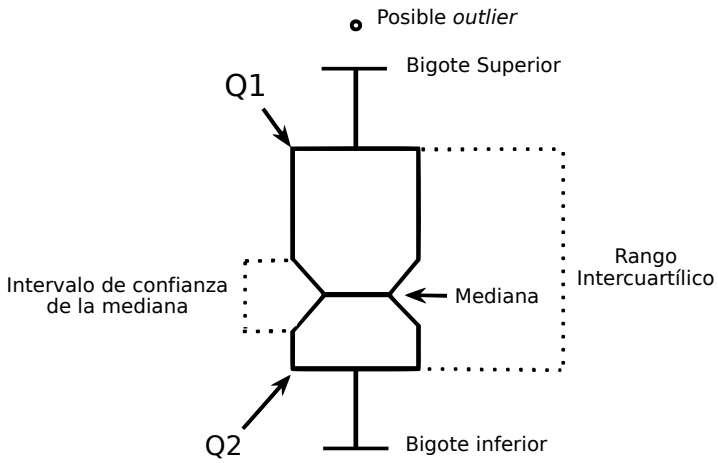

Figure 1: Diagrama de caja con bigotes e intervalos de confianza.

1.2 Construcción del diagrama de caja

La construcción del diagrama de caja puede ser de forma manual o automática a través de lenguajes como R y Python. Usualmente los pasos a seguir para la construcción manual son:

1. Ordenar los datos de menor a mayor;
2. Calcular los tres cuartiles (Q1, Q2 y Q3). Estos tres valores dividen el conjunto de datos en cuatro partes iguales y definen el tamaño de la caja a a partir del IQR con una linea en su interior que indica la mediana (Q2);
3. Calcular los límites inferior y superior de los bigotes para indicar los valores extremos (mínimo y máximo) apartir de líneas que se extienden desde el borde de la caja;
4. Una vez definidos los límites de los bigotes, éstos marcan la existencia de valores atípicos que serán los valores fuera del intervalo del límite

inferior y límite superior de los bigotes, usualmente se representan por puntos o círculos pequeños

En este trabajo se genero el diagrama de caja de forma automática. Usando el paquete *seaborn* de python. En particular, el calculo de la ubicación de las muescas es basado en una aproximación asintótica de Gauss (ver **?**).
Los límites del intervalo de confianza al rededor de la mediana M para un conjunto de n datos está dado por:

$$M \pm 1.58 \cdot IQR/\sqrt{n} \tag{1}$$

Esta definición conlleva a un 95% de confianza en la diferencia de las medianas cuando las muescas no se superponen.