Aula360
Aprendizaje para el siglo XXI

Apuntes Unidad 1

Cuartiles y diagramas de cajón

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

CMM Centro de Modelamiento Matemático

Siempre que se trabaja con datos cuyos valores se pueden ordenar, resulta interesante analizar su distribución. Una medida importante es la mediana, que divide a la muestra en dos partes que contienen al menos la mitad de los datos ¿Existirán más formas de dividir los cuartiles? La respuesta es sí.

CUARTILES

Los cuartiles son tres valores que dividen una muestra en, aproximadamente, 4 partes iguales cuando los datos están ordenados de menor a mayor. Estos tres valores se denotan por Q_1, Q_2 y Q_3, que representan al primer, segundo y tercer cuartil respectivamente.

El primer cuartil es un valor que nos dice que aproximadamente 25% es menor o igual al valor del cuartil. Por ejemplo, si tenemos que el primer cuartil de una muestra es 10, quiere decir que aproximadamente 25% de los valores de la muestra son menores o iguales a 10. De manera análoga podemos decir que el segundo cuartil es un valor que nos dice que aproximadamente un 50% de la muestra es menor o igual al valor del cuartil y que aproximadamente 75% de la muestra es menor o igual al valor del tercer cuartil.

Para calcular los cuartiles se deben seguir los siguientes pasos:

Paso 1: Ordenar los *n* datos en forma creciente

Paso 2: Calcular $\frac{n \cdot k}{4}$, con $k = \{1, 2, 3\}$.

Paso 3: Determinar que el *k* cuartil de la siguiente forma:

- Si ese valor resulta un número entero, Q_k es igual al promedio entre el dato que se ubica en esa posición y el dato siguiente.
- Si ese valor resulta un número no entero, Q_k es igual al dato que ocupa la posición $[\frac{n \cdot k}{4}] + 1$.

Observación: Se denomina parte entera de un número positivo a su truncamiento al entero y se denota con corchetes. Por ejemplo, [6,1] = 6.

Los cuartiles nos ayudan enormemente a entender la distribución de la muestra, incluso sin tener un gráfico que nos represente la forma de esta. A continuación encontramos una tabla que nos ayudará a profundizar en la interpretación de los cuartiles.

Primer cuartil Q_1	Al menos el 25 % de los datos son menores o iguales que Q_1.	Al menos el 75 % de los datos son mayores o iguales que Q_1.
Segundo cuartil Q_2 (mediana)	Al menos el 50 % de los datos son menores o iguales que Q_2.	Al menos el 50 % de los datos son mayores o iguales que Q_2.
Tercer cuartil Q_3	Al menos el 75 % de los datos son menores o iguales que Q_3.	Al menos el 25 % de los datos son mayores o iguales que Q_3.

DIAGRAMA DE CAJÓN

Un diagrama de caja, diagrama de cajón o diagrama de cajón con bigotes es una representación gráfica de la distribución de los datos a través de sus cuartiles, como se muestra a continuación:

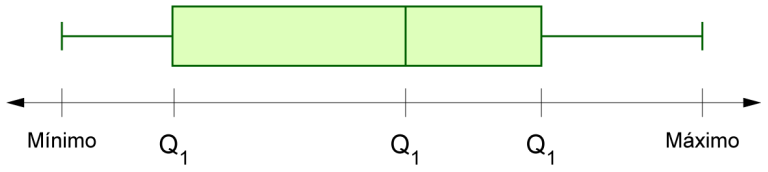

Figura 1: Diagrama de cajón

El cajón queda definido por los tres cuartiles de la distribución; y los bigotes, por los datos mínimo y máximo.

Es natural preguntarse, ¿Qué significado o importancia tiene la longitud del cajón? ¿Qué interpretación tiene la longitud de los bigotes?

El hecho de que una parte del cajón sea más ancha o delgada depende de la variabilidad de los datos. Mientras más delgada sea una parte del cajón, quiere decir que el 25% de los datos asociados a esa parte presentarán menos variabilidad. Por el contrario, si una parte del cajón es más ancha los datos asociados a esa parte presentarán una mayor variabilidad.

Esto mismo se aplica a la longitud de los bigotes, mientras más cortos sean se interpreta que la variabilidad de los datos de esos intervalos es menor y viceversa. Vale decir que hablar de variabilidad de datos está relacionado a hablar de dispersión de los datos.

Debido a lo anterior, los diagramas de cajón se consideran una representación simple y adecuada para visualizar la dispersión de los datos dentro de una distribución.

¿CÓMO ELABORAR UN DIAGRAMA DE CAJÓN?

Paso 1: Para elaborar un diagrama de cajón, primero debemos determinar el valor mínimo, máximo y los cuartiles a partir de los datos que tengamos.

Paso 2: Luego, debemos ubicar esos 5 valores en una recta numérica.

Paso 3: Usando los tres cuartiles construimos el cajón, tal como se muestra en la figura 1.

Paso 4: Por último, usando el mínimo y máximo, construimos los bigotes.

RELACIÓN ENTRE EL DIAGRAMA DE CAJÓN Y EL HISTOGRAMA

Los diagramas de cajón con bigotes, al igual que los histogramas, representan cómo se distribuyen los datos en una muestra o población, de manera que es natural que se puedan relacionar ambas representaciones.

A diferencia de los histogramas, los diagramas de cajón usan una sola dimensión o eje, lo que permite comparar las distribuciones de varias muestras en un solo gráfico.

Por otro lado, los diagramas de cajón permiten visualizar rápidamente cuán simétrica es la distribución de los datos, y entregan información inmediata sobre las medidas de posición más usadas: la mediana, los cuartiles, y los valores máximos y mínimos.

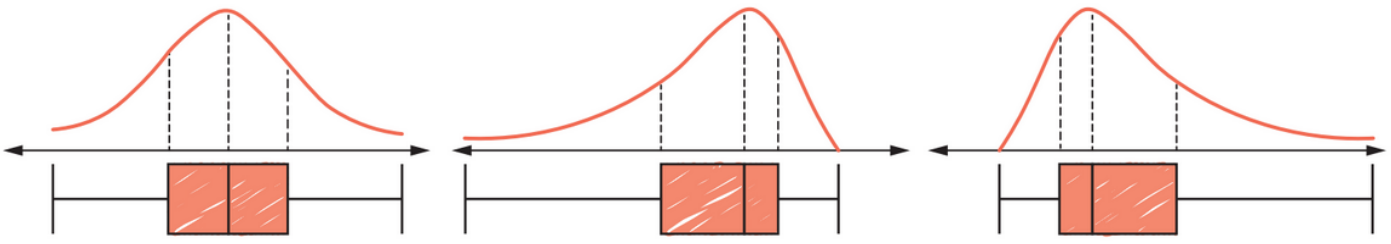

Figura 2: Relación entre diagrama de cajón y distribución de los datos.

El que una distribución esté cargada hacia uno de los dos lados, por ejemplo cargada a la izquierda, se interpreta diciendo que los datos de menor valor presentan una mayor

frecuencia que los datos de mayor valor. Por el contrario, si la distribución está cargada a la derecha se dice que los datos de mayor valor tienen frecuencias más altas.

DATOS ATÍPICOS

Los datos atípicos son observaciones cuyos valores son muy diferentes a otras observaciones del mismo grupo de datos, es decir, son datos extremadamente grandes o extremadamente pequeños en relación con el resto. Estos valores pueden distorsionar los resultados de los análisis, por lo que se deben identificar y tratar de manera adecuada.

Hay varias razones que podrían explicar la existencia de valores atípicos:

- Errores en la toma o registro de los datos, por ejemplo, al escribir los datos en una tabla.
- Existencia de datos correctos que, por alguna razón, tienen un valor muy alejado del resto.

Los datos atípicos, al ser valores alejados del centro de la distribución, pueden afectar la media, por lo que deben tenerse en cuenta para evitar malas interpretaciones. Sin embargo, debido a que hay varias causas que pueden explicar la presencia de datos atípicos, no es recomendable excluirlos del análisis de inmediato.

Una ventaja de los diagramas de cajón con respecto a otros tipos de representaciones de una distribución es que los datos atípicos se pueden visualizar directamente en el gráfico. Esto se hace dibujando el diagrama de cajón excluyendo los datos atípicos, y luego representándolos mediante puntos fuera de los bigotes del diagrama tal como se muestra a continuación.

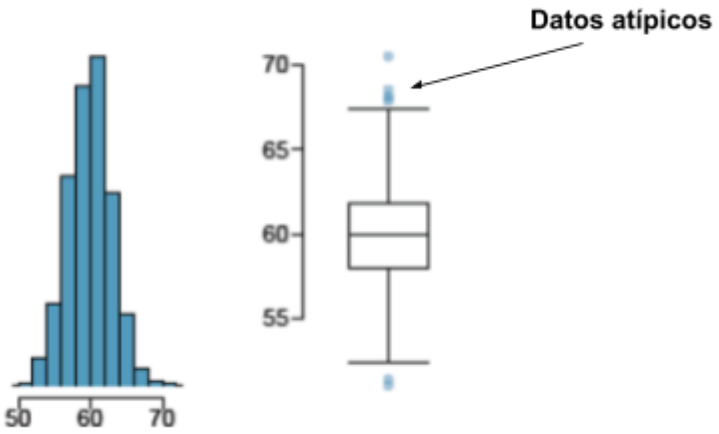

Figura 1: Datos atípicos y su representación en el diagrama de cajón

Para determinar si un valor es atípico o no, se usa un algoritmo que viene incluido en la mayoría de los programas que generan diagramas de cajón, de manera que su visualización es automática.

Al seguir analizando los diagramas de cajón y los datos que contempla, resulta interesante estudiar sobre la existencia de algún tipo de medida del ancho o largo del cajón. Para cuantificar esto, se utiliza el Rango Intercuartil (RIC), sobre el cual profundizaremos a continuación.

RANGO INTERCUARTIL

Por último, los diagramas de cajón permiten visualizar en qué rango se ubica el 50% de los datos centrales de la distribución. Este último intervalo corresponde a la diferencia entre Q_3 y Q_1, y se denomina **rango intercuartil (RIC).** El rango intercuartil (RIC) es una medida de la variabilidad o dispersión de un conjunto de datos. Corresponde al largo de la caja, es decir, a la distancia entre el primer y el tercer cuartil.

SÍNTESIS

- Los cuartiles Q_1, Q_2 y Q_3 son tres valores que dividen un conjunto de datos ordenados de menor a mayor en **cuatro** grupos de aproximadamente el mismo tamaño.

- Para determinar el cuartil Q_k , para k en {1, 2, 3}, se debe calcular $\frac{n \cdot k}{4}$ y luego considerar dos casos:
    - Si $\frac{n \cdot k}{4}$ es un número entero p, Q_k es igual al promedio entre los datos que se ubica en las posiciones p y $p + 1$.
    - Si $\frac{n \cdot k}{4}$ es un número no entero, Q_k es igual al dato que ocupa la posición $[\frac{n \cdot k}{4}] + 1$.

- El **diagrama de cajón** representa gráficamente los cuartiles, y el mínimo y máximo de una distribución. Además, esta representación simple permite visualizar la dispersión en una distribución de datos.
- Los diagramas de cajón permiten inferir la forma que tendrían los datos si se representarán con un histograma.
- Los diagramas de cajón permiten visualizar la simetría en la distribución de una variable.

- Una ventaja importante de los diagramas de cajón es que permiten visualizar dos o más distribuciones en un mismo gráfico, por lo que facilitan la comparación de muestras o poblaciones.
- Los valores atípicos son datos extremadamente pequeños o grandes en relación con el resto de los datos observados.

- La diferencia entre el tercer y primer cuartil se denomina rango **intercuartil** y es un indicador de cuán dispersos están los datos en torno a la mediana.

- El rango intercuartil es una medida más robusta que el rango, es decir, es menos susceptible a los valores atípicos o alejados del centro de la distribución.