PREWORK
SESIÓN 3

Introducción

En este prework, estudiarás las medidas de tendencia central más conocidas, algunas medidas de posición muy útiles como los cuartiles y medidas de dispersión. Es muy importante que conozcas estos conceptos básicos cuando estudias estadística, posteriormente en el work, trabajarás con datos, y los cálculos serán muy rápidos y fáciles de hacer con algunas funciones de R, pero primero necesitas entender el significado de estas medidas, para que, en el work, te puedas enfocar únicamente en cálculos fáciles e interpretaciones.

Objetivos

- Conocer las principales medidas de tendencia central, sus interpretaciones, así como ventajas y desventajas
- Conocer la utilidad de los cuantiles y algunas medidas de posición útiles y comunes en la práctica
- Conocer algunas medidas de dispersión y saber interpretarlas

Temas

- Medidas de tendencia central
- Medidas de posición
- Medidas de dispersión

Medidas de tendencia central

Dado un conjunto de datos, las llamadas medidas de tendencia central son números alrededor de los cuales se concentran los datos. La **media** o promedio es quizás la medida de tendencia central más conocida. Dado un conjunto de datos $\{x_1, \cdots, x_n\}$ que representa una muestra de alguna población, la media del conjunto se define como

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

es decir, el promedio de los datos.

Ejemplo. Un estudiante cursó seis materias en el semestre, obteniendo calificaciones de 8, 7, 10, 8, 9 y 7. Su calificación media o promedio semestral es de

$$\frac{8+7+10+8+9+7}{6} = 8.17$$

Ejemplo. Una pequeña compañía consultora tiene una secretaria, un empleado de limpieza, un mensajero y un economista. Sus salarios mensuales respectivos son de $5000, $4000, $3500 y $50000. El salario promedio de la compañía es de

$$\frac{5000+4000+3500+50000}{4} = 15625.$$

El lector podría pensar que este salario no es una medida *representativa* de los salarios del personal. ¿Qué sucede? La razón es que existe un dato, el salario del economista, que está totalmente fuera del rango de los demás salarios. Así, al realizar el promedio, este dato *jala* a los demás.

Para evitar problemas con datos alejados de los demás, como en el ejemplo anterior, se utiliza otra medida de tendencia central llamada mediana. La **mediana** es el valor que parte al conjunto de datos ordenados en dos. Para encontrar la mediana, los datos se ordenan de menor a mayor y,

- si el conjunto de datos es impar, la mediana es el valor que se encuentra a la mitad del conjunto,

- Si el conjunto de datos es par, la mediana es el promedio de los dos datos intermedios.

Ejemplo. Encontrar la mediana del conjunto de salarios {$5000, $4000, $3500 y $50000}. Los ordenamos de menor a mayor obteniendo,

$$\{3500,\ 4000,\ 5000,\ 50000\}.$$

Este es un conjunto par de datos de manera que la mediana es el promedio de los dos datos intermedios, es decir la mediana es

$$\frac{4000+5000}{2} = 4500.$$

En ocasiones, este dato puede ser una mejor representación de los salarios de la empresa que el promedio obtenido anteriormente en el ejemplo.

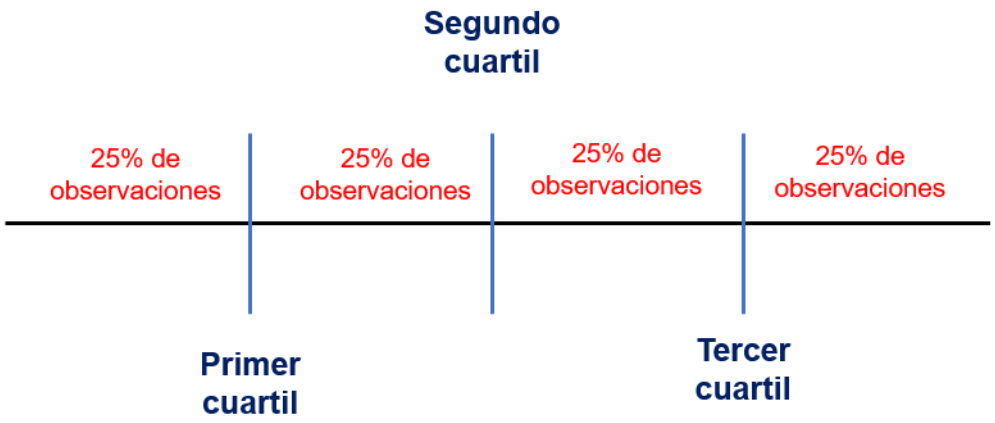
La media y la mediana no toman en cuenta la repetición de los datos, se define para este efecto, la moda. Ésta es simplemente el valor o categoría que ocurre con mayor frecuencia en un conjunto de datos. Es claro que puede haber más de una moda ya que puede haber más de un dato que se repita con la misma frecuencia. Para el caso de dos modas, decimos que la distribución de los datos

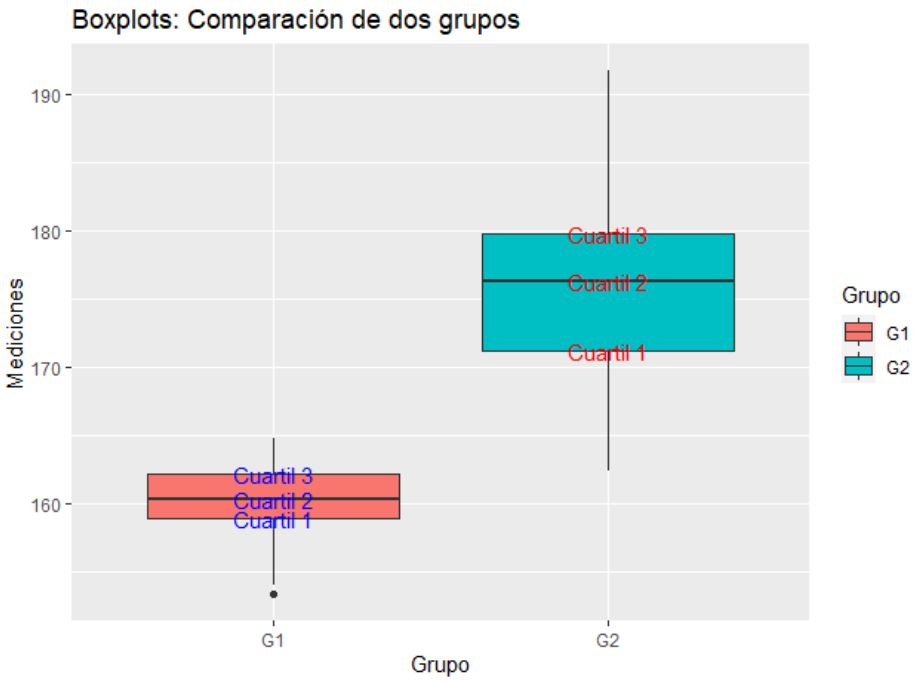
es bimodal. Cuando existen más de dos modas, la distribución se denomina multimodal.

Medidas de posición

De manera un poco informal pero útil para la práctica, podemos decir que el cuantil de orden p $(0 < p < 1)$ de un conjunto de mediciones, es un número que deja una proporción p de valores del conjunto por debajo de él. Por ejemplo, el cuantil de orden 0.43 dejaría un 43% de las observaciones por debajo de él.

Cuartiles. Los cuartiles son 3 números que dividen al conjunto de datos en cuatro partes iguales, es decir, debajo del primer cuartil se encuentra el 25% de las observaciones, el segundo cuartil es la mediana y el tercer cuartil es un número que tiene el 75% de las observaciones por debajo de él.





Deciles. Los deciles son 9 valores que dividen el conjunto de datos en 10 partes iguales.

Medidas de dispersión

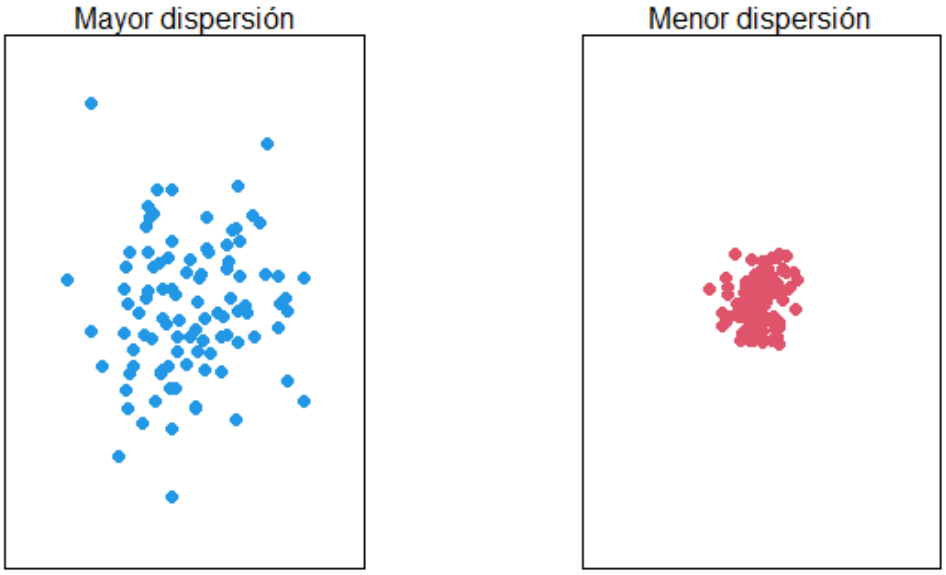
Consideremos a los conjuntos {5, 6, 7, 8, 9} y {1, 2, 7, 12, 13}. Ambos tienen media y mediana iguales a 7 y, sin embargo, nuestra intuición nos dice que los datos del segundo conjunto están más dispersos. ¿Cómo formalizar este concepto de dispersión? Una forma de hacerlo es considerando el **rango** o extensión de los datos que se define como la diferencia entre el dato más grande y el más pequeño.

Ejemplo. El rango del conjunto de datos {5, 6, 7, 8, 9} es de 9 - 5 = 4 y el rango del conjunto {1, 2, 7, 12, 13} es de 13 – 1 = 12. Observemos que el conjunto con el rango más grande es más disperso.

Una forma muy utilizada para medir la dispersión de un conjunto de datos es la llamada **desviación estándar**. Ésta mide qué tanto los datos se desvían de la media y se denota comúnmente por s_n. Esta desviación se construye a partir de su cuadrado conocido como **varianza**, como sigue: dado un conjunto de datos $\{x_1, \cdots x_n\}$ con media $\bar{x}$, la varianza se define como

$$s_n^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}.$$

Intuitivamente, $x_i - \bar{x}$ es la distancia a la media del dato x_i, ésta se eleva al cuadrado para tener siempre un valor positivo y se divide entre *n* para obtener el promedio de estas desviaciones.



Por definición, la varianza da un valor numérico para el promedio de los cuadrados de las distancias. Para que el número conserve las unidades originales de la variable, se toma la raíz cuadrada y se tiene así la desviación estándar

$$s_n = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}.$$

Esta desviación nos da una idea de que tan alejados están los datos de la media.

A la varianza de una población se le denota por σ^2 y a su desviación estándar por σ. Al igual que en el caso de la media, la varianza de una población puede inferirse a partir de las varianzas de las muestras. Desgraciadamente, en este caso se tiene una complicación: la varianza de la población no se aproxima bien por el valor esperado de las varianzas de las muestras y tiende a subestimarse. Decimos que s_n^2 es un **estimador con sesgo** de σ^2.

El problema de estimación de la varianza poblacional puede solucionarse utilizando, en lugar de las varianzas muestrales s_n^2, la siguiente expresión, en donde se divide entre *n – 1*, en lugar de *n:*

$$s_{n-1}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}.$$

En este caso, el valor esperado de los números s_{n-1}^2 para las diferentes muestras si es un buen estimador -sin sesgo- para la varianza poblacional.

Rango intercuartílico. Otra medida de dispersión muy útil es el **rango intercuartílico**, consulta en diversas fuentes acerca del rango intercuartílico y cuáles son sus ventajas como medida de dispersión.

Quiz

1. ¿Cuál es la media del siguiente conjunto de datos, redondeo a 2 dígitos?

195, 198, 199, 205, 212, 204, 202

a) 202.1429
b) 202.12
c) 202
d) 202.14

2. ¿Cuál es la mediana del conjunto de datos anterior?
e) 202.1429
f) 202.12
g) 202
h) 202.14

3. ¿Cuál es el mínimo valor que puede tomar la varianza?
a) 1
b) 31.1
c) 5.58
d) 0

4. Encuentre el rango intercuartílico para el conjunto de datos cuyos cuartiles son los siguientes:

Primer cuartil: 198.5, Segundo Cuartil: 202, Tercer cuartil: 204.5
a) 2.5
b) 6
c) 3.5
d) 2

BIBLIOGRAFÍA UTILIZADA

- Rumbos, B., (2009). Pensando antes de actuar: matemáticas para decidir. Instituto Tecnológico Autónomo de México.