Introducción a datos abiertos

Clase 05

Tabla de contenidos

Introducción

¿Por qué necesitamos calcular estadísticas?

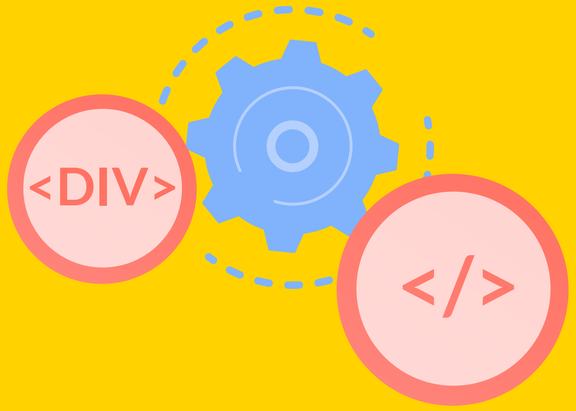
Estadística descriptiva: describir, sumarizar y comprender los datos.

Medidas de Tendencia Central

Distribución de los datos y asimetría

Medidas de Variabilidad

Medidas de tendencia central: Media



Dados los n números {x1,x2,...,xN} la media aritmética se define como:

$$\bar{x} = \frac{8 + 5 + (-1)}{3} = 4$$

Por ejemplo, para la muestra 8, 5 y -1, su media es:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Medidas de Tendencia Central: Mediana

La mediana es el valor del "medio" de una lista ordenada de datos (o el valor que separa la primera mitad y la segunda mitad de una distribución). Para una lista ordenada la mediana es calculada de diferente manera dependiendo de la cantidad de elementos de la misma:

Impar

[1, 2, 3, 5, 7, 8, 9, 10, 15]
9 elementos
La mediana es el valor de la posición 5 (la posición del “medio”)
Mediana = 7

Par

[-5, -1, 0, 1, 2, 3, 8, 20]
8 elementos
La mediana es la media de los valores en las dos posiciones centrales
Mediana = (1+2)/2 = 1.5

Medidas de Tendencia Central

Promedio con valores extremos
[1, 2, 3, 5, 7, 8, 9, 10, 150]

Promedio= 21.66

Mediana con valores extremos
[1, 2, 3, 5, 7, 8, 9, 10, 150]

Mediana= 7

Medidas de tendencia central:
Moda

La moda es el valor que aparece con mayor frecuencia o más veces en la distribución.

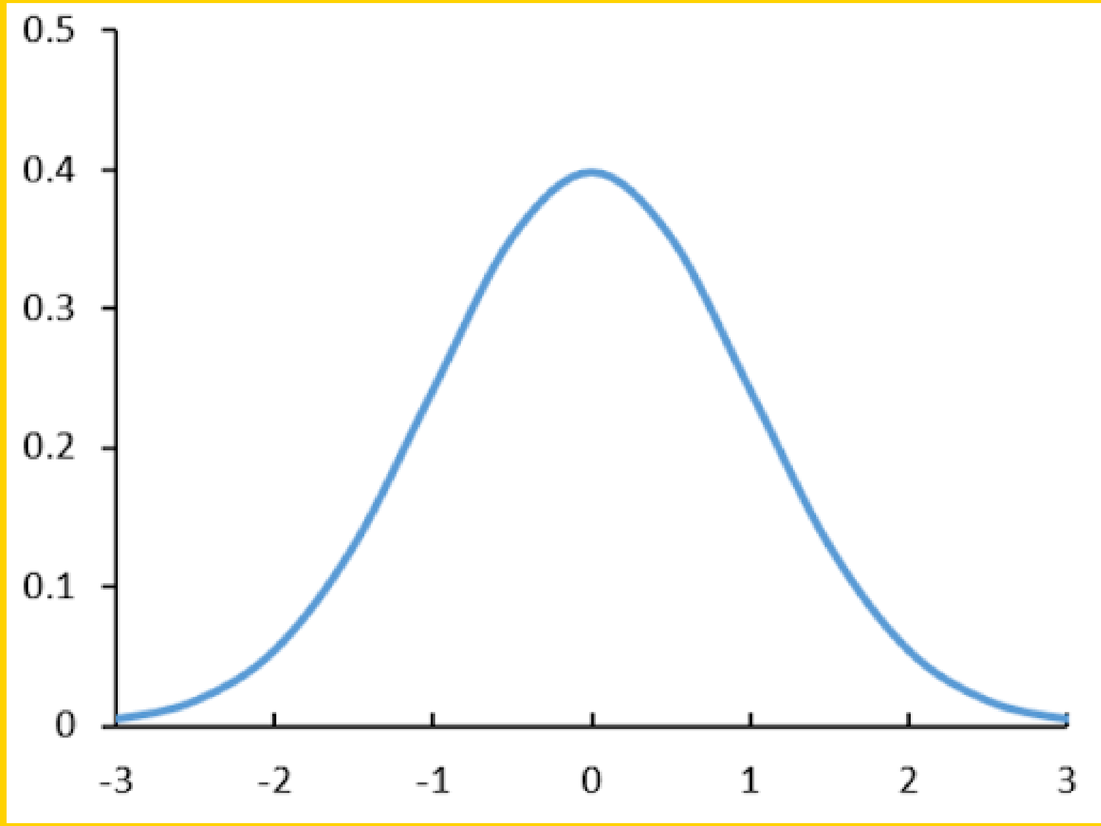
Por ejemplo

La moda de [0,1,1,2,2,2,2,3,3,4,4,4,5] es 2.

La moda no es necesariamente única. Puede ocurrir que haya dos valores diferentes que sean los más frecuentes.
Por ejemplo, para [10, 13, 13, 20, 20], tanto 13 como 20 son la moda.

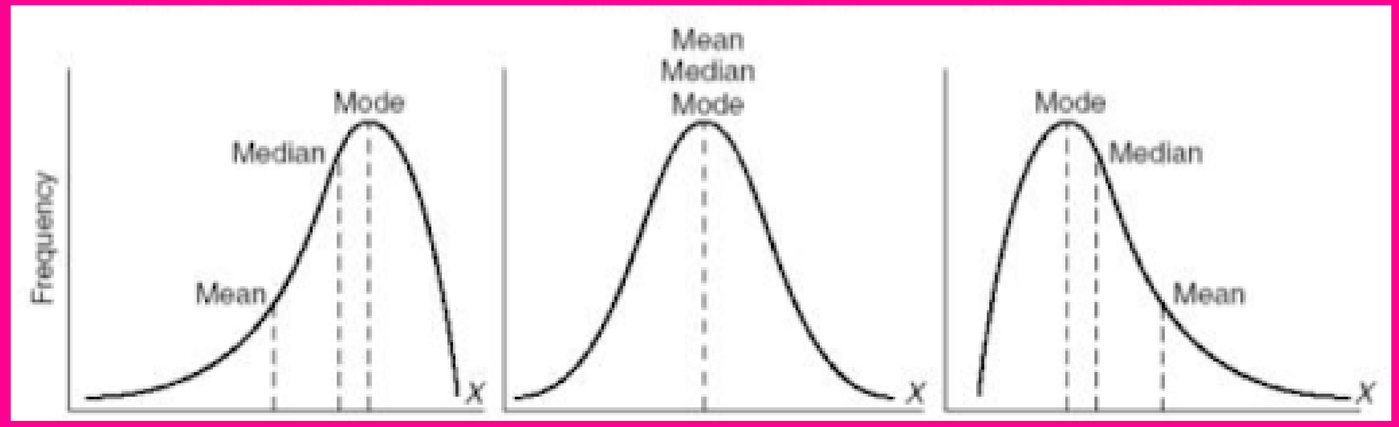
Distribuciones

Mediante las distribuciones de valores, mostramos cómo se acumulan las observaciones a lo largo de todo el rango de valores de la variable que estamos estudiando.



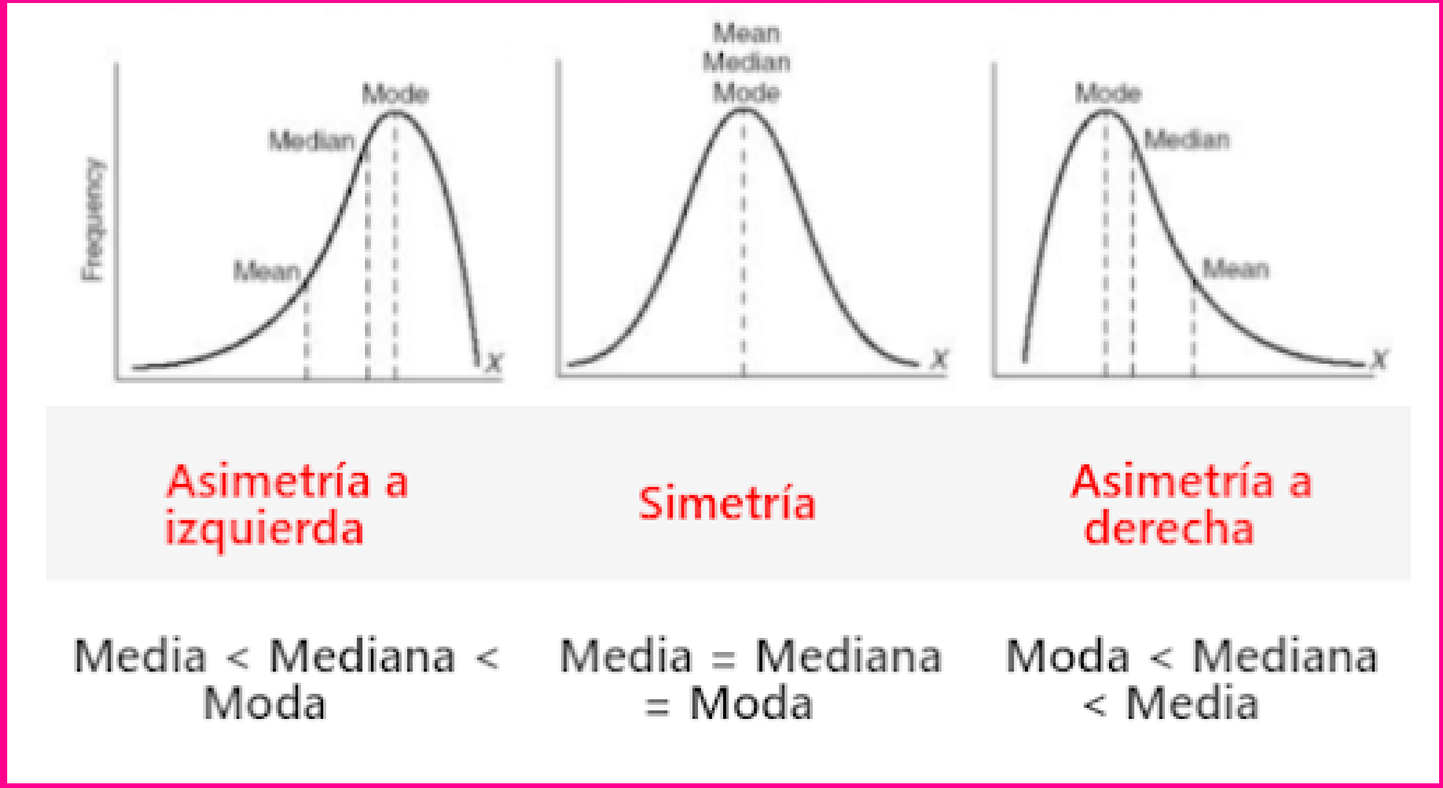
Distribución: Asimetría.

- Una distribución con asimetría a izquierda, significa que la cola de la izquierda es más larga que la de la derecha (gráfico de la izquierda).
- Por el contrario, una distribución con asimetría a derecha significa que la cola del lado derecho es más larga que la del lado izquierdo (gráfico de la derecha)
- A diferencia de éstas, una distribución simétrica no presenta este fenómeno dado que sus colas son de igual longitud (gráfico del centro).



Asimetría y Medidas de Tendencia Central

La media, mediana y moda son afectadas por la asimetría



Medidas de Dispersión o Variabilidad

Las medidas de variabilidad indican cómo están esparcidos los datos. Los mismos pueden categorizarse de acuerdo a:

-Rango
-Varianza
-Desvío estándar

Estas medidas proveen información complementaria a las medidas de tendencia central (media, mediana y moda).

Medidas de Dispersión o Variabilidad:

Rango

Sean $x1, x2, x3, \ldots$, los datos de una muestra ordenada en orden creciente, el rango es $xn - x1$

Rango intercuartil

Es la diferencia entre el primer y el tercer cuartil de la distribución. Acumula el 50% de la distribución y, a diferencia del rango, es un estadístico robusto, es decir que se ve poco afectado por valores extremos.

Medidas de Dispersión o Variabilidad: Varianza

La varianza es un valor numérico utilizado para describir cuánto varían los números de una distribución respecto a su media. Si tenemos un conjunto de valores de una variable, la varianza se calcula de la siguiente forma:

xi : cada dato
x: media de los datos
n: número de datos

Esto es el promedio de la diferencia elevada al cuadrado entre cada valor y la media

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

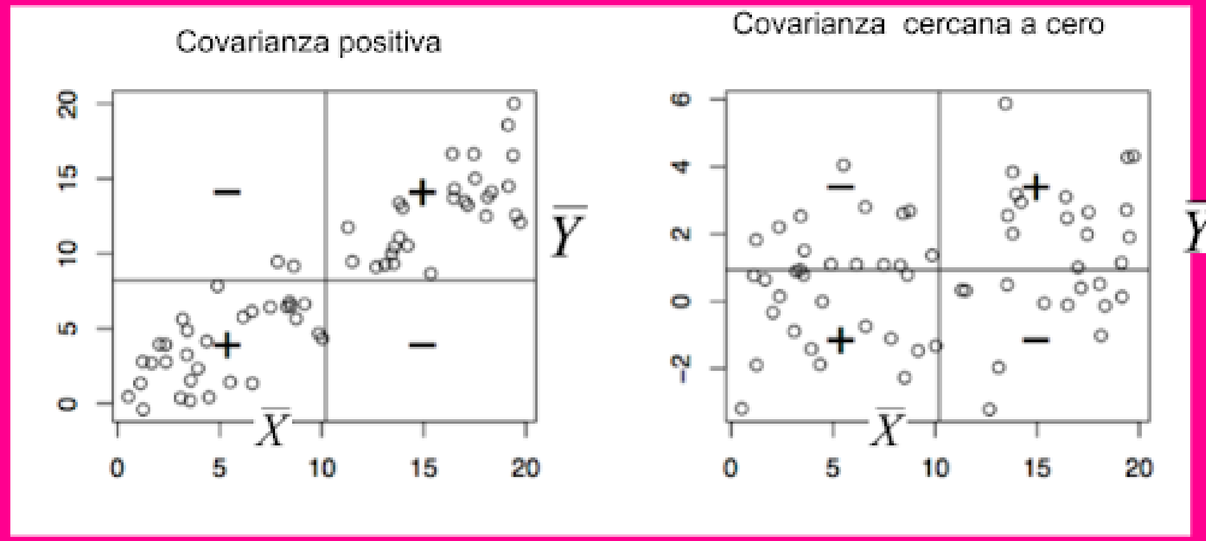
Medidas de Dispersión o Variabilidad: Desvío estándar

El desvío es una medida de la dispersión de los datos constituida por la raíz cuadrada de la varianza. No es la desviación promedio con respecto de la media.

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

Medidas de asociación lineal entre variables: Covarianza

La covarianza mide la asociación lineal entre ambas variables, es decir, qué tanto se asemeja la relación con una función lineal.
Decimos que dos variables X e Y, tienen covarianza positiva cuando tienden a encontrarse por encima de su media al mismo tiempo y tienen covarianza negativa cuando al mismo tiempo, tienden a encontrarse una por debajo y otra por encima. En cambio X e Y tienen covarianza cercana a cero cuando las variables pueden encontrarse por encima o por debajo de su media independientemente de lo que haga la otra.



Ejercicio de clase:

Dataset de recorridos de bicicleta 2021

- Cálculo de estadísticas descriptivas con numpy y pandas
- ¿Cómo podemos hacer para identificar distribuciones sin graficarlas?

Ejercicio práctico:

1- Cargar el recurso Molinetes - Agosto 2021 del dataset homónimo en google colaboratory
2- Calcular el promedio de pasajeros cada 15 minutos en el mes de agosto de 2021 en la estación Corrientes
3- Calcular el desvío estándar cada 15 minutos de pasajeros en el mes de agosto de 2021 en la estación corrientes
4- ¿Cuál es la cantidad total de pasajeros que viajaron el 1/8/2021?

Ejercicio práctico:

1- Usar query == para identificar la estación
2- Usar método .sum y .std para calcular estadísticas
3- Usar la query == para filtrar la fecha de interes.
4- Usar el método .sum() en la columna que se desea sumar