1. Promedio (Media)

El **promedio** o **media aritmética** es una medida de *tendencia central*, es decir, nos dice dónde se "centran" los datos.

Definición formal: Si tenemos (n) valores: (x_1, x_2, ..., x_n), el promedio se calcula como:

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

Ejemplo: Si las propinas son [5, 10, 15], el promedio es:

$$\left[\frac{5 + 10 + 15}{3} = 10\right]$$

El promedio resume el valor típico, pero **puede ser engañoso** si hay valores muy extremos (outliers).

2. Varianza y Desviación Estándar

Ambas son medidas de **dispersión**: nos indican qué tan separados o qué tan concentrados están los datos alrededor del promedio.

Varianza

La **varianza** mide el *promedio de las diferencias al cuadrado* respecto a la media:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Es decir:

- Si los datos están cerca del promedio, la varianza es baja.
- Si los datos están muy dispersos, la varianza es alta.

Desviación estándar

La **desviación estándar** es simplemente la **raíz cuadrada de la varianza**:

$$\sigma = \sqrt{Var(X)}$$

Se usa más que la varianza porque tiene las mismas **unidades** que los datos originales.

Ejemplo sencillo: Supongamos notas [5, 5, 5] → media = 5, desviación estándar = 0 (no hay variación). Si son [3, 5, 7] → media = 5, desviación estándar > 0 (hay dispersión).

A) Definición de Varianza

La **varianza** es una medida estadística que nos dice **cuánto se dispersan los datos alrededor de la media**.

- Si todos los datos están **muy cerca de la media**, la varianza es **pequeña**.
- Si los datos están **muy dispersos**, la varianza es **grande**.

B) Fórmula

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Qué significa cada parte:

- (x_i): cada valor individual de los datos.
- (\bar{x}): la media de los datos.
- (x_i - \bar{x}): cuánto se aleja cada dato de la media.
- ((x_i - \bar{x})^2): se eleva al cuadrado para que las diferencias negativas no se cancelen con las positivas.
- (\frac{1}{n}): se divide entre el número de datos para obtener un promedio de esas diferencias al cuadrado.

C) Ejemplo sencillo

Supongamos que tenemos las notas de **3 estudiantes**:

X = [4, 7, 10]

Paso 1: calcular la media

$$\overline{x} = \frac{4 + 7 + 10}{3} = \frac{21}{3} = 7$$

Paso 2: calcular las diferencias respecto a la media

- Para 4 → (4 - 7 = -3)
- Para 7 → (7 - 7 = 0)
- Para 10 → (10 - 7 = 3)

Paso 3: elevar al cuadrado esas diferencias

- $(-3)^2 = 9$
- $(0)^2 = 0$
- $(3)^2 = 9$

Paso 4: sacar el promedio

$$\overline{Var(X) = \frac{9 + 0 + 9}{3} = \frac{18}{3} = 6}$$

La **varianza es 6**.

D) Interpretación

- La varianza **no está en las mismas unidades** que los datos originales, porque trabajamos con cuadrados.
    - Ejemplo: si los datos están en "notas", la varianza está en "notas^2".
- Por eso, muchas veces se prefiere la **desviación estándar** (raíz cuadrada de la varianza), que vuelve a las unidades originales.

En este ejemplo: $\sigma = \sqrt{Var(X)} = \sqrt{6} \approx 2.45$

Esto significa que, en promedio, las notas se alejan **2.45 puntos** de la media (7).

E) ¿Para qué sirve la varianza?

La varianza y la desviación estándar son **fundamentales en estadística** porque nos dan una idea de la **variabilidad** de los datos.

- En un restaurante (ejemplo de *tips*), podríamos comparar si las propinas son **consistentes** (baja varianza) o **muy variables** (alta varianza).
- En educación, ver si todos los alumnos tienen notas parecidas (baja varianza) o hay mucha diferencia entre ellos (alta varianza).
- En ciencia de datos y machine learning, se usa para **normalizar, comparar distribuciones y detectar outliers**.

Resumen: La varianza **nos dice si los datos son homogéneos o heterogéneos**, y junto con la media, son la base de casi toda la estadística.

3. Estandarización (o normalización Z-score)

Es un proceso para transformar los datos de manera que tengan:

- Media = 0
- Desviación estándar = 1

Fórmula: $z = \frac{x-\overline{x}}{\sigma}$

Sirve para:

- Comparar variables que tienen **escalas diferentes** (ej. facturas en $100s y propinas en $10s).
- Facilitar algoritmos estadísticos y de machine learning que asumen datos en escalas similares.

Ejemplo: Si alguien gastó $20 en propina, pero el promedio de propinas es $10 y la desviación estándar es $5: $z = \frac{20-10}{5} = 2$ Significa que esa propina está **2 desviaciones estándar por encima de la media**.

Normalización vs estandarización

En datasets las variables a menudo vienen en escalas muy distintas (ej. edad en años [0–100], salario en miles [0–100000]). Muchos algoritmos estadísticos y de ML funcionan peor o convergen más despacio si las características tienen escalas diferentes. *Feature scaling* consiste en transformar las variables numéricas para que tengan una escala comparable.

Las dos transformaciones más usadas son:

- **Normalización (min–max scaling)** — reescalar a un rango fijo (p. ej. [0,1] o [-1,1]).
- **Estandarización (z-score)** — centrar en 0 y escalar por la desviación estándar (media = 0, desviación ≈ 1).

Fórmulas y explicación matemática

Min–Max (normalización)

Fórmula para reescalar a [0,1]:

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)}$$

Más general, para reescalar a ([a,b]):

$$x' = a + (x - \min(X)) \cdot \frac{b - a}{\max(X) - \min(X)}$$

Notas:

- Si max == min hay que manejar ese caso (p. ej. devolver 0 o una constante).
- Sensible a *outliers* (un valor extremo define el rango y "aplana" el resto).

Z-score (estandarización)

Fórmula:

$$z = \frac{x - \mu}{\sigma}$$

donde (\mu) es la media de la característica y (\sigma) su desviación estándar. Después de esto la característica tiene **media 0** y **desviación ≈ 1** (siempre que uses la misma convención para (\sigma)).

Estandarización (z-score):

1. Media: $(\mu = (10 + 20 + 30)/3 = 60/3 = 20.)$
2. Varianza (usando división por (N)): $(((10 - 20)^2 + (20 - 20)^2 + (30 - 20)^2)/3 = (100 + 0 + 100)/3 = 200/3 \approx 66.6666667.)$
3. Desviación: $\left(\sigma = \sqrt{200/3} \approx 8.1649658.\right)$
4. Z para cada valor:
    a. $(z_{10} = (10 - 20)/8.1649 \approx -1.2247449),$
    b. $(z_{20} = 0),$
    c. $(z_{30} \approx +1.2247449).$ $Resultado\ aproximado$: $([-1.2247,\ 0,\ +1.2247]).$

Diferencias prácticas y efectos en algoritmos

- **Rango**: Min–max fija un rango conocido (útil para redes neuronales con activaciones que esperan valores en [0,1] o [-1,1]). Z-score no fija un rango; conserva la forma de la distribución centrada en 0.
- **Outliers**: Min–max es muy sensible: un outlier puede comprimir el resto de observaciones. Z-score también se ve afectado por outliers porque inflan (\sigma), pero menos que min–max en cuanto a rango final.

4. La suma acumulada

En estadística y matemáticas, la **suma acumulada** (o *cumulative sum*) de una secuencia es otra secuencia en la que cada término es la suma de todos los elementos hasta esa posición.

Formalmente, si tenemos una lista de valores: $[x_1, x_2, x_3, \dots, x_n]$

La suma acumulada se define como: $y_k = \sum_{i=1}^{k} x_i\,, \quad k = 1,2, \dots, n$

Es decir:

- El primer valor acumulado es simplemente el primer elemento.
- El segundo es el primero + el segundo.
- El tercero es la suma de los tres primeros.

- Y así sucesivamente.

Ejemplo manual con ([2, 4, 6, 8])

- Primer valor: (2).
- Segundo: (2 + 4 = 6).
- Tercero: (2 + 4 + 6 = 12).
- Cuarto: (2 + 4 + 6 + 8 = 20).

Resultado manual: $[[2, , 6, , 12, , 20]]$

Con numpy.cumsum()

```
import numpy as np

valores = np.array([2, 4, 6, 8])
resultado = np.cumsum(valores)
print(resultado)  # [ 2  6 12 20]
```

El resultado es exactamente el mismo que al hacerlo a mano, pero más eficiente y con un código mucho más corto.

Interpretación estadística

El resultado te muestra cómo se **acumulan los valores a lo largo de la secuencia**.

Esto es útil en estadística y ciencia de datos para:

- **Series temporales**: ver la evolución acumulada de ventas, ingresos, precipitaciones, etc.
- **Distribuciones acumuladas**: la suma acumulada de frecuencias permite construir la **función de distribución acumulada (CDF)**, que indicaprobabilidad de que una variable aleatoria tome un valor menor o igual a un cierto punto.
- **Control de procesos**: detectar tendencias o acumulaciones en datos que puedan indicar un patrón.
- **Comparación de progresos**: por ejemplo, saber cuánto se lleva acumulado de un objetivo (ej. ahorro mensual).

5. LA CORRELACIÓN

Concepto básico

El índice de correlación mide **cuán fuerte y en qué dirección** se relacionan dos variables (por ejemplo, altura y peso, ingreso y gasto, etc.).

Su valor se denota comúnmente como **r** (si se trata del **coeficiente de correlación de Pearson**).

Rango de valores

$$-1 \leq r \leq 1$$

Valor de r	Interpretación
r = 1	Correlación positiva perfecta (cuando una variable aumenta, la otra también lo hace en proporción exacta).
r = -1	Correlación negativa perfecta (cuando una variable aumenta, la otra disminuye exactamente).
r = 0	No hay correlación lineal entre las variables.

Ejemplo intuitivo

- Si al subir la temperatura, las ventas de helados también aumentan, hay **correlación positiva** (r > 0).
- Si al subir la velocidad, el tiempo que tardas en llegar baja, hay **correlación negativa** (r < 0).

- Si no hay relación aparente entre el número de zapatos y la edad de una persona, **r ≈ 0**.

Fórmula del coeficiente de correlación de Pearson

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Donde:

- (x_i, y_i): valores individuales de las variables.
- (\bar{x}, \bar{y}): medias de cada variable.

Tipos de correlación

- **Correlación positiva** → ambas variables crecen o decrecen juntas.
- **Correlación negativa** → una crece mientras la otra decrece.
- **Correlación nula** → no hay relación lineal.
- **Correlación no lineal** → existe relación, pero no en línea recta (por ejemplo, forma de parábola).

Ejemplo: Calcular el índice de correlación (r de Pearson)

Supongamos que tenemos los siguientes datos:

Persona	X (horas de estudio)	Y (nota del examen)
A	2	65
B	3	70
C	4	75
D	5	85

E	6	90

Queremos saber **qué tan relacionadas están las horas de estudio con la nota del examen.**

A) Calculamos las medias

$$\left[\overline{X} = \frac{2+3+4+5+6}{5} = 4\right]\left[\overline{Y} = \frac{65+70+75+85+90}{5} = 77\right]$$

B) Restamos las medias a cada valor

X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
2	65	-2	-12	24	4	144
3	70	-1	-7	7	1	49
4	75	0	-2	0	0	4
5	85	1	8	8	1	64
6	90	2	13	26	4	169

C) Sumamos las columnas necesarias

$$\sum(X - \bar{X})(Y - \bar{Y}) = 24 + 7 + 0 + 8 + 26 = 65$$

$$\sum(X - \bar{X})^2 = 4 + 1 + 0 + 1 + 4 = 10$$

$$\sum(Y - \bar{Y})^2 = 144 + 49 + 4 + 64 + 169 = 430$$

D) Aplicamos la fórmula

$$r = \frac{65}{\sqrt{10 \times 430}} = \frac{65}{\sqrt{4300}} = \frac{65}{65.57} \approx 0.992$$

Resultado:

$$r \approx 0.99$$

Esto indica una **correlación positiva muy fuerte** entre las horas de estudio y la nota del examen. (Es decir, a más horas de estudio, mejor calificación.)