

Estadística aplicada

Introducción a la estadística

Clase 01

Temario del curso

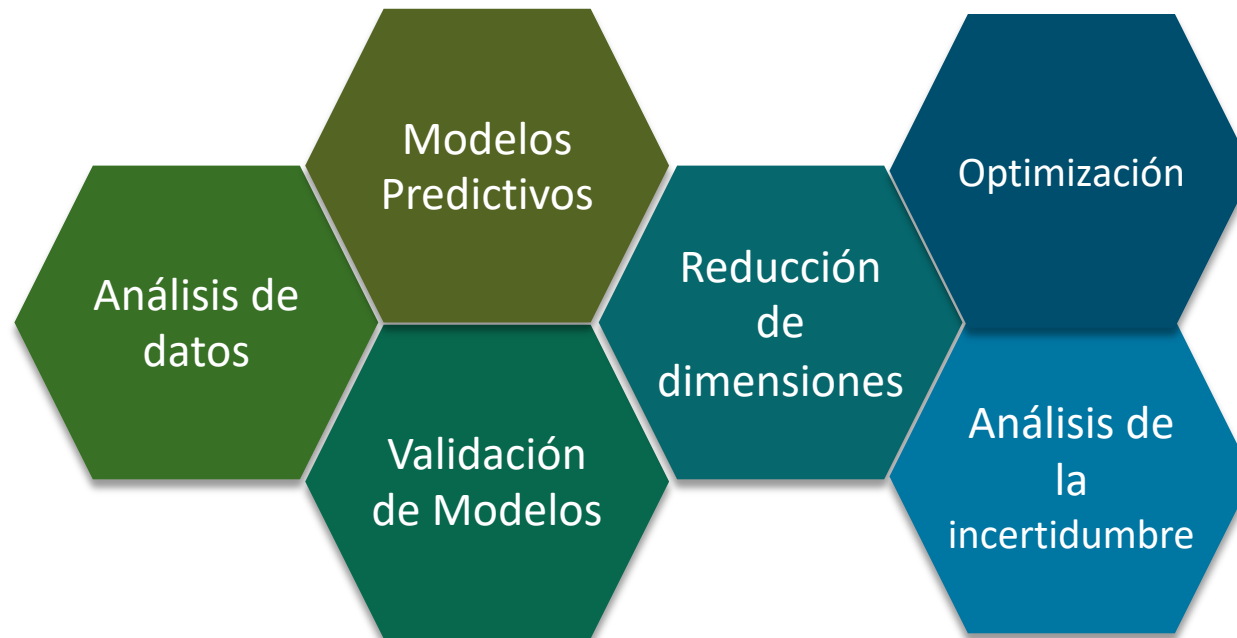
- Introducción a la Estadística
- Estadística Descriptiva
- Gráficos y Visualización de Datos
- Probabilidad
- Distribuciones de Probabilidad
- Correlación y Regresión
- Métodos de Muestreo
- *Aplicaciones de la Estadística en Redes Neuronales*
- Ejercicio práctico

Importancia de la estadística en Redes Neuronales

¿Por qué es importante?

La estadística es **crucial** para la inteligencia artificial porque **proporciona las herramientas y métodos** necesarios para **analizar datos, construir y validar modelos**, hacer inferencias, manejar la incertidumbre y optimizar algoritmos. Sin la estadística, muchos de los avances en IA no serían posibles.

Aplicaciones de la estadística en IA



¿Qué es la estadística?

¿Qué es la estadística?

La estadística es una **rama de las matemáticas** que se encarga de **recolectar, analizar, interpretar, presentar y organizar datos**. Es una herramienta fundamental en muchas disciplinas, como la economía, la biología, la ingeniería, la psicología, la sociología, **la inteligencia artificial** y muchas otras, ya que **permite tomar decisiones informadas** basadas en datos y tendencias.

Descriptiva

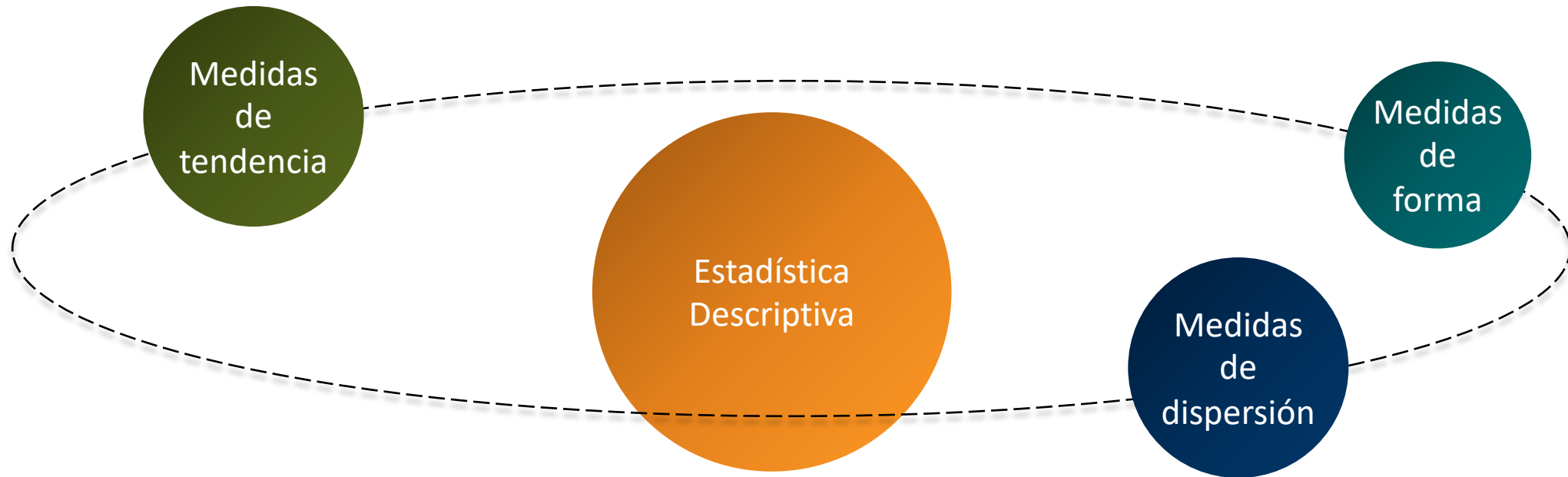
Esta área se enfoca en **describir y resumir un conjunto de datos**. Utiliza **medidas** como la media, la mediana, la moda, la desviación estándar, y **gráficos** como histogramas y diagramas de dispersión para representar la información de manera comprensible.

Inferencial

Esta área se centra en hacer inferencias o **predicciones** sobre una población **a partir de una muestra de datos**. Utiliza técnicas como la estimación de intervalos de confianza, pruebas de hipótesis, análisis de regresión, y análisis de varianza **para llegar a conclusiones sobre la población de interés**.

Estadística descriptiva

Esta área se enfoca en **describir y resumir un conjunto de datos**. Utiliza **medidas** como la media, la mediana, la moda, la desviación estándar, y **gráficos** como histogramas y diagramas de dispersión para representar la información de manera comprensible.



Medidas de tendencia: Media

¿Qué es la media?

La media aritmética, es una medida de **tendencia central** que indica el valor promedio de un conjunto de datos. Se utiliza para representar el valor típico de un conjunto de datos y es una de las medidas más comunes y útiles en la estadística descriptiva

¿Cómo se calcula?

$$Media(x) = \frac{1}{n} \sum_{i=1}^n x_i$$

Otros tipos de media

Media Geométrica

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Media armónica

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

Media ponderada

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Media cuadrática

$$\bar{x}_q = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}$$

Medidas de tendencia: Mediana

¿Qué es la mediana?

Es una medida de tendencia central que **representa el valor del medio en un conjunto de datos ordenados**. A diferencia de la media, que puede verse afectada por valores extremadamente altos o bajos (outliers), **la mediana proporciona una mejor representación del centro de un conjunto de datos asimétricos** o con valores atípicos.

¿Cómo se calcula?

1

Ordeno de menor a mayor la muestra

2

Elijo el elemento que se encuentra en el medio

Si el número de elementos de la muestra es **impar**, por ejemplo 99 elemento. Tomo el que ocupe la posición $(n+1) / 2$ es decir el 50

Si el número de elementos de la muestra es **par**, por ejemplo 100 elemento. Tomo la media de los valores que ocupen la posición $n/2$ y $((n/2)+1)/2$, la media entre los valores de las posiciones 50 y 51

Medidas de tendencia: Moda

¿Qué es la Moda?

La moda en estadística se refiere al valor que aparece con mayor frecuencia en un conjunto de datos. Es simplemente **el número que se repite más** a menudo. La moda es útil porque proporciona una **idea rápida de cuál es el valor más típico** en un conjunto de datos. Sin embargo, a diferencia de la media y la mediana, la moda no necesariamente representa un valor central o típico

¿Cómo se calcula?

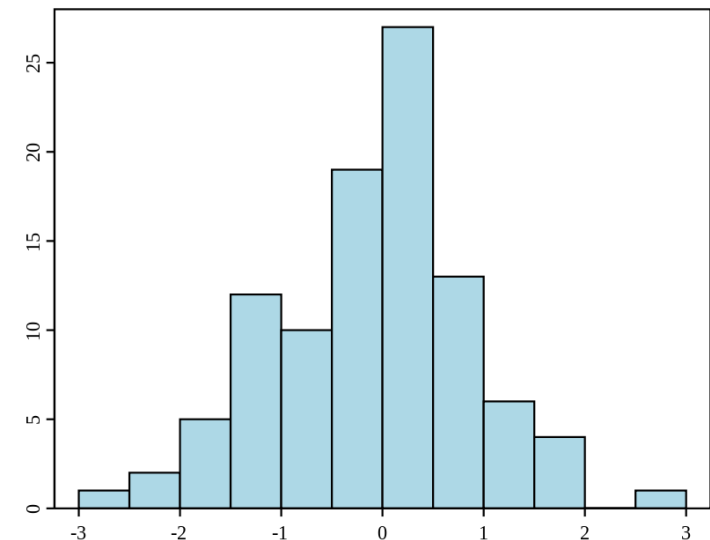
1

Se cuentan las repeticiones de los elementos distintos de la muestra

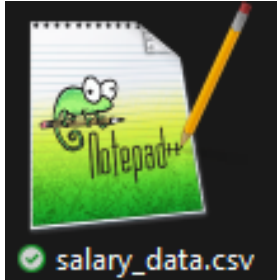
2

El elemento que más repeticiones tenga es la moda

Histograma



Dataset de trabajo



El conjunto de datos contiene un informe salarial fabricado sobre el que se pueden realizar análisis para intentar estimar el salario en función de las condiciones dadas. Hay 100.000 puntos de datos, 50.000 de ellos son mujeres y los otros 50.000 son hombres.

Columnas

ID	# income	# age	A gender	# education_level
Identifier of the candidate	The annual income declared by the person	Age of the person at the moment of the test	Gender declared by the person	Education level declared by the user (0: primary complete, 1: secondary complete, 2: tertiary complete, 3: post)

Enlace a dataset

<https://www.kaggle.com/datasets/micheldc55/anual-salary-reports-survey>

Programar Medidas de tendencia



EstadisticaBasica_1.py

- Media
- Mediana
- Moda

Medidas de Dispersión: Rango

¿Qué es el Rango?

El rango es una medida de **dispersión** que indica la **diferencia entre el valor máximo y el valor mínimo de un conjunto de datos**. Es una de las formas más sencillas de medir la variabilidad de los datos

¿Cómo se calcula?

1

Se identifica el mayor de los valores $X_{máx}$ de la muestra

2

Se identifica el menor de los valores X_{min} de la muestra

3

La diferencia entre ambos valores es el rango:

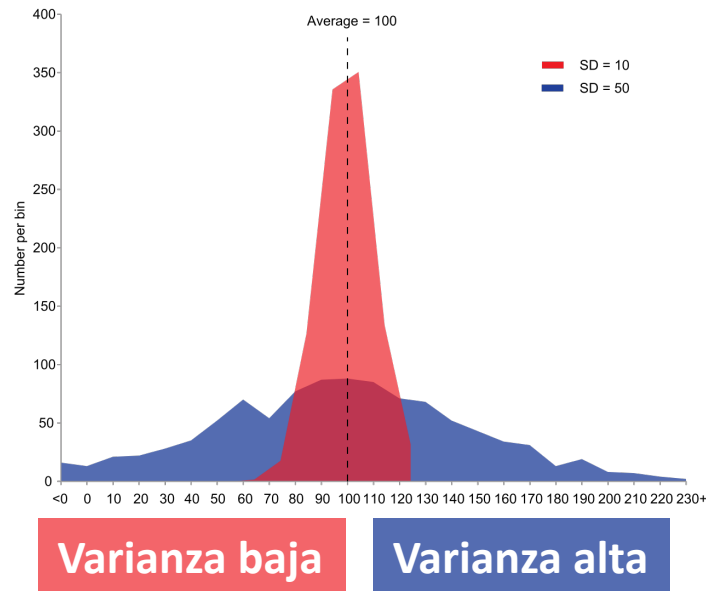
$$Rango = X_{máx} - X_{min}$$

Medidas de Dispersión: Varianza

¿Qué es el Varianza?

La varianza es una medida de dispersión que **indica cuánto varían los datos en un conjunto con respecto a la media**. Específicamente, la varianza **mide la media de las diferencias al cuadrado entre cada valor del conjunto y la media del conjunto**.

- Una **varianza alta** indica que **los datos están muy dispersos** alrededor de la media.
- Una **varianza baja** indica que **los datos están más concentrados** cerca de la media.



¿Cómo se calcula?

- 1 Se calcula la media de la población

$$\text{Media} = \frac{1}{n} \sum_{i=1}^n x_i$$

- 2 Sumamos las distancias entre cada elemento y la media elevados al cuadrado

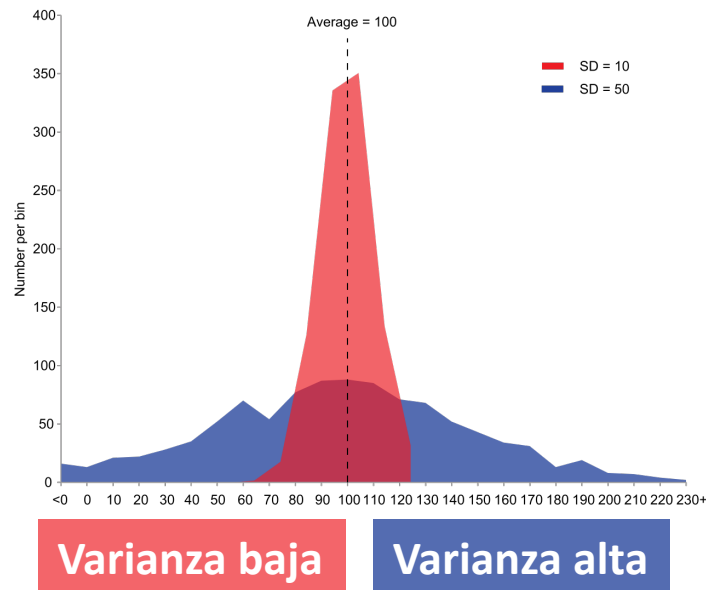
$$\sigma^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \text{Media})^2$$

Medidas de Dispersión: Desviación Estándar

¿Qué es la desviación estándar?

La desviación estándar es una medida de dispersión **que indica cuánto se desvían, en promedio, los valores de un conjunto de datos con respecto a su media**. A diferencia de la varianza, la **desviación estándar tiene las mismas unidades** que los datos originales, lo que facilita la interpretación.

- Una **desviación estándar alta** indica que **los datos están muy dispersos** alrededor de la media.
- Una **desviación estándar baja** indica que **los datos están más concentrados** cerca de la media.



¿Cómo se calcula?

- 1 Se calcula la media de la población

$$\text{Media} = \frac{1}{n} \sum_{i=1}^n x_i$$

- 3 Hago la raíz cuadrada de la varianza

- 2 Sumamos las distancias entre cada elemento y la media elevados al cuadrado

$$\sigma^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \text{Media})^2$$

$$\sigma = \sqrt{\sigma^2}$$

Medidas de Dispersión: Percentiles

¿Qué es son los percentiles?

Los percentiles son medidas estadísticas que **dividen un conjunto de datos ordenados en 100 partes iguales**. Cada percentil indica **el valor debajo del cual se encuentra un cierto porcentaje de los datos**. Por ejemplo, **el percentil 25 (P25) es el valor debajo del cual se encuentra el 25% de los datos**, el percentil 50 (**P50**) es el valor debajo del cual se encuentra el 50% de los datos (**también conocido como la mediana**)

¿Por qué son importantes?

1

Análisis de distribución: Los percentiles permiten comprender cómo se distribuyen los datos. Por ejemplo, conocer el percentil 90 puede ayudarte a saber qué valor supera el 90% de los datos.

2

Identificación de valores atípicos: Los percentiles pueden ayudar a identificar valores atípicos o extremos en los datos. Por ejemplo, los valores por debajo del percentil 5 o por encima del percentil 95 pueden considerarse atípicos.

Programar Medidas de Dispersión



EstadisticaBasica_2.py

- Rango
- Varianza
- Desviación Estándar
- Percentiles

Medidas de Forma: Asimetría

¿Qué es la Asimetría?

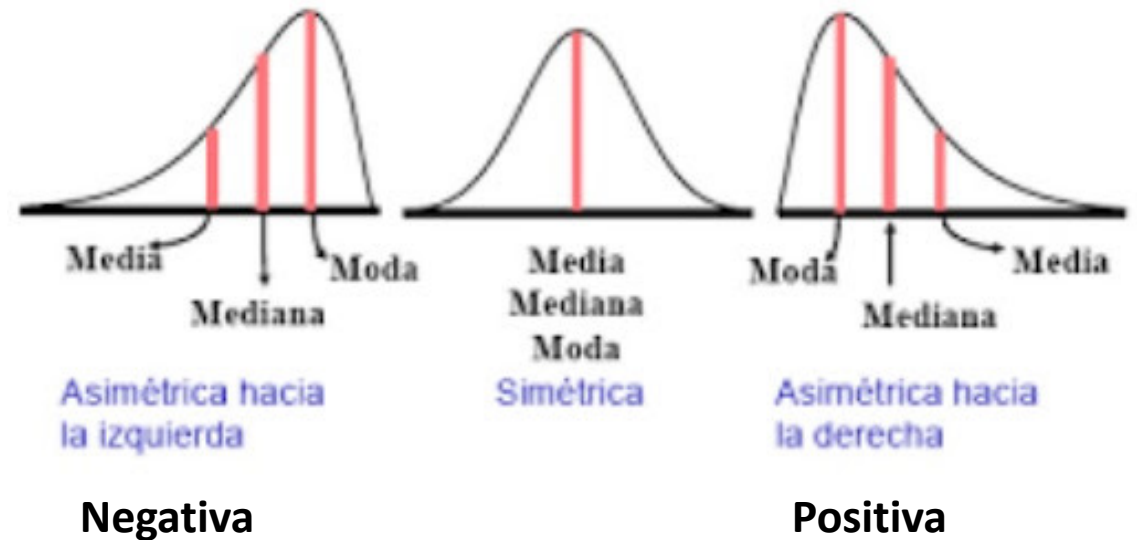
Es una medida que describe la **falta de simetría en la distribución** de los datos. Indica la dirección y el grado de desbalance en una distribución. **En una distribución simétrica, los datos están distribuidos de manera uniforme alrededor de la media.** Sin embargo, en una **distribución asimétrica, los datos pueden estar más concentrados en uno de los extremos.**

Asimetría Negativa (o Sesgo a la Izquierda):

1. La **cola** de la distribución es más larga en el lado **izquierdo**.
2. La **mayoría de los valores** se encuentran en la parte **derecha de la distribución**.
3. Ejemplo: edades a la jubilación en una población.

Asimetría Positiva (o Sesgo a la Derecha):

1. La **cola** de la distribución es más larga en el lado **derecho**.
2. La **mayoría de los valores** se encuentran en la parte **izquierda de la distribución**.
3. Ejemplo: ingresos en una población donde pocos individuos tienen ingresos muy altos.



Medidas de Forma: Asimetría

¿Cómo se calcula la asimetría?

La asimetría se mide comúnmente usando el coeficiente de asimetría. Se pueden utilizar varias fórmulas para calcularlo, pero una fórmula comúnmente utilizada es la del **coeficiente de asimetría de Pearson**:

Fórmula e interpretación

$$\text{Asimetría} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s} \right)^3$$

donde:

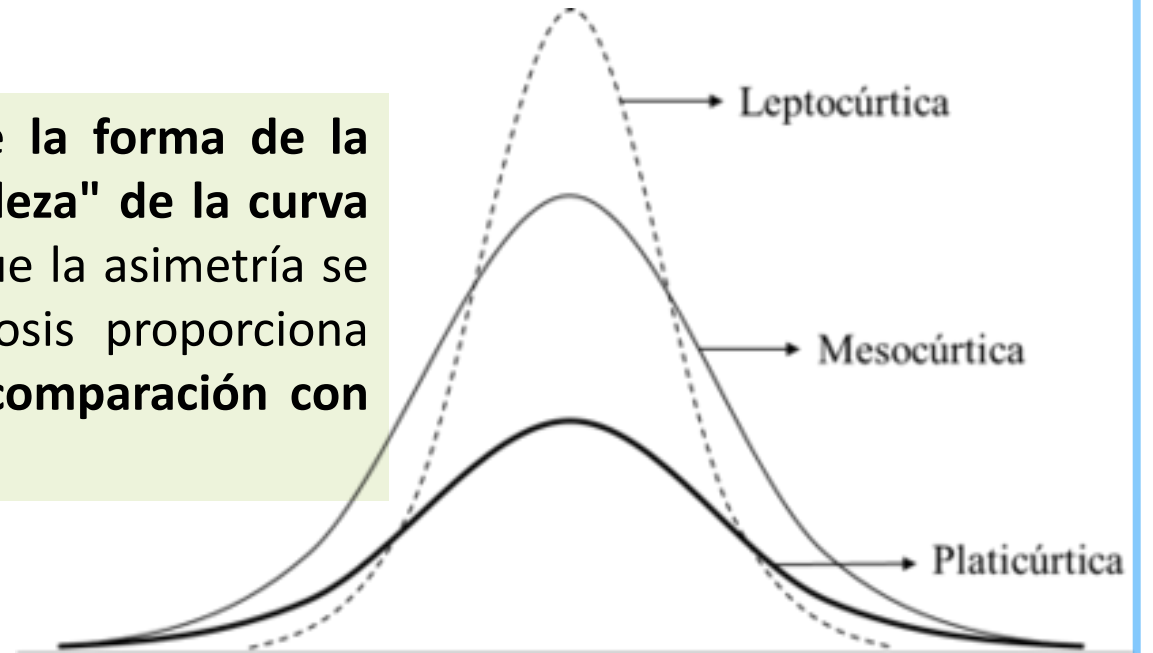
- n es el número de datos.
- X_i es cada valor en el conjunto de datos.
- \bar{X} es la media del conjunto de datos.
- s es la desviación estándar del conjunto de datos.

La asimetría es una medida importante porque proporciona información **sobre la forma de la distribución de los datos**. Ayuda a los analistas a comprender **si los datos están sesgados hacia un lado y cómo esto puede afectar los resultados** y las interpretaciones de los análisis estadísticos. Por ejemplo, **una alta asimetría puede indicar la presencia de valores atípicos** que pueden influir significativamente en la media y otras medidas estadísticas.

Medidas de Forma: Curtosis

¿Qué es la Curtosis?

La curtosis en estadística es una medida describe **la forma de la distribución de los datos**, específicamente la **"agudeza" de la curva en su pico y la "pesadez" de sus colas**. Mientras que la asimetría se centra en la simetría de la distribución, la curtosis proporciona información sobre **el perfil de la distribución en comparación con una distribución normal** (campana de Gauss)



Tipos de Curtosis

1 Mesocúrtica

- **Valor de curtosis = 0**
- Describe una distribución que tiene una forma **similar a la distribución normal**.

2 Leptocúrtica

- Valor de curtosis > 0
- La distribución tiene un pico más alto y **colas más pesadas que la distribución normal**.
- Indica más valores extremos
- Ejemplo: Distribuciones con **valores atípicos más frecuentes**.

3 Platicúrtica

- Valor de curtosis < 0
- La distribución tiene **un pico más bajo y colas más ligeras** que la distribución normal.
- Indica menos valores extremos
- Ejemplo: Distribuciones más planas y anchas

Medidas de Forma: Curtosis

¿Cómo se calcula la Curtosis?

La fórmula para la curtosis es **momento de cuarto orden** de una distribución de datos

Fórmula de cálculo


$$\text{Curtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

donde:

- n es el número de datos.
- X_i es cada valor en el conjunto de datos.
- \bar{X} es la media del conjunto de datos.
- s es la desviación estándar del conjunto de datos.

El término -3 se resta para que la curtosis de una distribución normal sea 0 (exceso de curtosis). Si no se resta, la curtosis de una distribución normal es 3.

Programar Medidas de Forma

 EstadísticaBasica_3.py

- Asimetría
- Curtosis