

guayerd

Fundamentos IA

Análisis con Python

Clase 7

En colaboración con
IBM SkillsBuild





¡Bienvenidos!

¿Nos presentamos?

- ¿Qué recuerdan de la clase anterior?
- ¿Qué esperan aprender?
- ¿Tienen alguna pregunta?

Contenidos

Por temas

05

- Copilot Chat y prompts
- Demo asincrónica

06

- Limpieza y transformación

07

- Estadística aplicada

08

- Visualización

Objetivos de la clase



- Estadística descriptiva básica
- Distribuciones de datos
- Correlaciones

Análisis con Python

Estadística aplicada

Plataforma Skill Build: Python



eLearning


Data Visualization with Python

3 horas  1.849 ★★★★★ 150



eLearning

Utilizar la IA generativa para el desarrollo de software

1 hora  34.080 ★★★★★ 2.316

Estadística aplicada

Conjunto de **técnicas para entender y resumir datos**.

- Describe características principales
- Detecta patrones y tendencias
- Mide relaciones entre variables
- Soporta la toma de decisiones





La **estadística** es el arte y la ciencia de reunir datos, analizarlos, presentarlos e interpretarlos.

Esto ayuda a las personas que deben tomar decisiones una mejor comprensión del entorno, permitiéndoles así tomar mejores decisiones con base en mejor información.

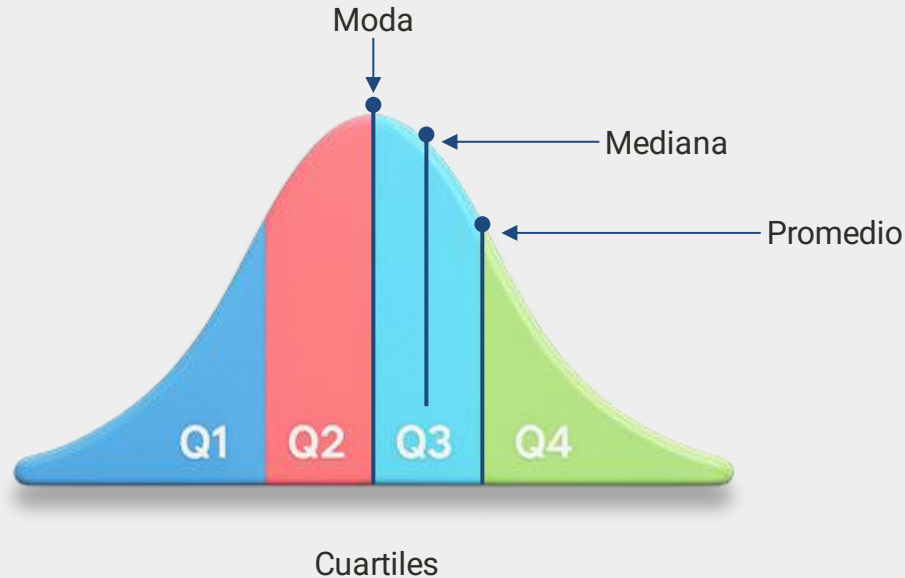
Exploración con Pandas

Herramientas para exploración y análisis estadístico

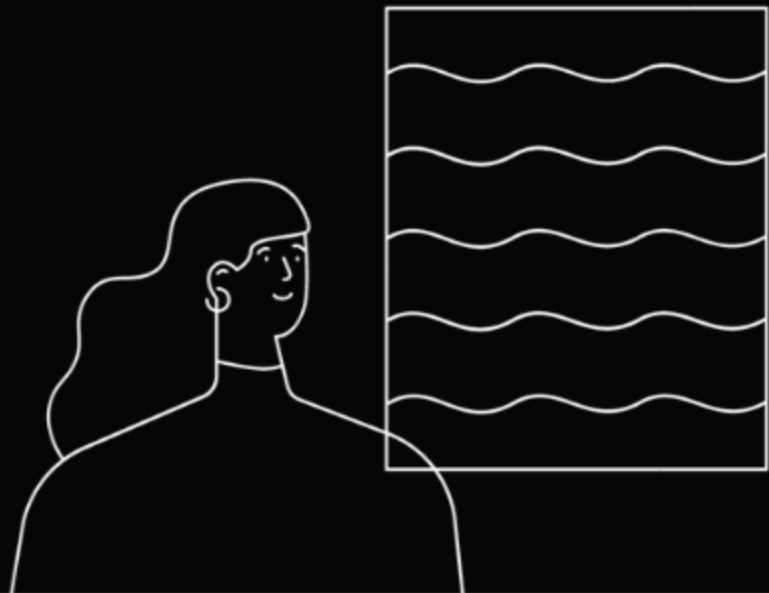
Función	Propósito	Resultado
<code>.describe()</code>	Resumen completo	Todas las estadísticas principales
<code>.info()</code>	Información general	Tipos y valores nulos
<code>.value_counts()</code>	Frecuencias	Conteo por categoría
<code>.groupby().agg()</code>	Estadísticas agrupadas	Métricas por segmento

Estadística descriptiva

Ejemplo



La mayor parte de la información estadística en periódicos, revistas, informes de empresas y otras publicaciones consta de datos que se resumen y presentan en una forma fácil de leer y de entender. A estos resúmenes de datos, que pueden ser tabulares, gráficos o numéricos se les conoce como **estadística descriptiva**.





Inferencia estadística

Una de las principales contribuciones de la estadística es emplear datos de una muestra para hacer estimaciones y probar hipótesis acerca de las características de una población mediante un proceso al que se le conoce como **inferencia estadística**.

Población

Cuando se examina un grupo entero o universo completo de observaciones.



Muestra

Cuando se examina una pequeña parte del grupo.

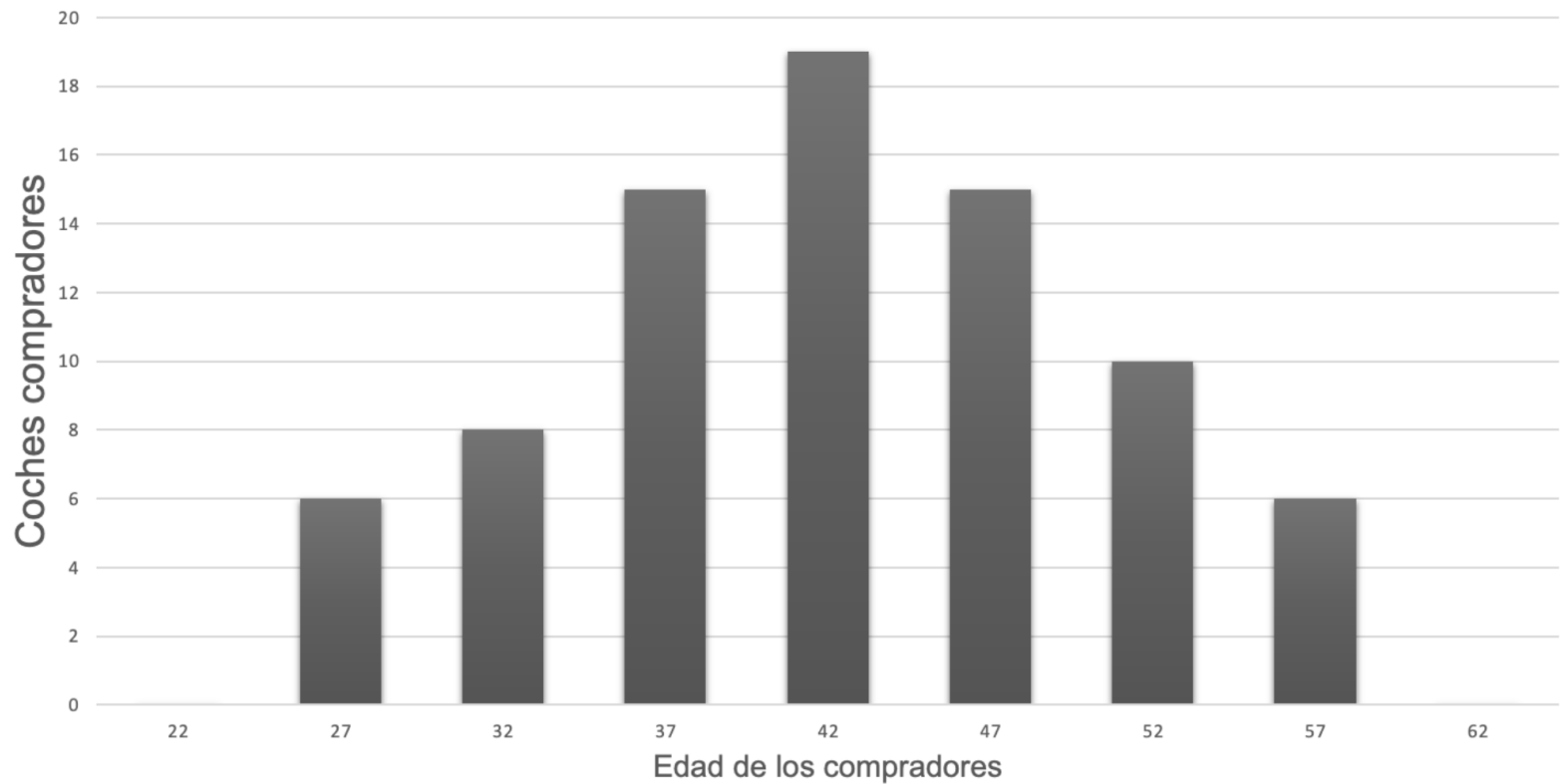


Distribución de frecuencias

- ✓ Forma de presentación de los datos que facilita su tratamiento conjunto y permite una comprensión diferente de ellos.
- ✓ Es una tabla de datos con base en observaciones (frecuencias).
- ✓ La frecuencia es el número de casos que pertenecen a un valor determinado.

Histograma

Gráfico de la distribución de frecuencias, que se construye con rectángulos de superficie proporcional al producto de la amplitud por la frecuencia absoluta (o relativa) de cada uno de los intervalos de clase.



Tendencia central

Se refiere al **punto medio** de una distribución.

El **sesgo** se produce cuando al trazar una línea vertical que pase por el punto más alto de la curva dividirá su área en dos partes que no son iguales.

Distribuciones de datos

Muestran la forma en que se organizan los valores dentro de un conjunto de datos.

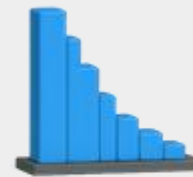
- **Normal:** campana simétrica
- **Sesgada:** cola hacia un lado
- **Bimodal:** dos picos de frecuencia
- **Multimodal:** múltiples picos de frecuencia
- **Uniforme:** frecuencias similares



Normal



Sesgada a la izquierda



Sesgada a la derecha



Uniforme



Bimodal



Multimodal

Identificación de distribución

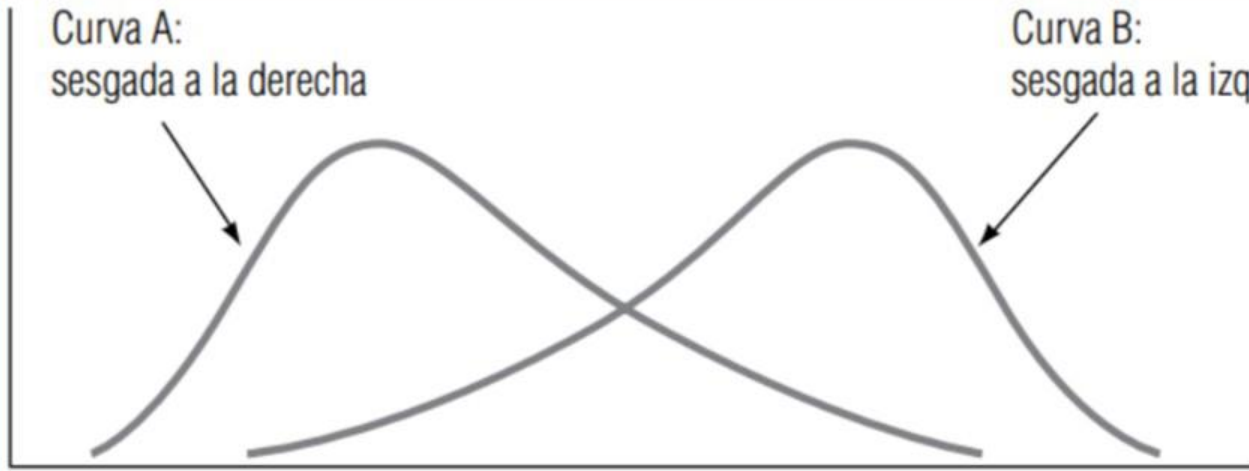
El tipo de distribución se deduce comparando media y mediana.

Tipo de Distribución	Relación Media–Mediana	Ejemplo Real
Normal	Media \approx Mediana	Alturas, pesos
Sesgada	Media muy diferente de mediana	Ingresos, precios
Bimodal	Depende de los picos	Horarios de tráfico
Multimodal	Variable según picos	Preferencias múltiples
Uniforme	Media \approx Mediana	Números aleatorios

Sesgos

Curva A:
sesgada a la derecha

Curva B:
sesgada a la izquierda



Media aritmética (Promedio)

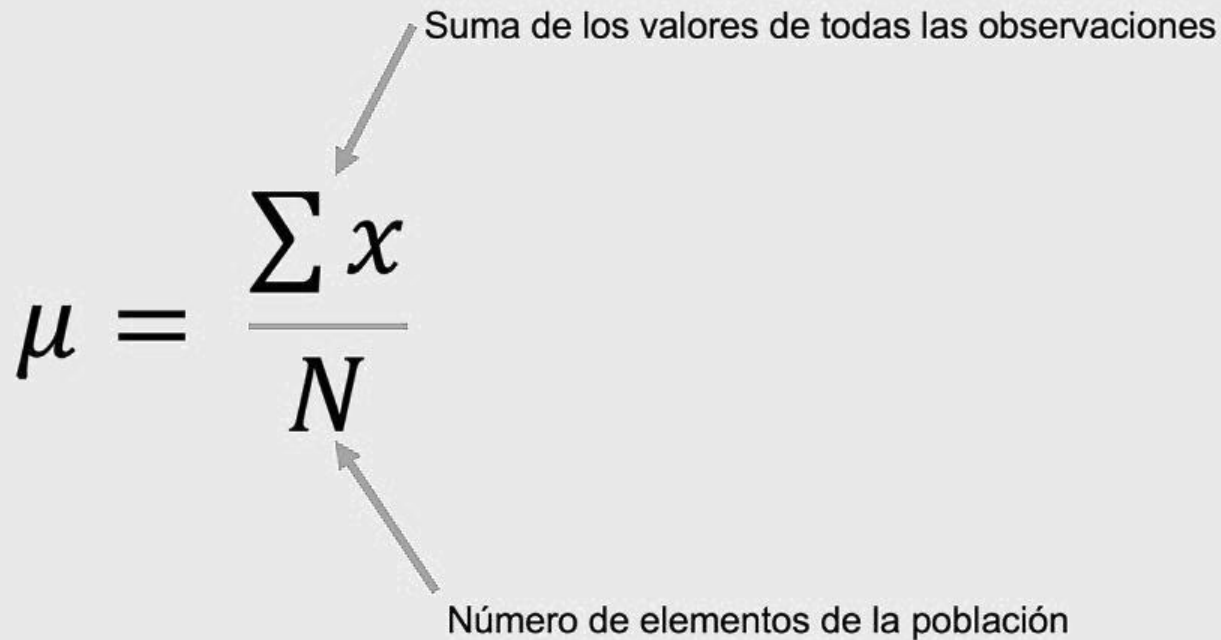
- ✓ Es la suma de los valores de todas las observaciones, dividido la cantidad de elementos de la muestra.

Media aritmética Población

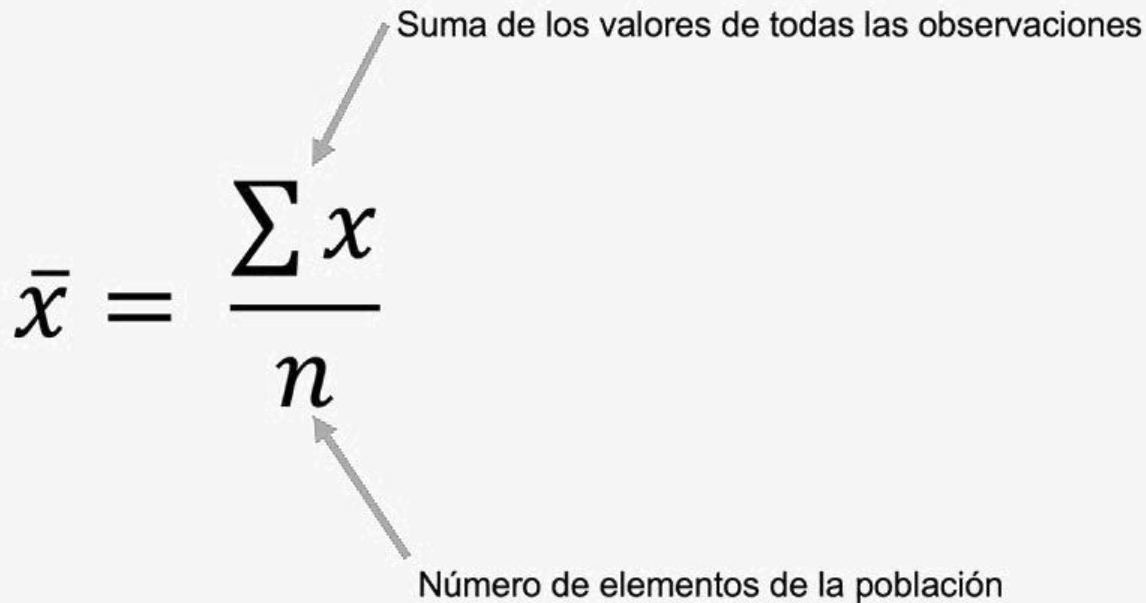
$$\mu = \frac{\sum x}{N}$$

Suma de los valores de todas las observaciones

Número de elementos de la población

The diagram shows the formula for the population arithmetic mean, $\mu = \frac{\sum x}{N}$. An arrow points from the text 'Suma de los valores de todas las observaciones' to the summation symbol \sum in the numerator. Another arrow points from the text 'Número de elementos de la población' to the variable N in the denominator.

Media aritmética de la muestra



The diagram shows the formula for the sample mean, $\bar{x} = \frac{\sum x}{n}$. An arrow points from the text 'Suma de los valores de todas las observaciones' to the summation symbol \sum . Another arrow points from the text 'Número de elementos de la población' to the variable n .

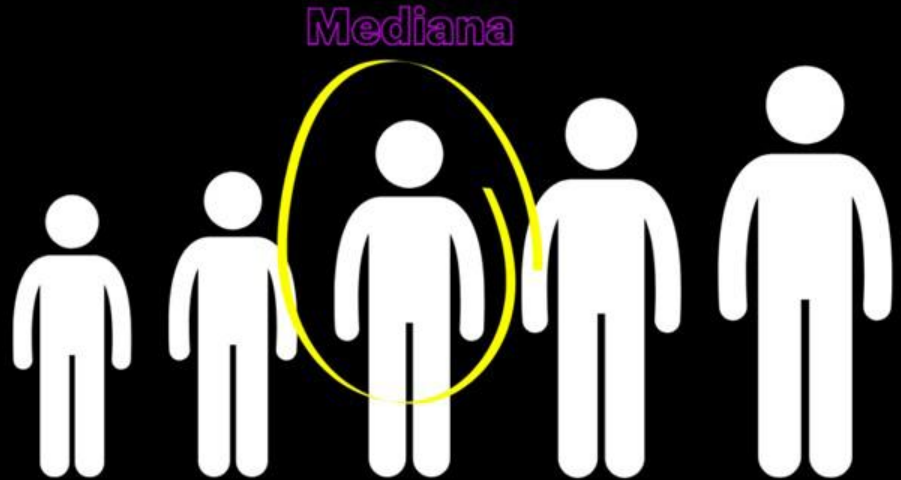
$$\bar{x} = \frac{\sum x}{n}$$

Suma de los valores de todas las observaciones

Número de elementos de la población

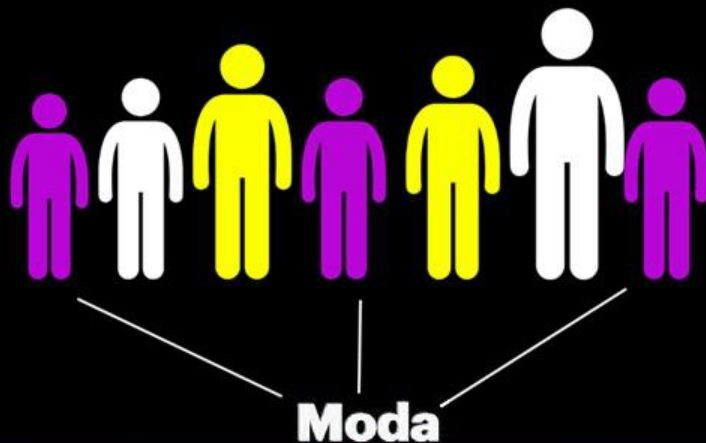
La mediana

- ✓ Mide la observación central del conjunto.
- ✓ Para hallar la mediana de un conjunto de datos, primero se organizan en orden descendente o ascendente.
- ✓ El elemento que está más al centro del conjunto de números, la mitad de los elementos están por arriba de este punto y la otra mitad está por debajo.

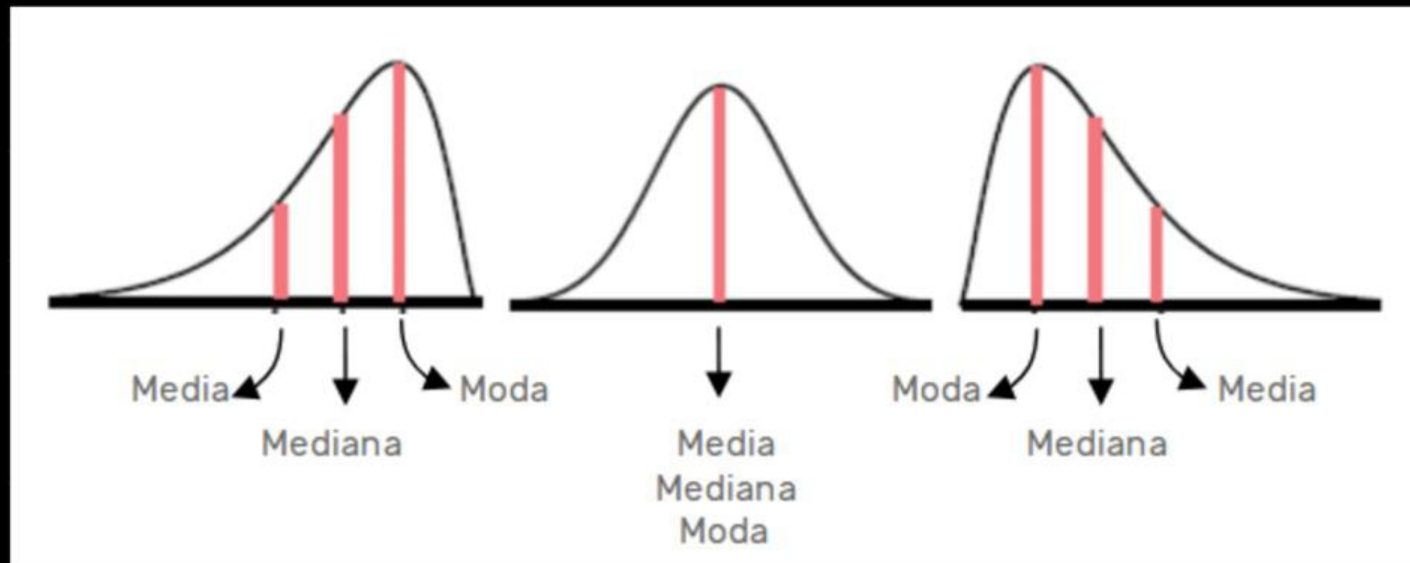


Moda

- ✓ La moda es el valor que más se repite en el conjunto de datos.



Media, Mediana, Moda



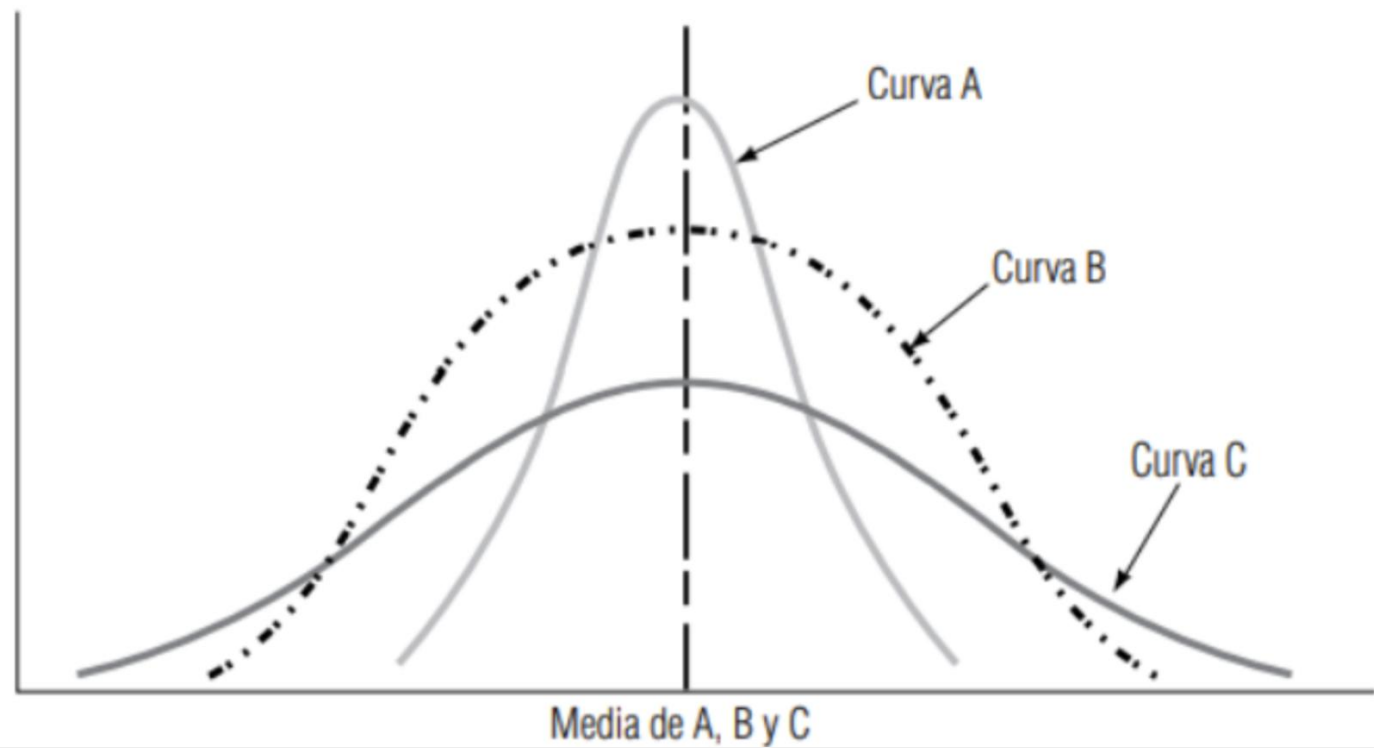
Al igual que sucede con cualquier conjunto de datos, la media, la mediana y la moda sólo nos revelan una parte de la información que debemos conocer acerca de las **características de los datos**. Para aumentar nuestro entendimiento del patrón de los datos, debemos medir también su



Dispersión

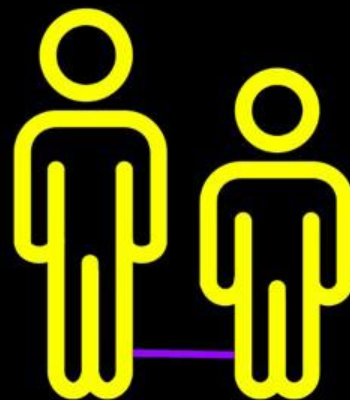
Separación

Variabilidad



El rango

El rango es la diferencia entre el más alto y el más pequeño de los valores observados



La Varianza

- ✓ Es la suma de los cuadrados de las distancias entre la media y cada elemento de la población, dividido entre el número total de observaciones.

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} = \frac{\sum x^2}{N} - \mu^2$$

La desviación estándar

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}} = \sqrt{\frac{\sum x^2}{N} - \mu^2}$$

Medidas descriptivas

Resumen numérico de características principales

Medida	Comando	Descripción
Media	<code>df['columna'].mean()</code>	Promedio aritmético
Mediana	<code>df['columna'].median()</code>	Valor central ordenado
Moda	<code>df['columna'].mode()</code>	Valor más frecuente
Desviación estándar	<code>df['columna'].std()</code>	Dispersión promedio

Medidas de posición

Ubicación de valores en la distribución

Medida	Comando	Interpretación
Mínimo	<code>df['columna'].min()</code>	Valor más bajo
Máximo	<code>df['columna'].max()</code>	Valor más alto
Cuartiles	<code>df['columna'].quantile([0.25, 0.5, 0.75])</code>	Divide datos en 4 partes
Rango	<code>df['columna'].max() - df['columna'].min()</code>	Amplitud total

¿Qué significan estos datos?

Dataset de salarios (en miles USD)



Media: 45, mediana: 38, moda: 35

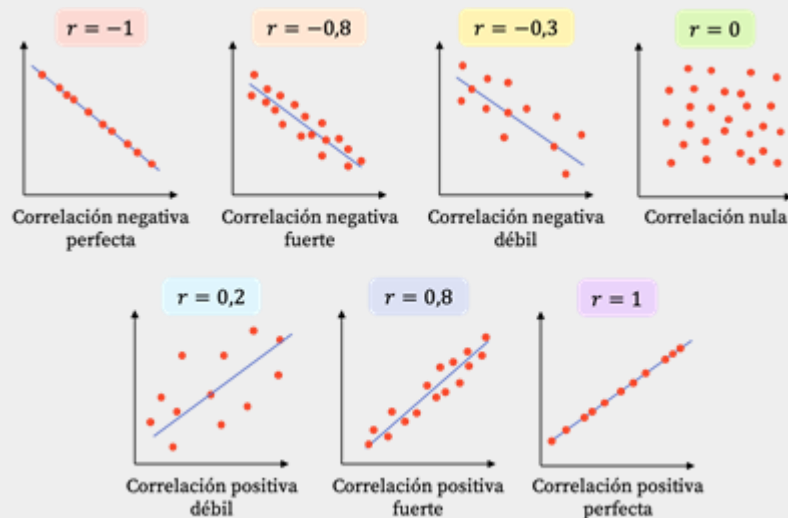
Desviación estándar: 15

Rango: 80 (min: 20, máx: 100)

Correlaciones

Medida de **qué tanto dos variables cambian** juntas.

- Valores entre -1 (inverso) y +1 (directo)
- 0 indica que no hay relación lineal
- **Comando:** `df[['var1', 'var2']].corr()`



Errores comunes de interpretación

Media vs. Mediana

- Diferencias grandes indican presencia de valores extremos
- Siempre reportar ambas métricas

Correlación \neq Causación

- Una correlación alta no implica causalidad
- Considerar variables ocultas que puedan explicar la relación

Evaluación de confiabilidad

Desviación estándar

- **Baja:** datos consistentes, resultados predecibles
- **Alta:** datos dispersos, mayor incertidumbre

Outliers (valores extremos)

- Analizar antes de eliminarlos
- Pueden ser errores de medición o información relevante

Rendimiento E-commerce



- Identificar mes con mayor eficiencia (ventas/gasto publicidad)
- Determinar mes con mejor tasa de conversión y analizar causa
- Calcular ticket promedio (ventas/productos) por mes
- Evaluar relación entre visitantes y ventas

Mes	Ventas (\$)	Visitantes	Conversión (%)	Gasto Publicidad (\$)	Productos Vendidos
Ene	45,000	15,000	3.2	8,500	450
Feb	52,000	18,200	2.9	9,800	520
Mar	38,000	12,500	3.8	7,200	380
Abr	61,000	20,500	3.1	11,200	610
May	48,000	16,800	2.7	9,500	480

Proyecto

Tienda Aurelion

- **Documentación:** notebook Markdown
- **Desarrollo técnico:** programa Python
- **Visualización de datos:** dashboard en Power BI
- **Presentación oral:** problema, solución y hallazgos



Análisis estadístico descriptivo

Trabajo en equipo



1. Calcular **estadísticas básicas**
2. Identificar **tipo de distribución**
3. Calcular **correlaciones** entre variables principales
4. Analizar **outliers**
5. **Interpretar resultados** para el problema de negocio
6. **Documentar con Copilot** cada paso y resultado



Retro

¿Cómo nos vamos?

- ¿Qué fue lo más útil de la clase?
- ¿Qué parte te costó más?
- ¿Qué te gustaría repasar o reforzar?