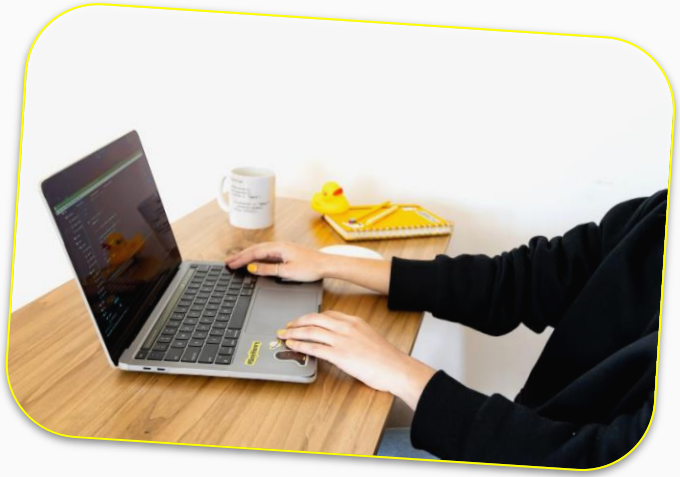
HENRY

Estadística Descriptiva

Data Science



Agenda

- Tipos de estadística
- Población y muestra
- Distribución de frecuencias
- Histograma
- Tendencia
- Media, mediana y moda
- Explicación de la HW

OBJETIVOS DE LA CLASE

Al finalizar esta lecture estarás en la capacidad de...

- ➜ Conocer los conceptos fundamentales de la Estadística.
- ➜ Comprender el uso de la estadística aplicada con Python

¡COMENCEMOS!

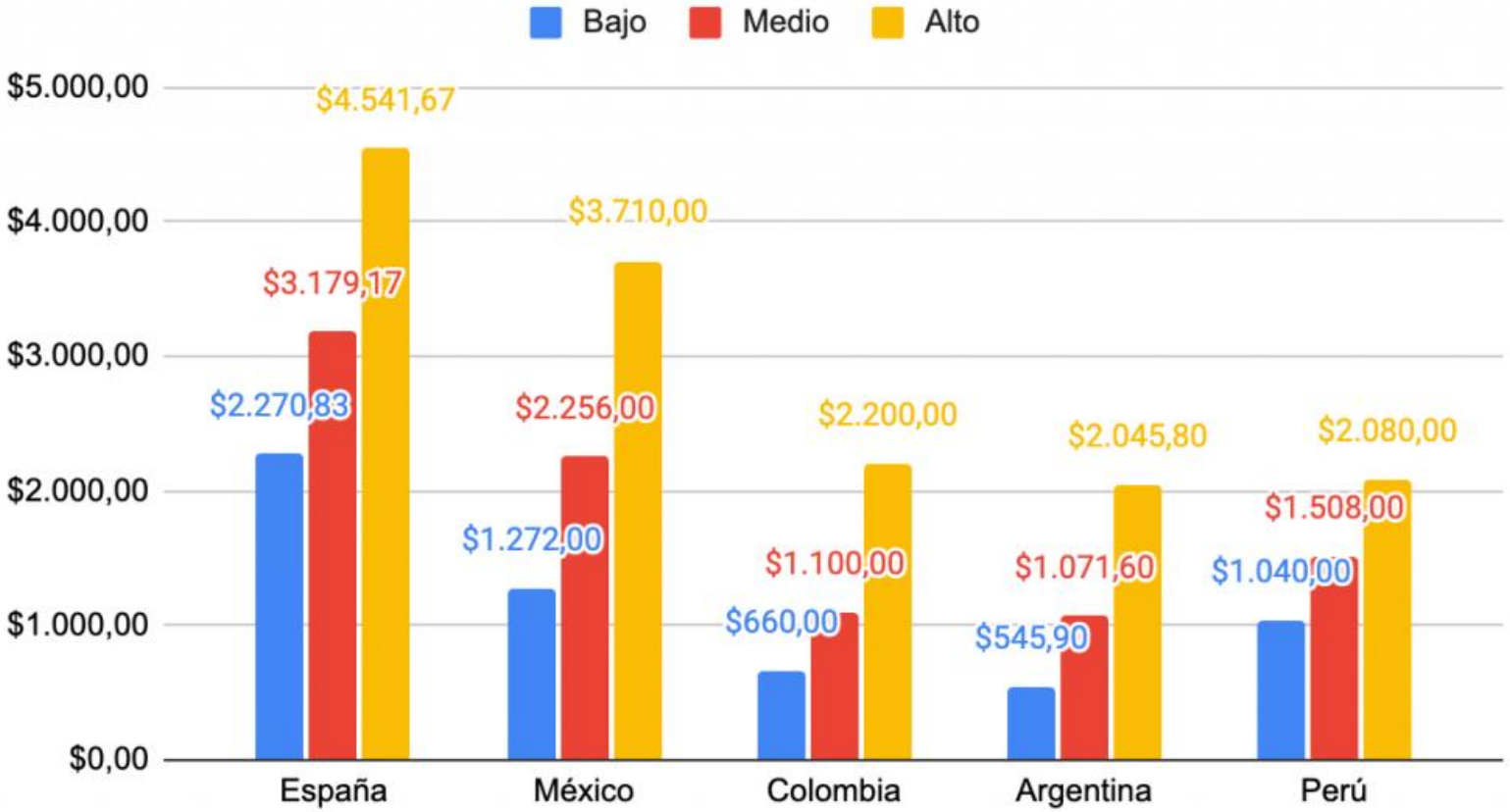
Estadística


25%
51%
38%
65%
Rol
Ent
Pos1
Page ↑
Page ↓

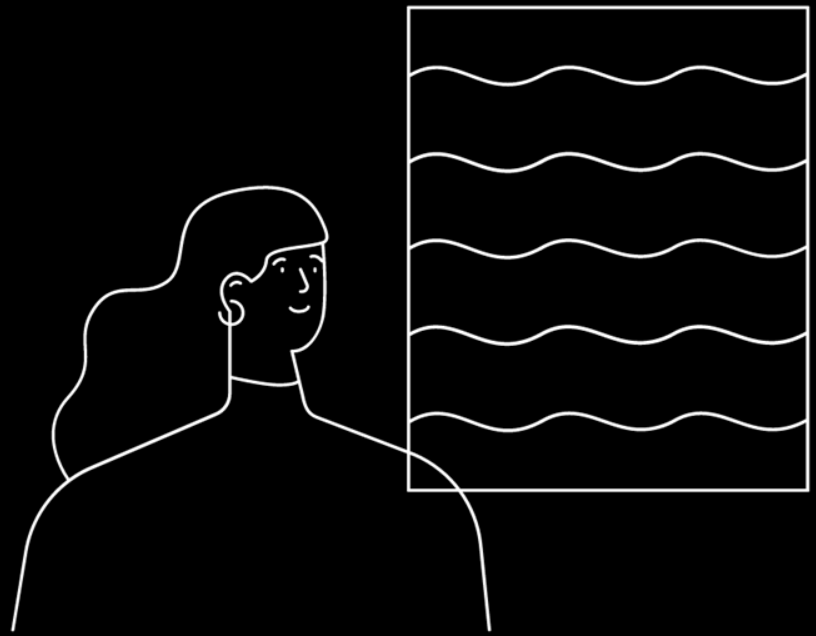
La **estadística** es el arte y la ciencia de reunir datos, analizarlos, presentarlos e interpretarlos.

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

Sueldos Data Engineer convertido a dólares



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.

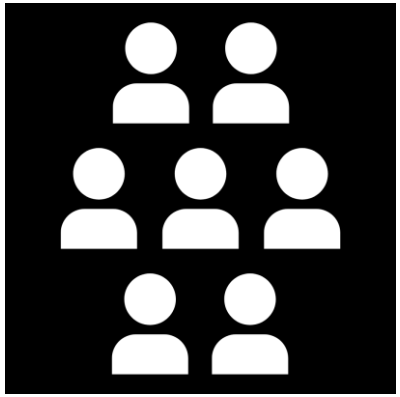


Inferencia estadística

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 y muestra

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



Muestra
Cuando se examina una pequeña parte del grupo.

Población

Se puede hablar de la población de viviendas de un barrio; de la población de comprobantes contables de una empresa; de la población de alumnos en Henry, etc.

Distribución de frecuencias

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.

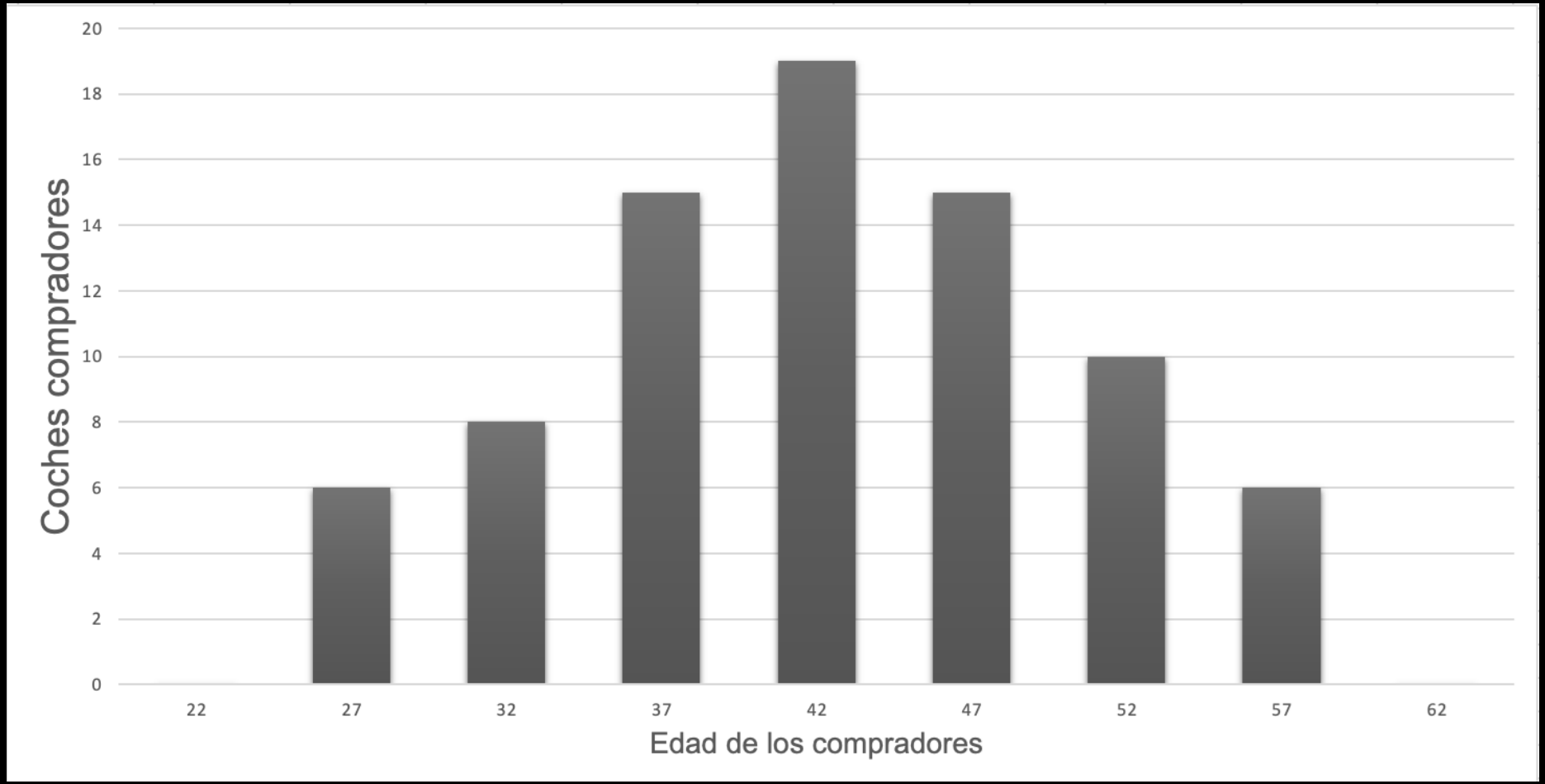
Edades de los compradores de automóviles

Edades	N° de autos f	Verdadero Límite $\underline{VL}$	Punto Medio X_i	Frecuencia Acumulada menor que $F_i^{(-)}$	Frecuencia Acumulada mayor que $F_i^{(+)}$	Frecuencia relativa h	Frecuencia Relativa Acumulada Menor que $H_i^{(-)}$
25-29	6	24,5	27	6	80	7,50%	100,00%
30-34	9	29,5	32	15	74	11,25%	92,50%
35-39	15	43,5	37	30	65	18,75%	81,25%
40-44	18	39,5	42	48	50	22,50%	62,50%
45-49	15	44,5	47	63	32	18,75%	40,00%
50-54	10	49,5	52	73	17	12,50%	21,25%%
55-59	7	54,5	57	80	7	8,75%	100,00%
	80					100,00%	

Histograma

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.


Coches compradores
20
18
16
14
12
10
8
6
4
2
0
22
27
32
37
42
47
52
57
62
Edad de los compradores

Tendencia

Tendencia central

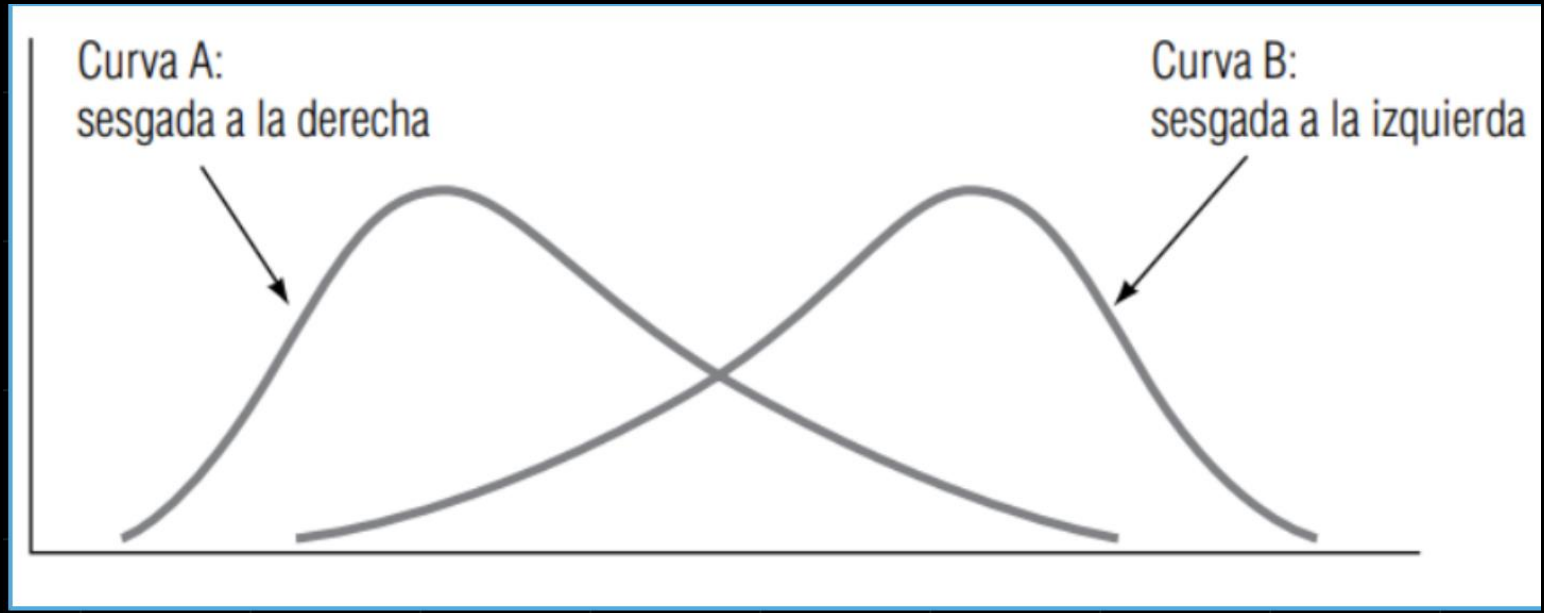
Se refiere al **punto medio** de una distribución.

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.

Tendencia central

Cuando se da el caso de que cada parte es una imagen de espejo de la otra, esta curva se denomina simétrica. Si la curva está sesgada hacia la derecha, se considera positivamente sesgada y si el sesgo se pronuncia hacia la izquierda, se denomina negativamente sesgada.

Sesgos

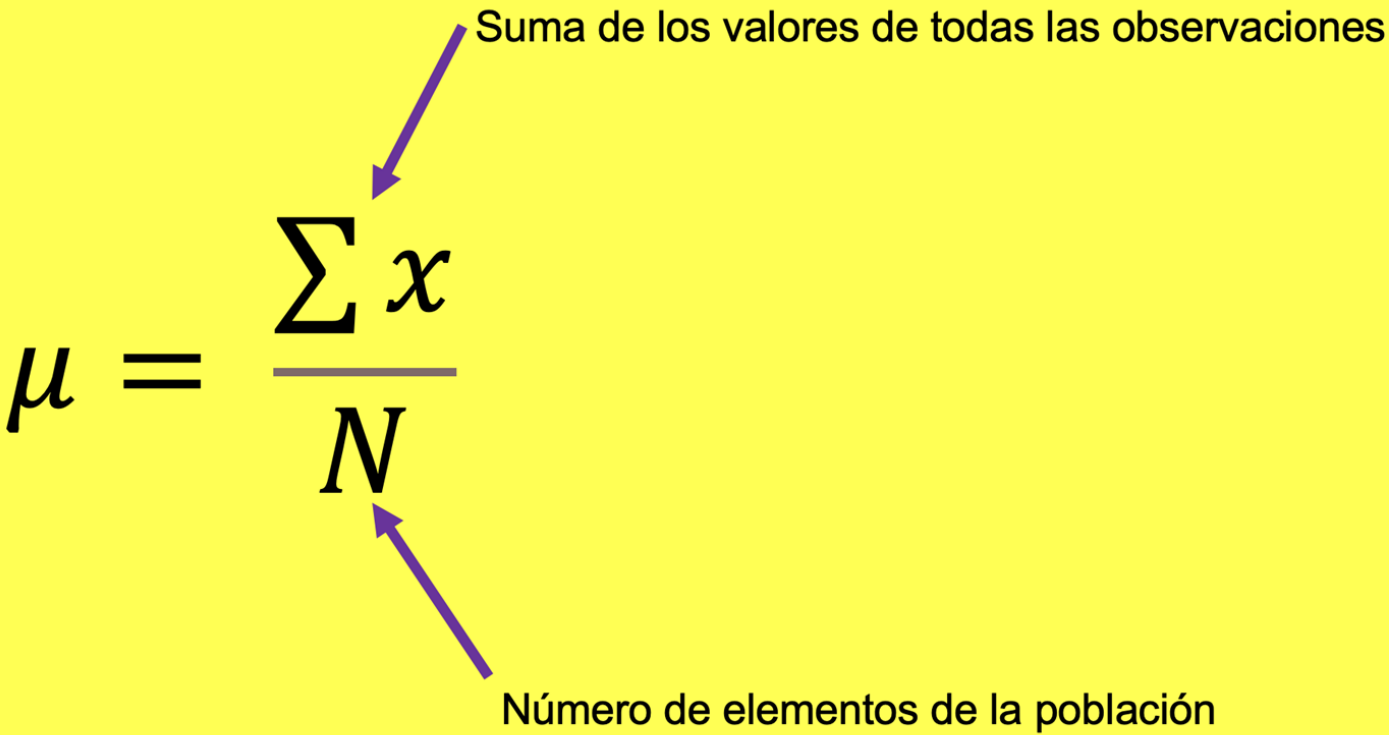


Media, Mediana, Moda

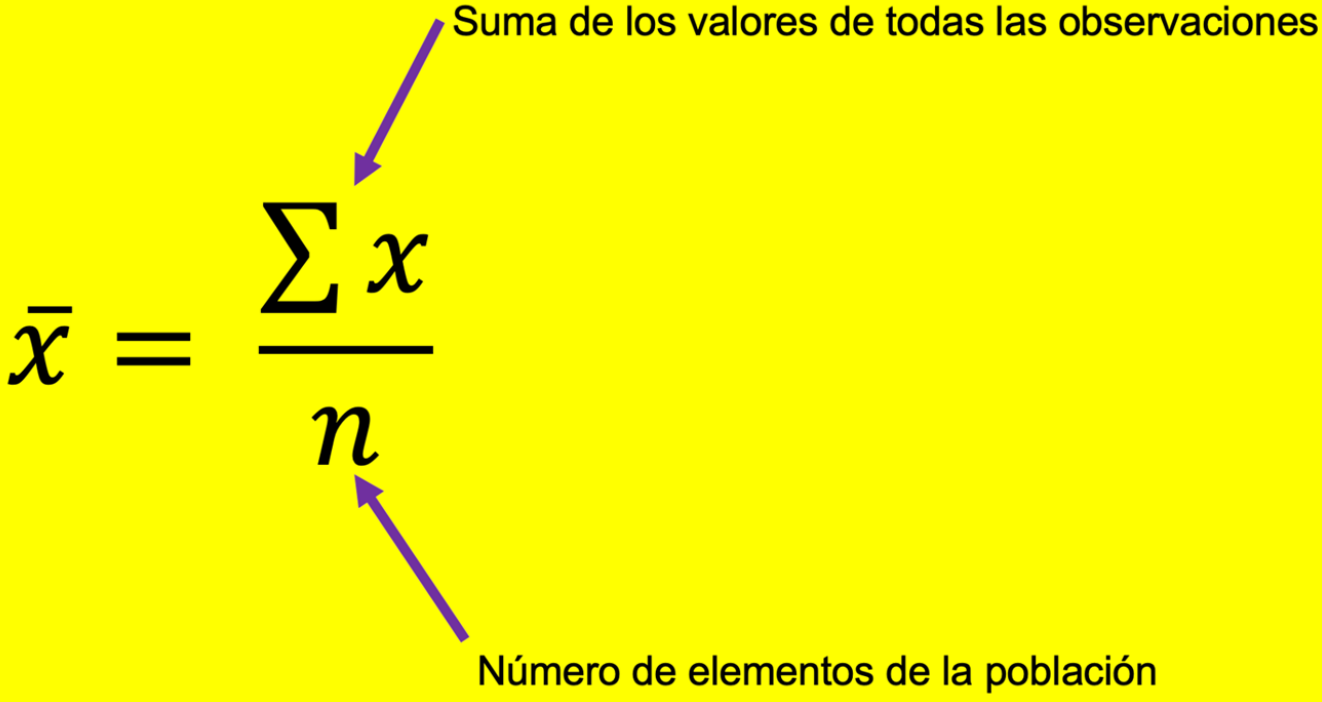
Media aritmética (Promedio)

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

Media aritmética Población



Media aritmética de la muestra



Ventajas

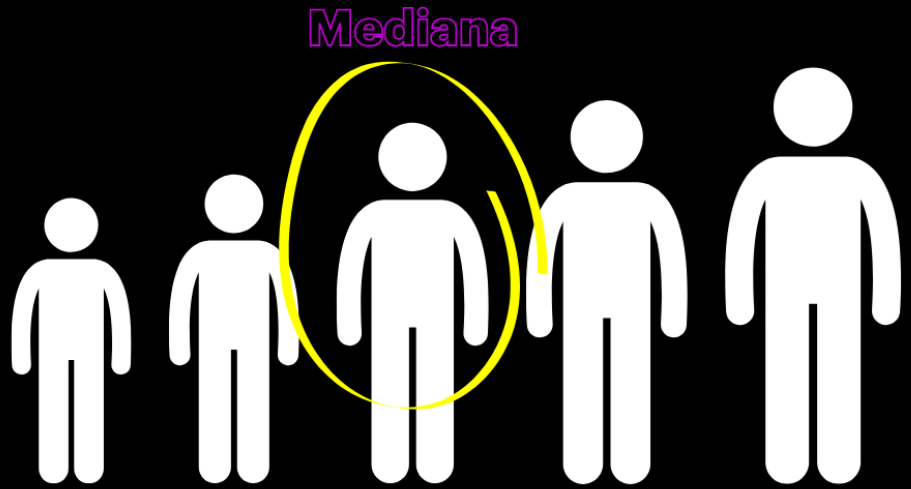
- ✓ Un solo número que representa a un conjunto de datos completo.
- ✓ Concepto familiar.
- ✓ Es única.
- ✓ Es útil para la comparación de medias de varios conjuntos de datos.

Desventajas

- ✗ Puede verse afectada por valores extremos.
- ✗ Resulta un cálculo tedioso.
- ✗ Cuando existen valores de clase extremos abiertos ("60 años o más", "18 años o menos", etc.) no se puede calcular.

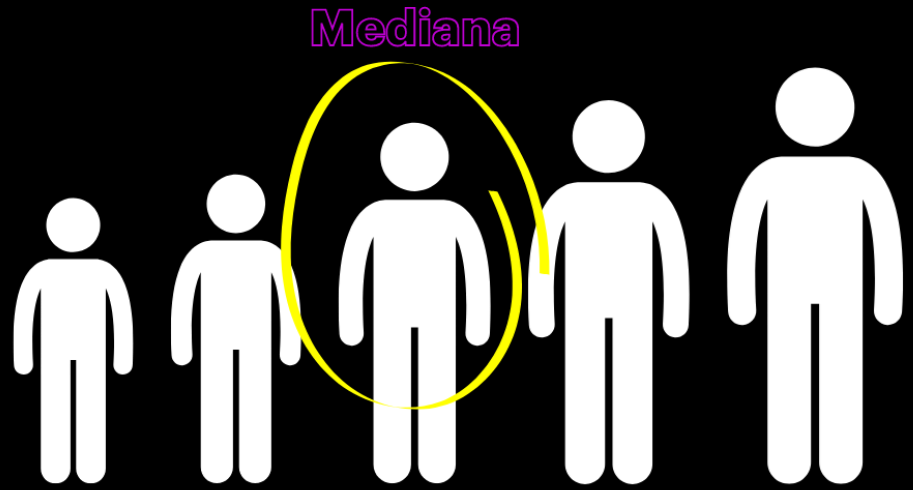
La mediana

- ✓ 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.



A tomar en cuenta

Si el conjunto de datos contiene un número impar de elementos, el de en medio en el arreglo es la mediana; si hay un número par de observaciones, la mediana es el promedio de los dos elementos de en medio.



Ventajas

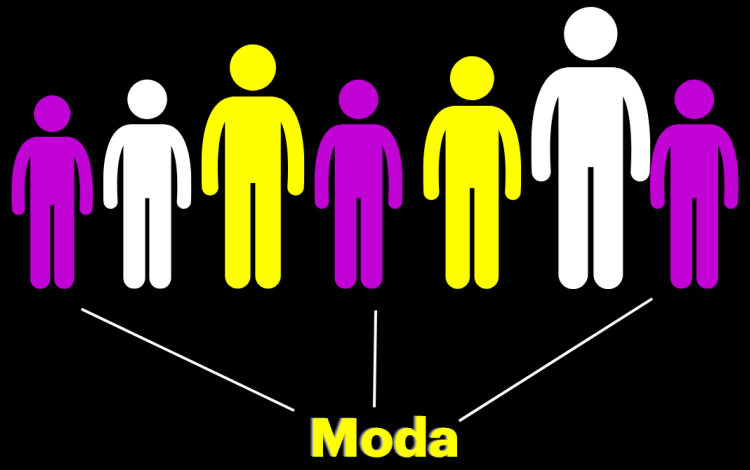
- ✓ No se ve afectada por valores extremos.
- ✓ Es fácil de entender y se calcula a partir de cualquier tipo de datos.
- ✓ La podemos encontrar incluso cuando nuestros datos son descripciones cualitativas, en lugar de números.

Desventajas

- ✗ Ciertos procedimientos estadísticos que utilizan la mediana son más complejos que aquellos que utilizan la media.
- ✗ Debemos ordenar los datos antes de llevar a cabo cualquier cálculo.

Moda

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



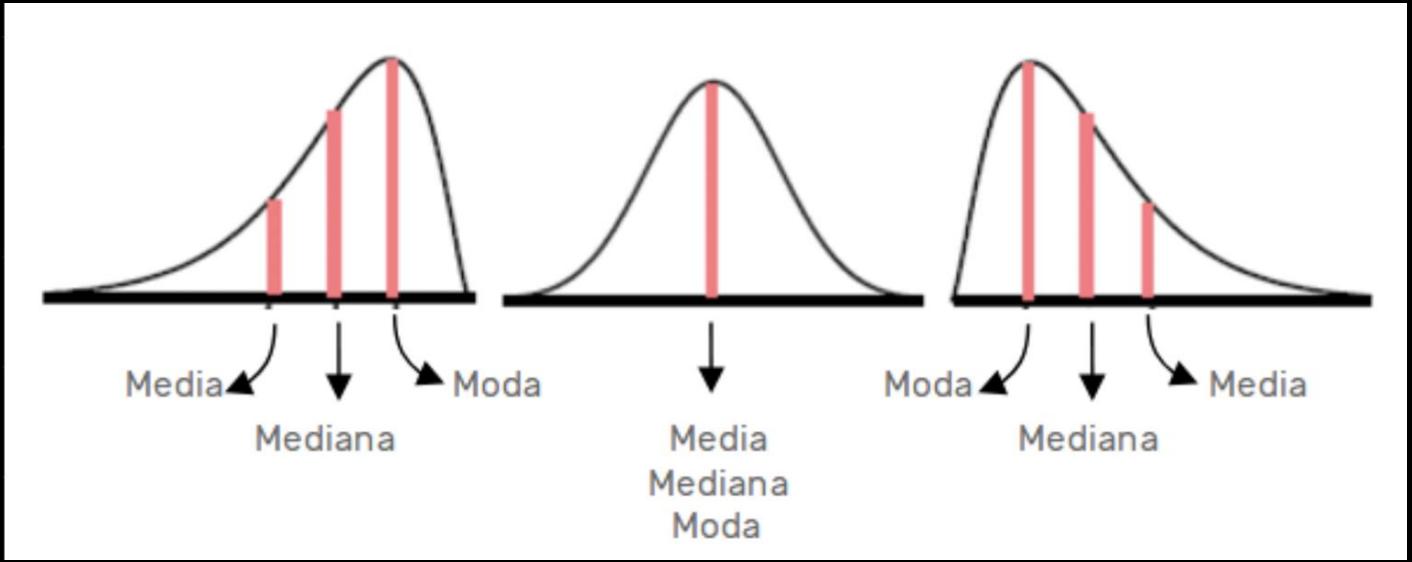
Ventajas

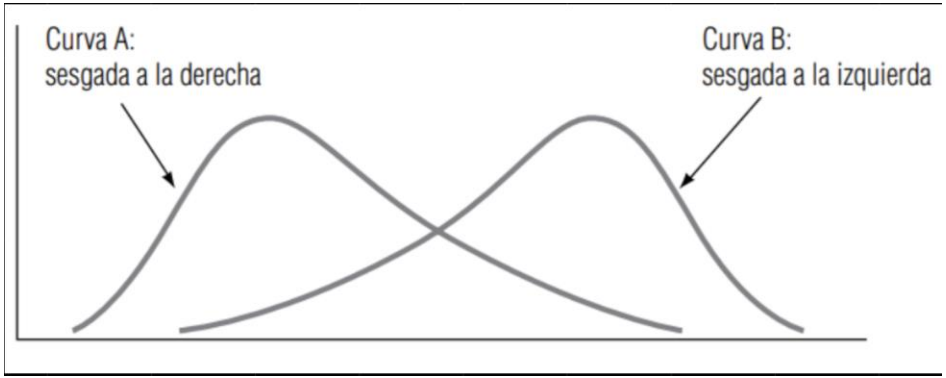
- ✓ Se puede utilizar como una posición central para datos tanto cualitativos como cuantitativos.
- ✓ La mediana, los valores extremos no afectan indebidamente a la moda.
- ✓ La podemos utilizar aun cuando una o más clases sean de extremo abierto.

Desventajas

- ✗ No se utiliza tan a menudo como medida de tendencia central.
- ✗ A veces, no existe un valor modal
- ✗ Conjuntos de datos contienen dos, tres o más modas, es difícil interpretarlos y compararlos.

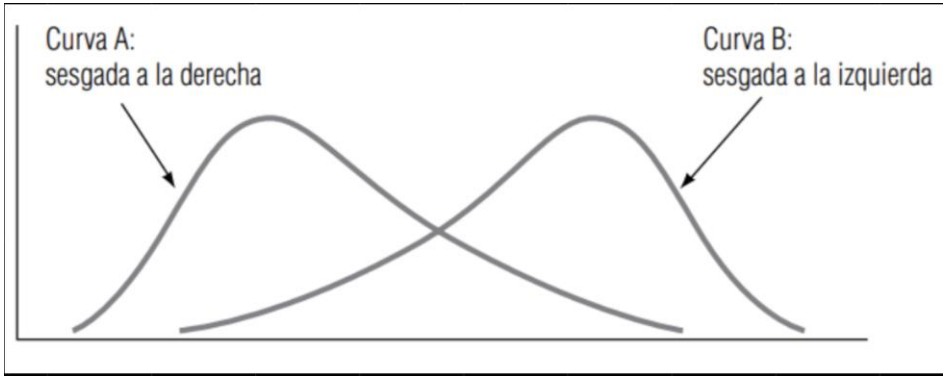
Media, Mediana, Moda





Sesgo Negativo

La moda se encuentra en el punto más alto de la distribución, la mediana está a la izquierda y la media se encuentra todavía más a la izquierda de la moda y la mediana.



Sesgo Positivo

La moda se encuentra en el punto más alto de la distribución, la mediana está a la derecha de la moda y la media se encuentra todavía más a la derecha de la moda y la mediana.

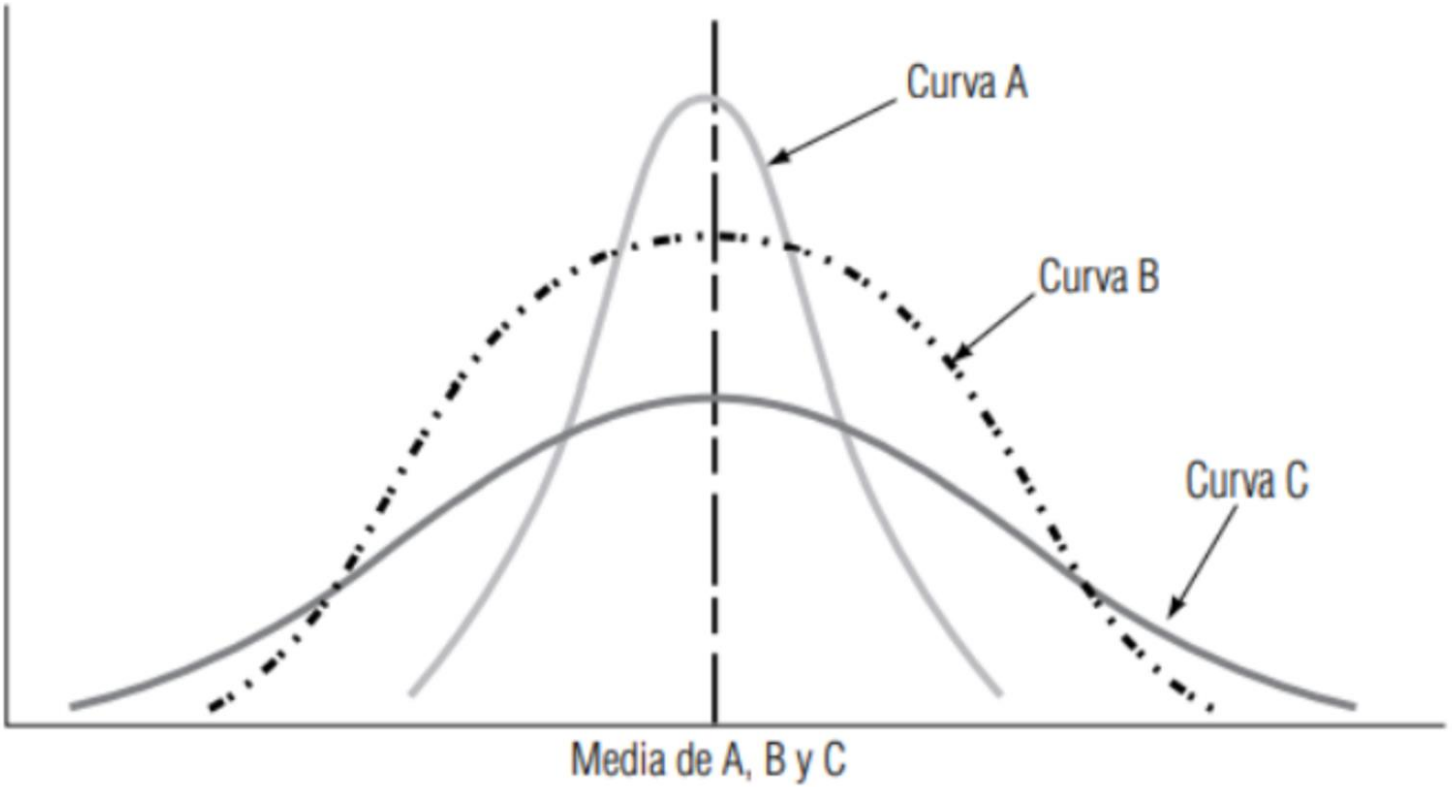
Dispersión, rango, varianza

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

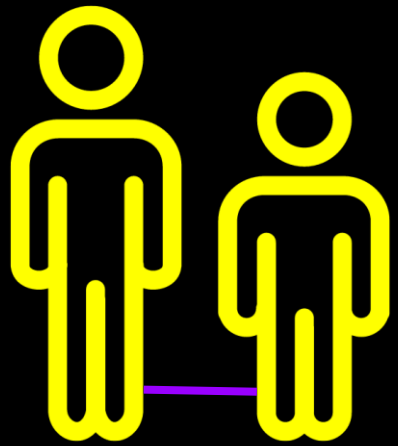
Separación

Variabilidad


Curva A
Curva B
Curva C
Media de A, B y C

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.

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

HENRY

La desviación estándar

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

Coeficientes de variación

Desviación estándar de la población

$$\text{Coeficiente de variación de la población} = \frac{\sigma}{\mu}(100)$$

Media de la población


HENRY