METODOLOGÍA DE LA INVESTIGACIÓN

MSc. IQ. Sebastián BECERRA ROJAS

ANALISIS DE DATOS: INTRODUCCIÓN, HERRAMIENTAS Y SOFTWARE

ANÁLISIS DE DATOS

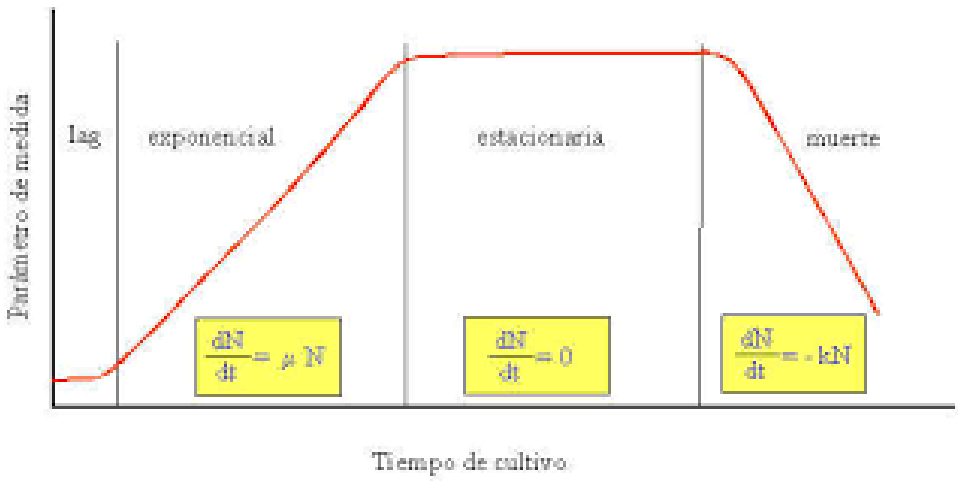
Tras la obtención de los datos necesarios en la investigación, llega el momento de analizarlos, en donde se le da un sentido, orden, descripción e interpretación de dicha información.

Peso	Altura	Edad	Sexo	Nombres
77	1.63	23	1	Pepe
58	1.63	23	2	Ana
89	1.85	26	1	Manolo
55	1.62	23	1	Rafa
47	1.60	26	2	María
60	1.63	26	2	Auxi
54	1.70	22	1	Germán
58	1.65	23	2	Celia
75	1.78	26	2	Carmen
65	1.70	24	1	Juan
82	1.77	28	1	Dani
85	1.83	42	1	Antonio
75	1.74	25	2	Belinda
65	1.65	26	2	Sara

RESPONDER A LA PREGUNTA Y OBJETIVOS DE LA INVESTIGACIÓN

ANÁLISIS DE DATOS

La principal utilidad de analizar correctamente los datos que se obtienen en los procesos de investigación, es encontrar patrones, relaciones, y como las variables de investigación fluctúan con el tiempo.



$$Ln(X) = Ln(X_o) + \mu t$$

Donde:
X = Concentración de sustrato.
μ = Tasa crecimiento M.O.
t = Tiempo.

ANÁLISIS DE DATOS

Tomemos el rol de un investigador el cual quiere emplear un cierto M.O. para la producción de un fármaco revolucionario, por lo que necesita determinar la tasa de crecimiento del mismo para poder escalar el proceso. Al medir experimentalmente las concentraciones del sustrato obtiene:

Tiempo [h]	X [mg/L]
0	7.75
1.04	11.25
1.72	14.30
2.36	17.65
2.86	20.55
3.48	24.10
4.12	26.50
4.74	27.20
4.96	27.25

ANÁLISIS DE DATOS

Gracias a este ejemplo nos podemos dar cuenta que las herramientas para poder “hacer hablar” nuestros datos de investigación, es TODA la teoría que hemos estudiado, y que estudiaremos, a lo largo de nuestro proceso académico.

ANÁLISIS DE DATOS

Adicionalmente, para poder describir diferentes relaciones o patrones dentro de los datos que se obtienen en la investigación científica, debemos hacer uso de otro tipo de herramientas, en este caso matemáticas, la ***estadística.***

ESTADÍSTICA DESCRIPTIVA	ESTADÍSTICA INFERENCIAL
Describir los datos obtenidos, para poder organizarlos y clasificarlos.	Observa los datos y extrae conclusiones por medio de inferencias.

ANÁLISIS DE DATOS

Estadística descriptiva:
Trata a los datos mediante el recuento, ordenación y clasificación de los datos tomados en la aplicación de la prueba piloto o fase experimental.

Construcción de tablas

Parámetros estadísticos

Representación gráfica

ANÁLISIS DE DATOS

Estadística descriptiva:

No olvidar que para el tratamiento de dichos datos, debemos tener en cuenta la gran gama de tipos de datos que podemos obtener en cada una de las investigaciones.

- Cuantitativas
  - Discretas
  - Continuas
- Cualitativas

ANÁLISIS DE DATOS

Estadística descriptiva:
Gracias a la propia definición de la estadística descriptiva, no se van a emplear herramientas de la probabilidad, sino medidas que describan el comportamiento del los datos.

- Centralización
- Dispersión
- Forma
- Relación entre variables

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de centralización):
Siendo estas la piedra angular del análisis estadístico, gracias a que, al ser empleadas en un conjunto de datos, señalan cual es el centro estadístico del mismo

- Promedio
- Mediana
- Cuantiles
- Moda

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de centralización):
Para poder visualizar la utilidad de las medidas de centralización, contestemos, **de forma voluntaria**, las siguientes preguntas:

- Edad en años.
- Altura en metros.
- Asignatura preferida.
- Color del iris ocular.
- Artículos científicos leídos en el semestre.

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de centralización):

Promedio:

Descrito también como media aritmética, es el valor que representan todos los datos si se repartieran de forma equitativa. Matemáticamente se define como:

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

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de centralización):
Moda:
Se define como el valor dentro de los datos que aparece con una mayor frecuencia. Es importante tener en cuenta que la moda es funcional para datos discretos, en el caso de datos continuos, *es necesario discretizarlos*

Edad [años]	Altura [m]	Asignatura pref	Color Iris	Art. Cie. Leídos 2025-2

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de centralización):
Mediana:
Es el dato que divide de forma equitativa el conjunto de datos, al estar organizados, generando dos subconjuntos en donde, en cada uno, se aloja el 50% de los datos.

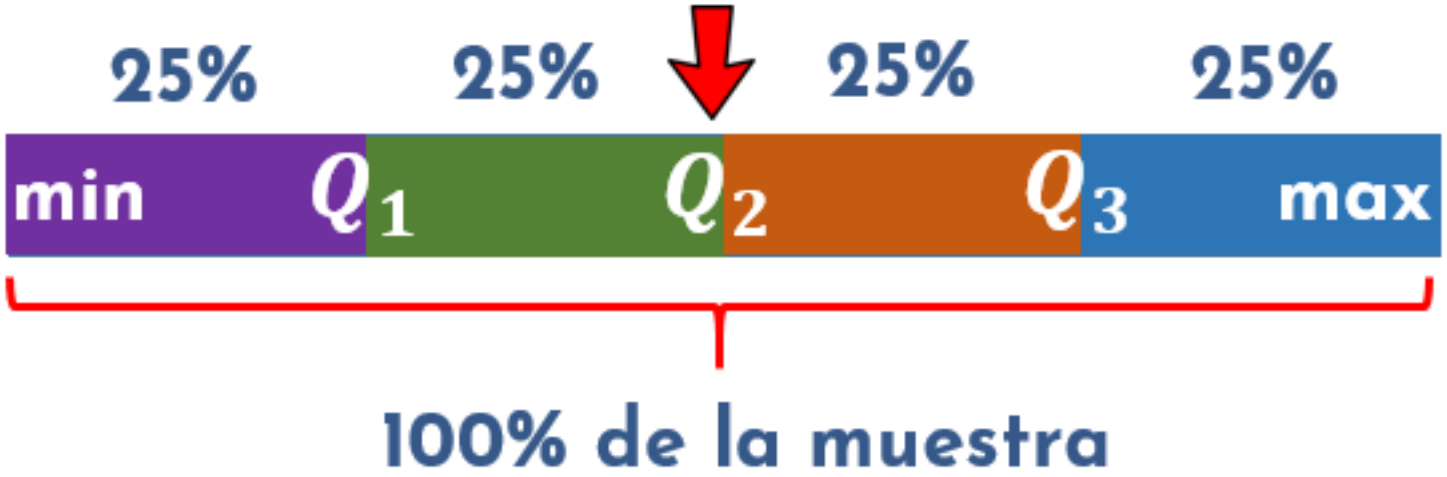
Cantidad de datos impar	Cantidad de datos par
Dato que queda en la mitad de la lista ordenada.	Promedio entre los dos valores que quedan en la mitad de la lista ordenada.

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de centralización):

Cuantiles:

Son valores que dividen los datos organizados en partes iguales, por lo que es una medida de organización de los datos, generando rangos en donde se dividen los diferentes grupos de datos.

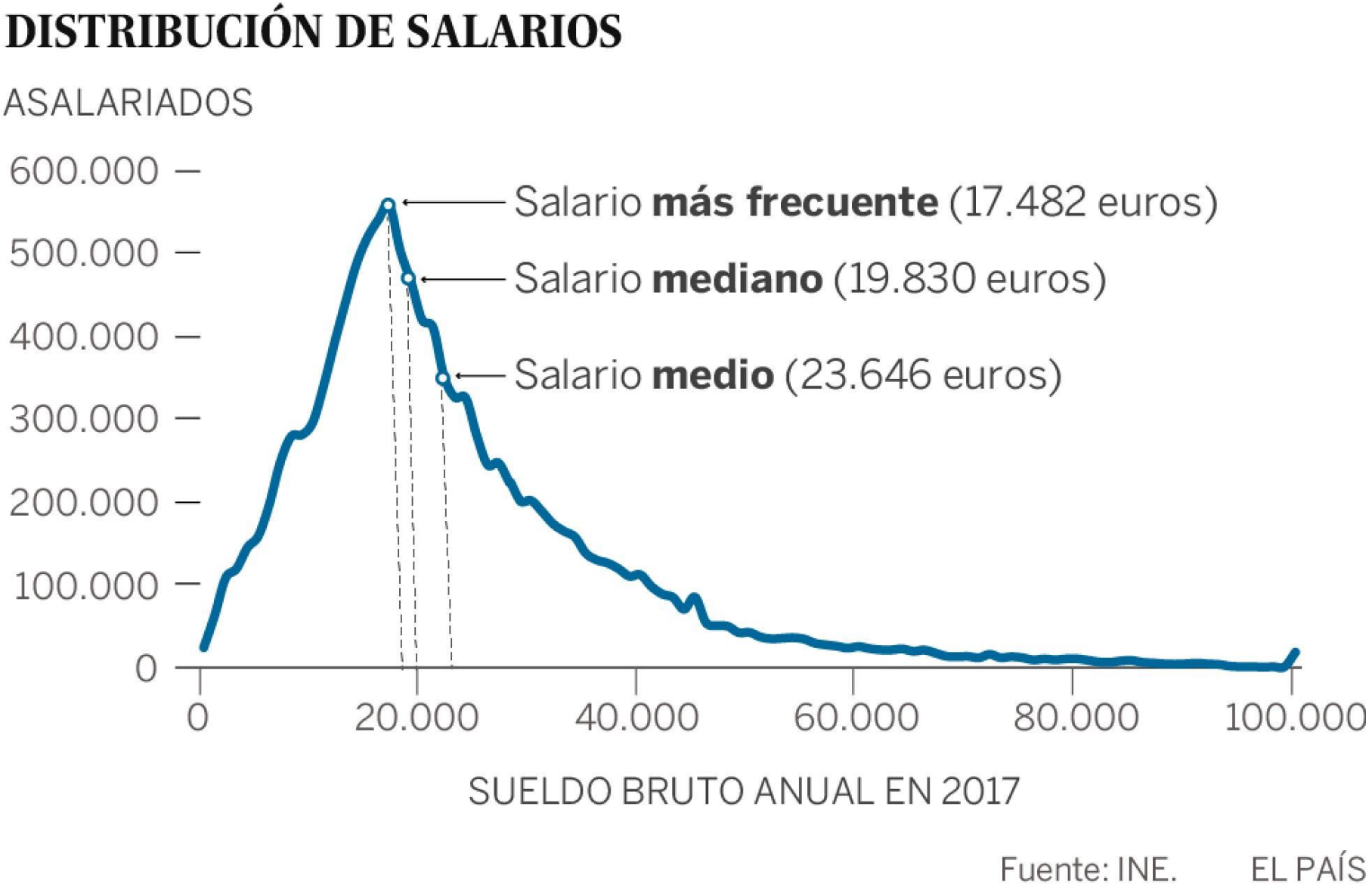


ANÁLISIS DE DATOS

Estadística descriptiva (medidas de centralización):
Cuantiles:
Para determinar los valores que determinan los rangos de los cuartiles, seguiremos el siguiente modelo matemático.

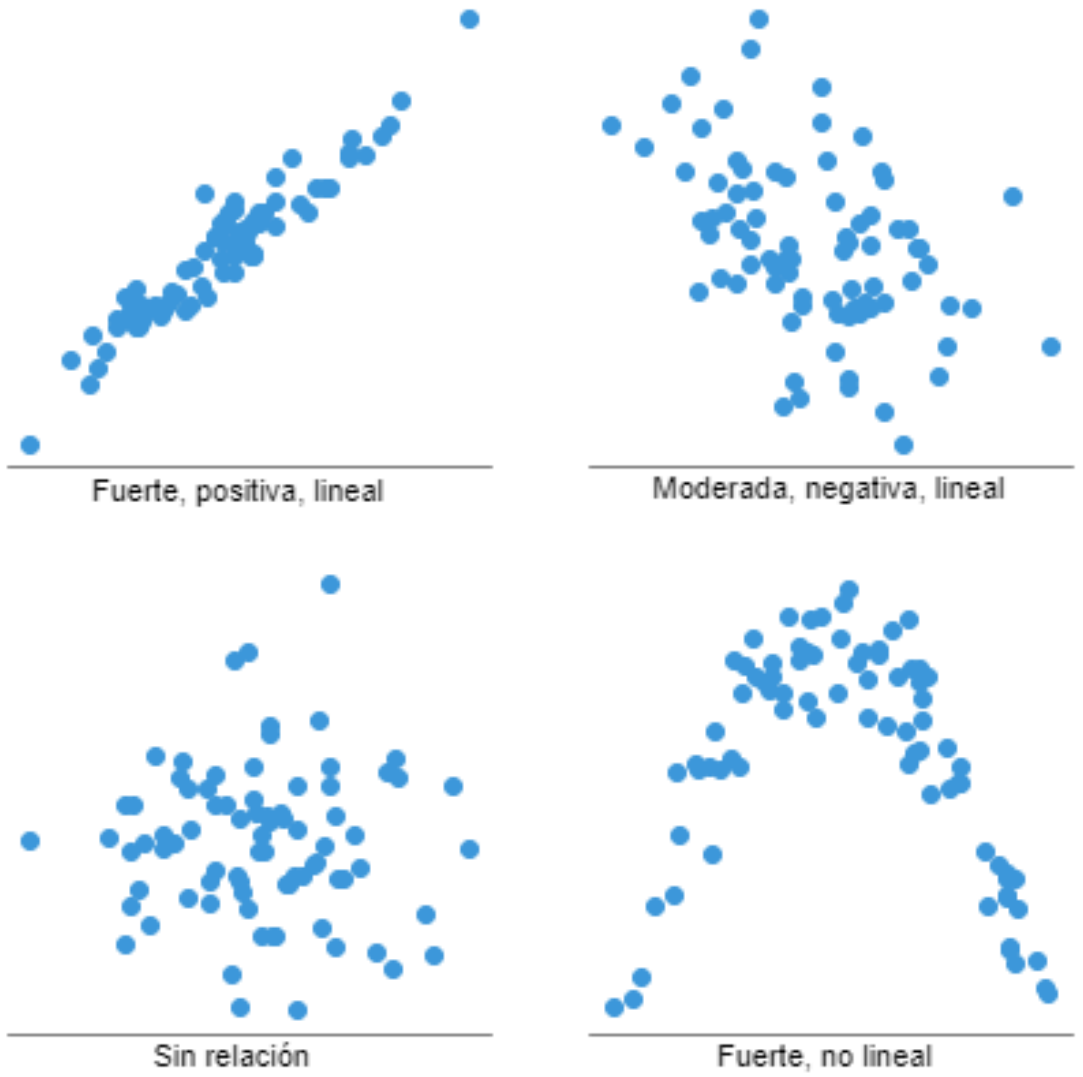
$$Q_k = \left(\frac{k}{4}\right) n = \left(\frac{\%k}{100}\right) n$$

ANÁLISIS DE DATOS



ANÁLISIS DE DATOS

Estadística descriptiva (medidas de dispersión): Llamadas también medidas de variabilidad, definen la variabilidad del grupo de datos, indicando de forma numérica si los datos está alejados de la media o no.



ANÁLISIS DE DATOS

Estadística descriptiva (medidas de dispersión):
Al igual que ocurre con las medidas de centralización, las medidas de dispersión también se basan en ecuaciones matemáticas para determinar los diferentes valores, y describir los datos.

- Rango
- Desviación estandar
- Coeficiente de variación
- Varianza

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de dispersión):
Rango:
Como primer dato informativo se definen los valores extremos de las variables estudiadas, de forma tal que podamos saber dentro de valores va a trabajar la distribución estadística.

$$Rango = X_{máx} - X_{min}$$

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de dispersión):

Varianza:

Determina que tan “alejados” están los datos de su promedio aritmético, por lo que, dependiendo del valor obtenido, se puede afirmar:

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

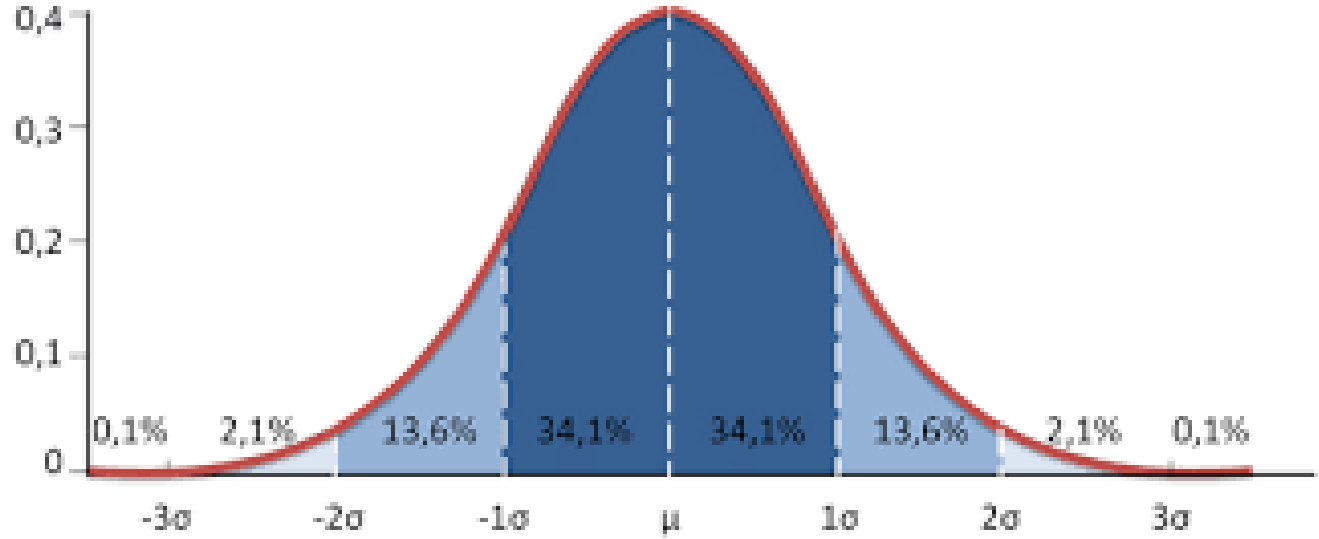
Valores cercanos a 0

Valores alejados del 0

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de dispersión):
Desviación estándar:
Siendo la raíz cuadrada de la varianza, mide la separación de los datos con respecto a la media, por lo que podemos agrupar los datos en diferentes zonas:

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



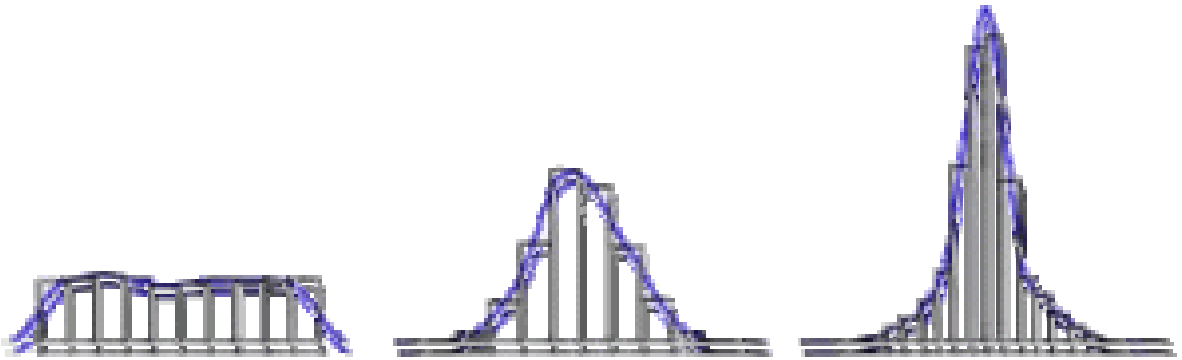
ANÁLISIS DE DATOS

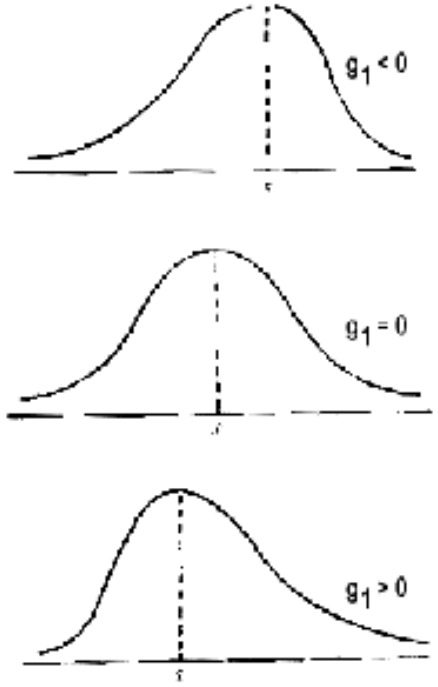
Estadística descriptiva (medidas de dispersión):
Coeficiente de variación:
Permite realizar la comparación entre diferentes grupos de datos, incluso si los datos no están relacionados entre si. Matemáticamente se define como el cociente entre la desviación estándar y la media aritmética:

$$CV = \frac{\sigma}{|\bar{x}|}$$

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de forma):
Muestran la forma de la distribución, al ser graficada, describiendo características particulares como simetría, nivel de concentración de datos, y apuntamiento de los mismos.

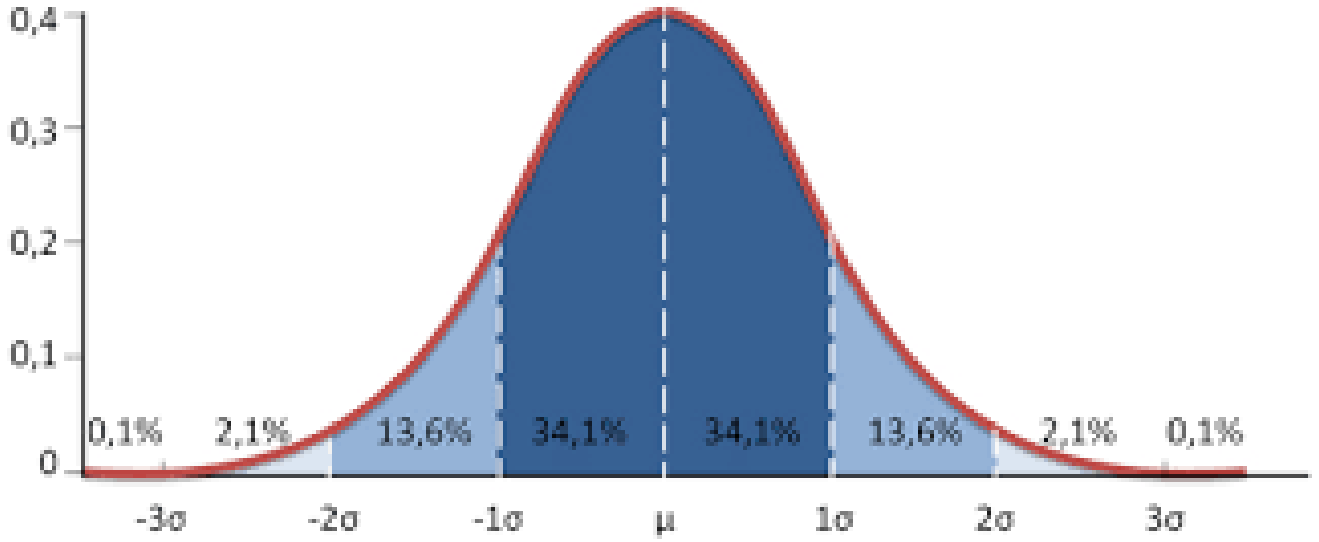




ANÁLISIS DE DATOS

Estadística descriptiva (medidas de forma):

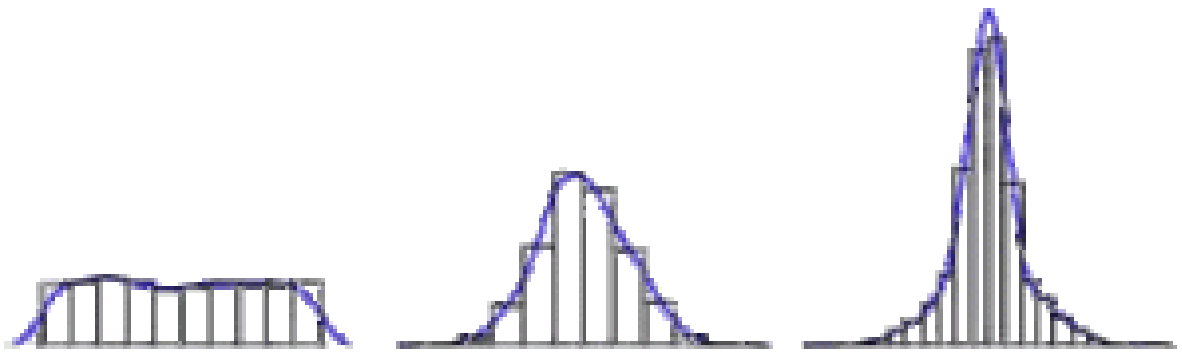
Es importante tener en cuenta, que las diferentes medidas de forma se basan en la distribución normal, en donde los valores tienen una simetría perfecta respecto a la media de los valores, además de representar el primer acercamiento al análisis estadístico más profundo.



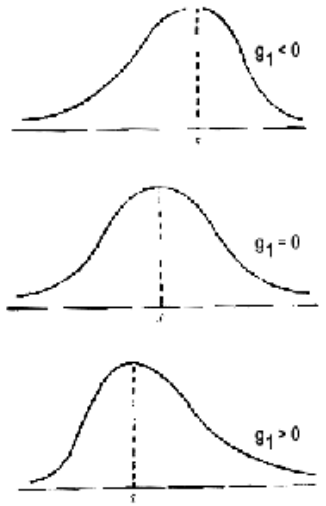
ANÁLISIS DE DATOS

Estadística descriptiva (medidas de forma):
Dentro de las medidas de forma que se pueden estudiar, se tienen dos en particular:

Curtosis



Asimetría



ANÁLISIS DE DATOS

Estadística descriptiva (medidas de forma):
Curtosis:
Medida de la frecuencia con la que se forman valores atípicos dentro de grupo de datos, generando tres posibles casos:

Mesocúrtico

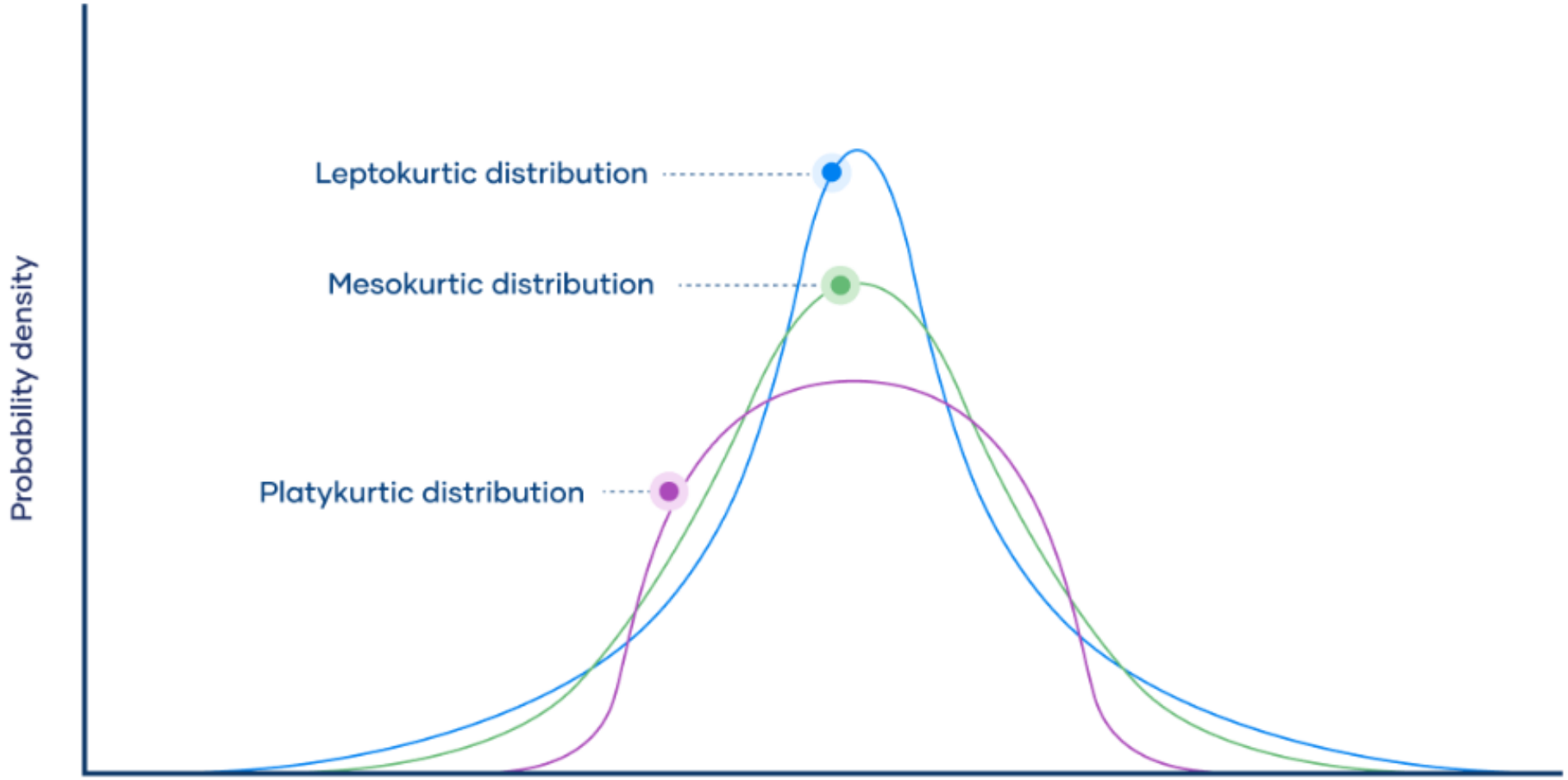
Platicúrtico

Leptocúrtico

	Categoría		
	Mesocúrtico	**Platicúrtico**	**Leptocúrtico**
Cola	De cola media	De cola fina	De cola gorda
Frecuencia de valores atípicos	Medio	Bajo	Alto
Curtosis	Moderado (3)	Bajo (< 3)	Alto (> 3)
Exceso de curtosis	0	Negativo	Positivo
Ejemplo de distribución	Normal	Uniforme	Laplace

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de forma):
Curtosis:



ANÁLISIS DE DATOS

Estadística descriptiva (medidas de forma):

Curtosis:

Matemáticamente, la curtosis se define como:

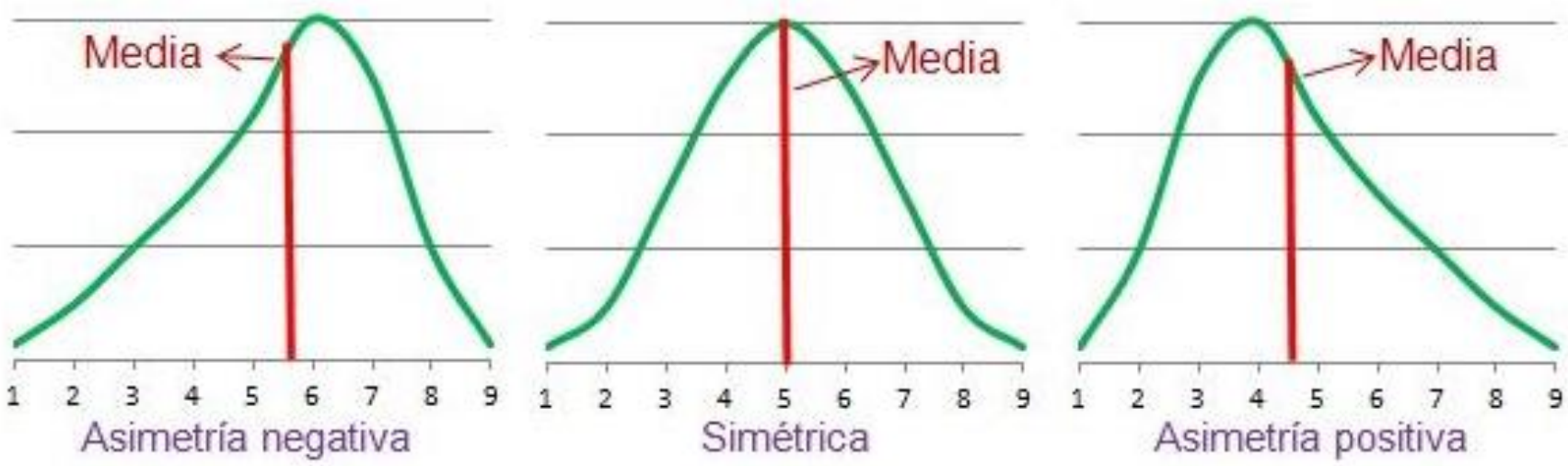
$$Curtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} * \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4 - \frac{3(n-1)^4}{(n-2)(n-3)}$$

No obstante, dentro de softwares especializados para el cálculo directo de la curtosis.

Mesocúrtico	Platicúrtico	Leptocúrtico
Cu = 0	Cu < 0	Cu > 0

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de forma):
Asimetría:
Indica la tendencia de los datos a agruparse, o no, en valores cercanos a la media aritmética, descubriendo así patrones de agrupación de los diferentes valores de las variables estudiadas.



ANÁLISIS DE DATOS

Estadística descriptiva (medidas de forma):

Asimetría:

No obstante existen diferentes coeficientes de asimetría, dentro de la investigación cuantitativa, el más ampliamente usado es el de Fischer:

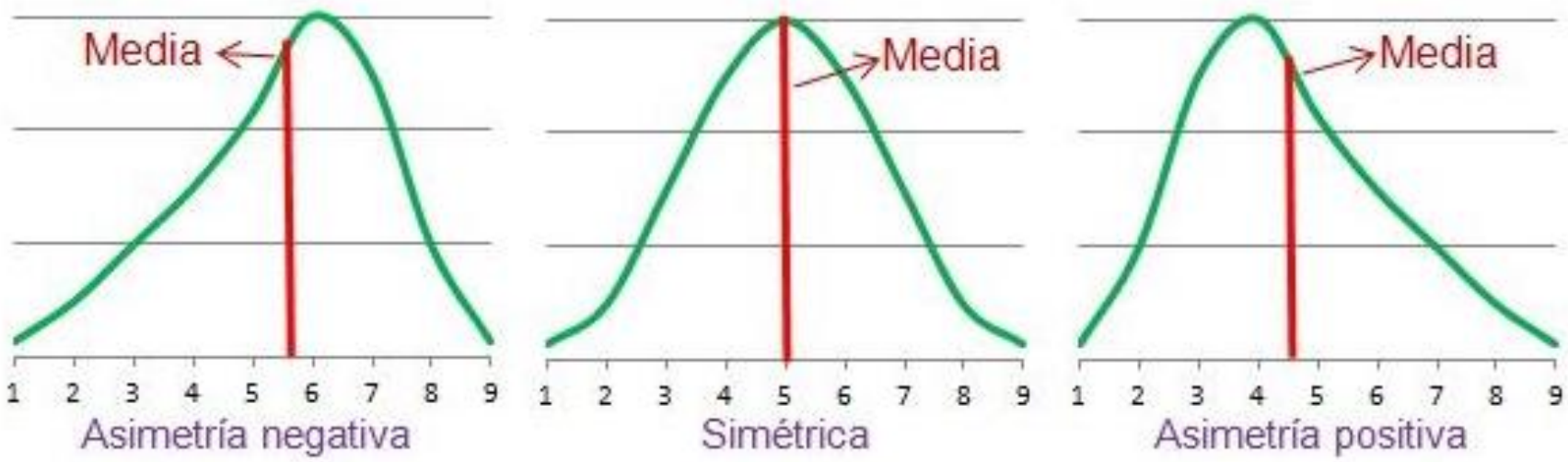
$$CA = \frac{n}{(n-1)(n-2)} * \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma}\right)^3$$

ANÁLISIS DE DATOS

Estadística descriptiva (medidas de forma):

Asimetría:

Dependiendo el valor que se obtenga en el coeficiente de asimetría de Fischer, se pueden obtener tres posibles opciones:



CA < 0	CA = 0	CA > 0

ANÁLISIS DE DATOS

Estadística descriptiva (Relación entre variables):
Dentro de la investigación cuantitativa uno de los puntos mas valiosos dentro del análisis de variables, es determinar la relación entre ellas, para poder verificar, o no nuestras hipótesis planteadas en el proceso investigativo.

$$Concentración\ Comp.Act\ VS\ Efectividad$$

ANÁLISIS DE DATOS

Estadística descriptiva (Relación entre variables):
Para poder esclarecer la relación entre ellos se emplea, lo que en estadística se conoce como análisis de regresión.

Análisis de regresión

Relación matemática entre dos, o mas variables planteadas dentro de la hipótesis.

Importancia de los factores - ¿Qué factores se pueden ignorar? – Interacciones entre factores

ANÁLISIS DE DATOS

Estadística descriptiva (Relación entre variables):
Dentro del análisis en la estadística aplicada a los procesos investigativos cualitativos, se debe determinar:

Modelo de regresión

Lineal simple
Lineal múltiple
No lineal

Verificación del modelo de regresión

R^2

ANÁLISIS DE DATOS

Estadística descriptiva (Relación entre variables):

R^2:

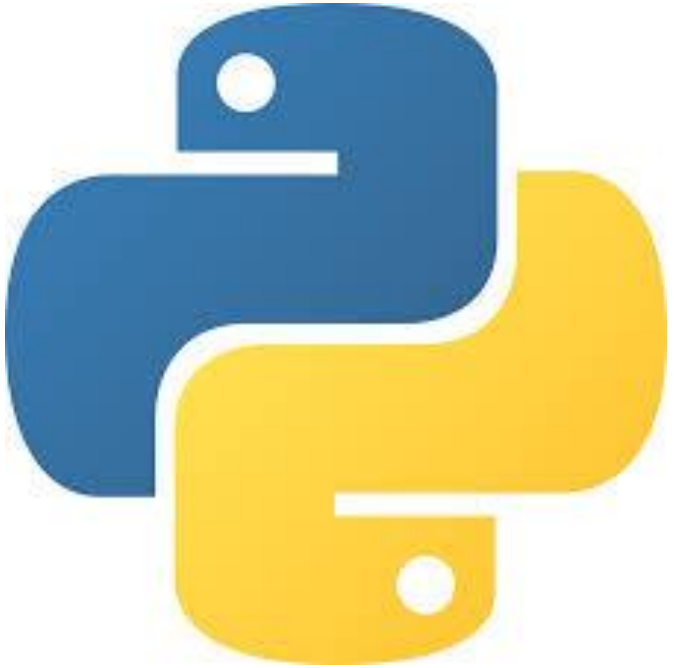
Medida estadística que relaciona que tan próximos están los datos al modelo de regresión ajustada, en donde matemáticamente se relaciona de la siguiente manera:

$$R^2 = \frac{\sum_{t=1}^{T} (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2}$$

Valor estimado del modelo respecto a la media.

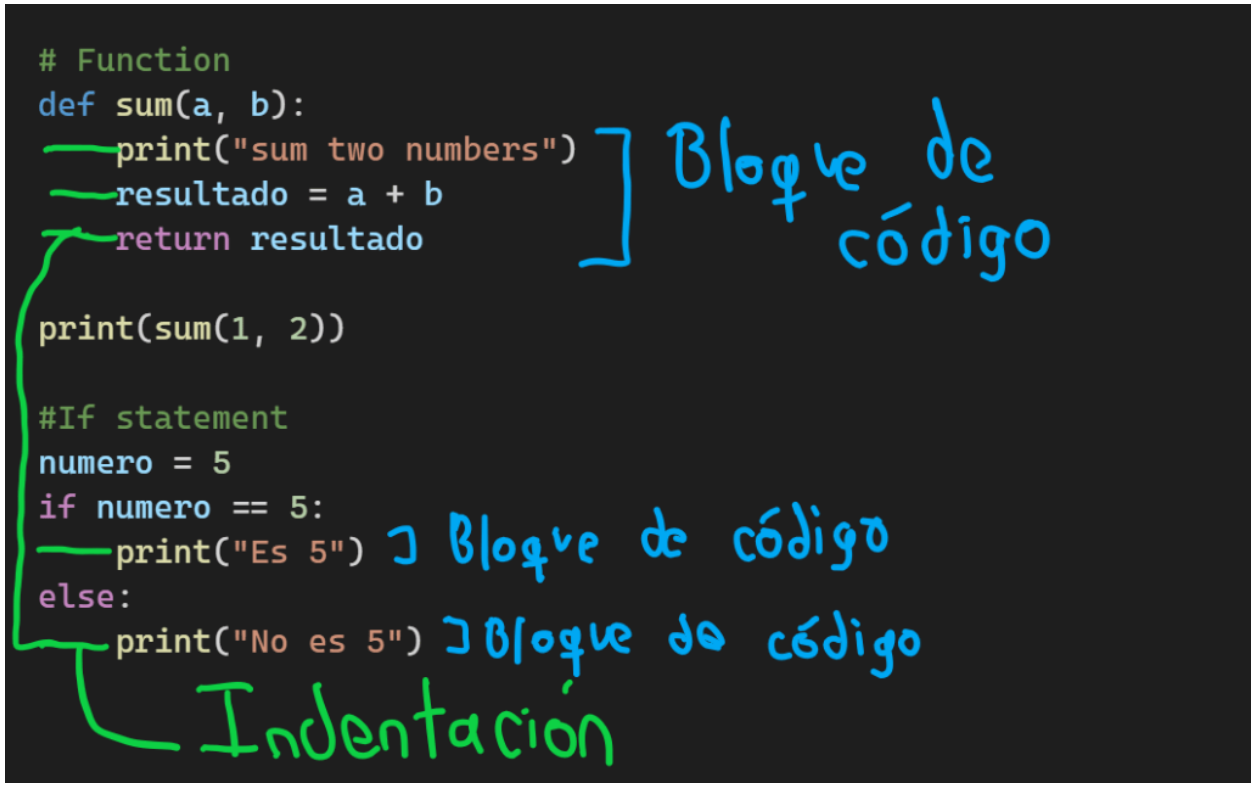
SOFTWARES

Actualmente el proceso de analizar los datos recolectados e



SOFTWARES

Python es un lenguaje de programación ampliamente usado en la actualidad para procesos de Machine Learning, modelamiento de procesos químicos y análisis de datos, caracterizado por su facilidad de aprenderlo, y entender el código generado



SOFTWARES

Para poder aprovechar todo el potencial que tiene Python, se emplean diferentes librerías para poder adicionar funciones a el código base, expandiendo enormemente las capacidades de poder analizar datos.

SOFTWARES

Visualización de datos

Manipulación y analisis de datos

Trabajo con objetos matemáticos como vectores o matrices.

SOFTWARES

Como gran ventaja dentro de los diferentes softwares de análisis de datos que se pueden encontrar en el mercado es su uso completamente gratuito, por lo que existe una gran comunidad en internet que aporta actualizaciones y un amplio soporte para el trabajo colaborativo.

SOFTWARES

Ejercicio:
De la base de datos dada compare tres países cualquiera, y compare los valores de "*Government Effectiveness: Estimate" vs "Total greenhouse gas emissions including LULUCF (Mt CO2e)",* entre ellos.

Exponga, tras una breve investigación en línea, y debate con su compañero, la relación que tienen estas dos variables, y como se relacionan entre los países seleccionados.