Análisis Exploratorio y Programación Estadística

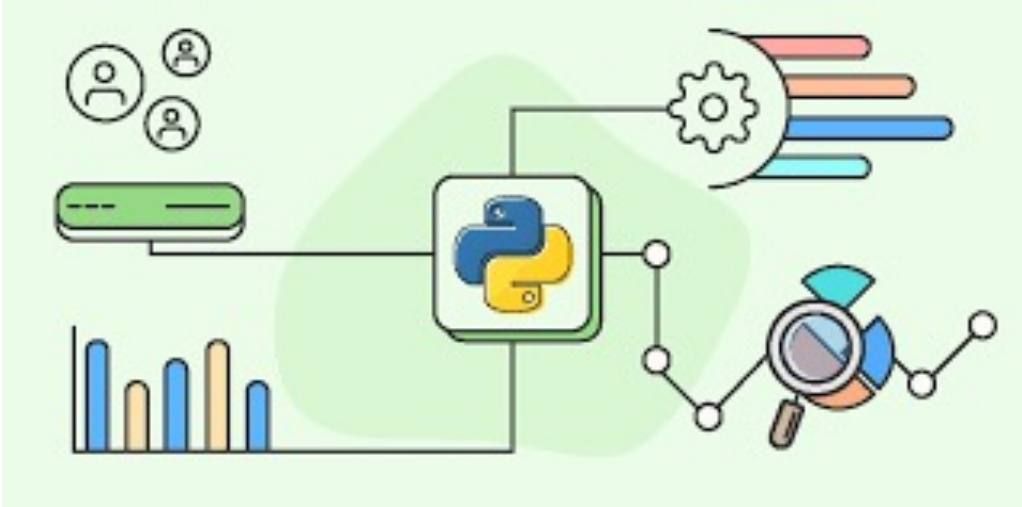
Estadística Descriptiva

Especialización en Ciencia de Datos

2021

Objetivos

- Utiliza los conceptos básicos de estadística descriptiva.
- Caracterizar un conjunto de datos de una población.



Contenido:

1. Estadística: definición
2. Tipos de variables
3. Medidas de tendencia central
4. Representación de datos
5. Rangos de dispersión
6. Librerías Python

1. Estadística

¿Para que sirve la Estadística?

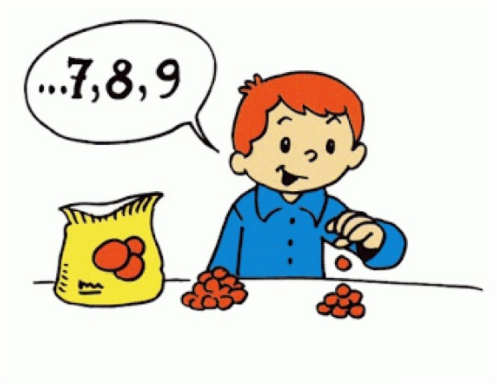
La Ciencia se ocupa en general de fenómenos observables

La Ciencia se desarrolla observando hechos, formulando leyes que los explican y realizando experimentos para validar o rechazar dichas leyes

Los modelos que crea la ciencia son de tipo determinista o **aleatorio (estocástico)**

La **Estadística** se utiliza como **tecnología al servicio** de las ciencias donde la variabilidad y la incertidumbre forman parte de su naturaleza

Estadística



Estadística

La Estadística es la Ciencia de la

Descriptiva

- **Sistematización, recogida, ordenación y presentación** de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico, con objeto de

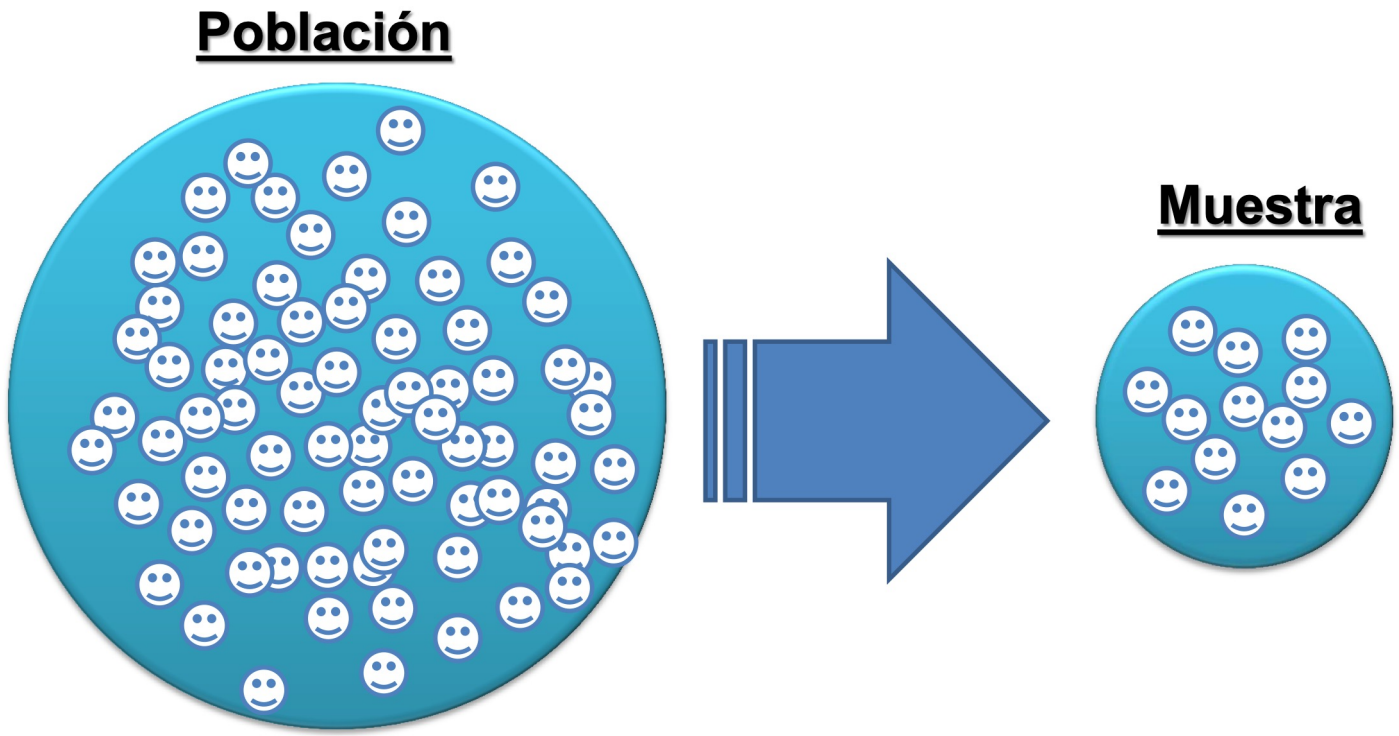
Probabilidad

- **deducir las leyes** que rigen esos fenómenos,

Inferencia

- y poder de esa forma hacer previsiones sobre los mismos, tomar **decisiones** u obtener **conclusiones**.

Población y Muestra



Población (‘*population*’) es el conjunto sobre el que estamos interesados en obtener conclusiones (hacer inferencia).

Normalmente es demasiado grande para poder abarcarlo.

Muestra (‘*sample*’) es un subconjunto suyo al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones)

Debería ser “representativo”

Esta formado por miembros “seleccionados” de la población (individuos, unidades experimentales).

Tipos de estadística

Estadística univariada (estudia una sola variable), bivariada (estudia la relación entre dos variables), y multivariada (estudia tres o más variables).

Estadística

Descriptiva → **Describir la muestra**

Inferencial → **Infiere conclusiones a partir de los datos que describen la muestra**

La diferencia radica en que la estadística descriptiva procede a resumir y organizar esos datos para facilitar su análisis e interpretación, y la estadística inferencial procede a formular estimaciones y probar hipótesis acerca de la población a partir de esos datos resumidos y obtenidos de la muestra.

Estadística descriptiva

Incluye la tabulación, representación y descripción de conjuntos de datos.

A partir de ellos se puede organizar, simplificar y resumir información básica.

Los datos pueden ser de variables cuantitativas o categóricas.

2. Tipos de variables

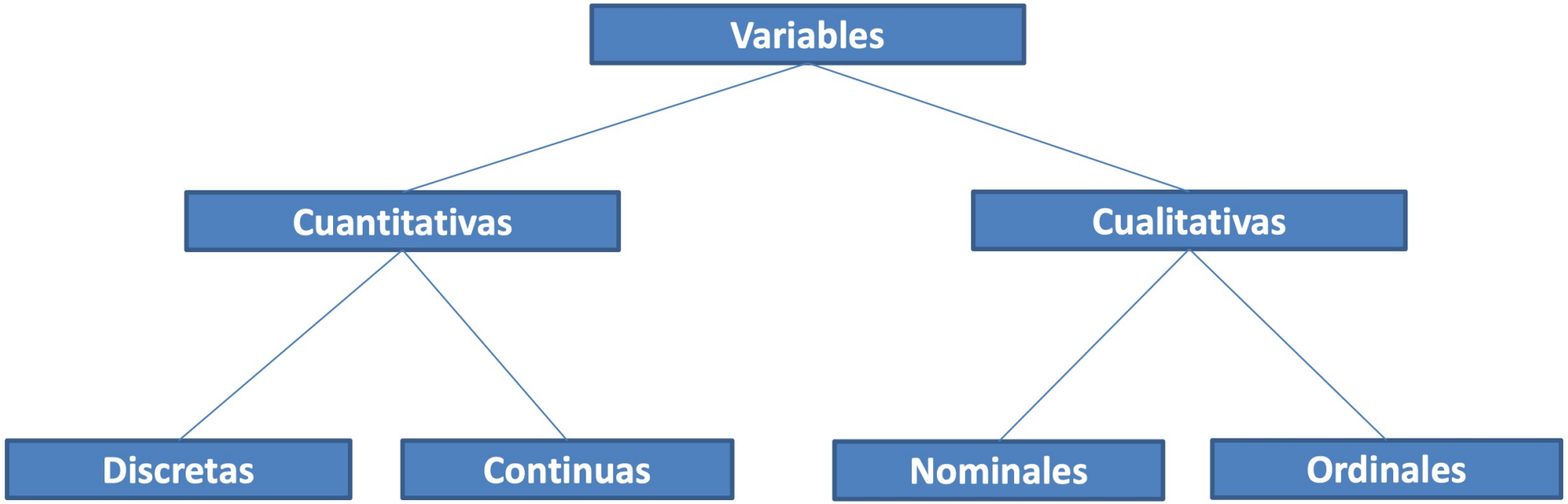
Variables

Una **variable** es una característica observable *que varía entre los diferentes individuos* de una población. La información que disponemos de cada individuo es resumida en **variables**.

En los individuos de la *población* , de uno a otro ***es variable***:

El grupo sanguíneo
{A, B, AB, O}
Su nivel de felicidad “declarado”
{Deprimido, Ni fu ni fa, Muy Feliz}
El número de hijos
{0,1,2,3,...}
La altura
{1’62 ; 1’74; ...}

Tipos de variables



Datos cuantitativos

Cuantitativas o Numéricas
Si sus valores son numéricos (tiene sentido hacer operaciones algebraicas con ellos)

- Discretas: Si toma valores enteros
  - Número de hijos, Número de cigarrillos, Num. de “cumpleaños”
- Continuas: Si entre dos valores, son posibles infinitos valores intermedios.
  - Altura, Presión intraocular, Dosis de medicamento administrado, edad

Datos cuantitativos

- Los posibles valores de una variable suelen denominarse modalidades.
- Las modalidades pueden agruparse en clases (intervalos)
  - Edades:
    - Menos de 20 años, de 20 a 50 años, más de 50 años
  - Hijos:
    - Menos de 3 hijos, De 3 a 5, 6 o más hijos
- Las modalidades/clases deben forman un sistema exhaustivo y excluyente
  - Exhaustivo: No podemos olvidar ningún posible valor de la variable
    - **Mal: ¿Cuál es su color del pelo: (Rubio, Moreno)?**
    - Bien: ¿Cuál es su grupo sanguíneo?
  - Excluyente: Nadie puede presentar dos valores simultáneos de la variable
    - Estudio sobre el ocio
      - Mal: De los siguientes, qué le gusta: (deporte, cine)
      - Bien: Le gusta el deporte: (Sí, No)
      - Bien: Le gusta el cine: (Sí, No)
      - Mal: Cuántos hijos tiene: (Ninguno, Menos de 5, Más de 2)

Datos cualitativos

Cualitativas
Si sus valores (*modalidades*) no se pueden asociar naturalmente a un número (no se pueden hacer operaciones algebraicas con ellos)

- **Nominales**: Si sus valores no se pueden ordenar
  - Sexo, Grupo Sanguíneo, Religión, Nacionalidad, Fumar (Sí/No)
- **Ordinales**: Si sus valores se pueden ordenar
  - Mejoría a un tratamiento, Grado de satisfacción, Intensidad del dolor

Los datos cualitativos (nominales u ordinales) se cuantifican como recuentos del **número de casos** observados para cada categoría, y suelen expresarse habitualmente como **porcentajes** u otro tipo de **cocientes**.
Ej. *La proporción de mujeres con síndrome X es del 82 % (55 de 67)*

Datos cualitativos

Es buena idea codificar las variables como números para poder procesarlas con facilidad en un ordenador.

Es conveniente asignar "etiquetas" a los valores de las variables para recordar qué significan los códigos numéricos.

Sexo (Cualit: Códigos arbitrarios)

1 = Hombre

2 = Mujer

Raza (Cualit: Códigos arbitrarios)

1 = Blanca

2 = Negra,...

Felicidad Ordinal: Respetar un orden al codificar.

1 = Muy feliz

2 = Bastante feliz

3 = No demasiado feliz

Se pueden asignar códigos a respuestas especiales como

0 = No sabe

99 = No contesta...

Estas situaciones deberán ser tenidas en cuentas en el análisis.

Datos perdidos ('missing data')

	sexo	raza	región	feliz	vida	herma	hijos	educ	edad	ed
1	Mujer	Blanca	Nor-E	Muy feliz	Excitante	1	2	12	61	No p
2	Mujer	Blanca	Nor-E	Bastante	Excitante	2	1	20	32	
3	Hombre	Blanca	Nor-E	Muy feliz	No proced	2	1	20	35	
4	Mujer	Blanca	Nor-E	No conte	Rutinaria	2	0	20	26	
5	Mujer	Negra	Nor-E	Bastante	Excitante	4	0	12	25	No
6	Hombre	Negra	Nor-E	Bastante	No proced	7	5	10	59	
7	Hombre	Negra	Nor-E	Muy feliz	Excitante	7	3	10	46	
8	Mujer	Negra	Nor-E	Bastante	No proced	7	4	16	No	

Vista de datos | Vista de variables

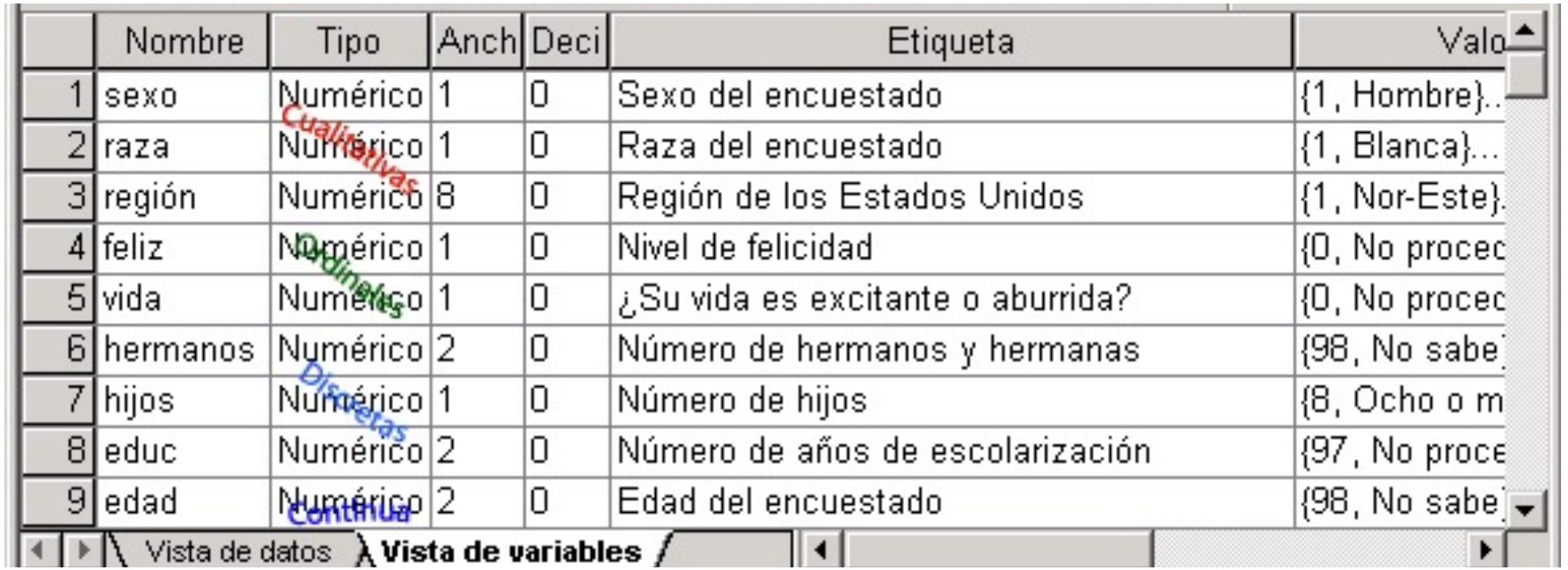
	sexo	raza	región	feliz	vida	herma	hijos	educ	edad	ed
1	2	1	1	1	1	1	2	12	61	
2	2	1	1	2	1	2	1	20	32	
3	1	1	1	1	0	2	1	20	35	
4	2	1	1	9	2	2	0	20	26	
5	2	2	1	2	1	4	0	12	25	
6	1	2	1	2	0	7	5	10	59	
7	1	2	1	1	1	7	3	10	46	
8	2	2	1	2	0	7	4	16	99	

Vista de datos | Vista de variables

Datos cualitativos

Aunque se codifiquen como números, debemos recordar siempre el verdadero tipo de las variables y su significado cuando vayamos a usar programas de cálculo estadístico.

No todo está permitido con cualquier tipo de variable.



	Nombre	Tipo	Anch	Deci	Etiqueta	Valo
1	sexo	Numérico	1	0	Sexo del encuestado	{1, Hombre}..
2	raza	Numérico	1	0	Raza del encuestado	{1, Blanca}...
3	región	Numérico	8	0	Región de los Estados Unidos	{1, Nor-Este}.
4	feliz	Numérico	1	0	Nivel de felicidad	{0, No procec
5	vida	Numérico	1	0	¿Su vida es excitante o aburrida?	{0, No procec
6	hermanos	Numérico	2	0	Número de hermanos y hermanas	{98, No sabe
7	hijos	Numérico	1	0	Número de hijos	{8, Ocho o m
8	educ	Numérico	2	0	Número de años de escolarización	{97, No proce
9	edad	Numérico	2	0	Edad del encuestado	{98, No sabe

Cualitativas
Ordinales
Discretas
Continua

Vista de datos | Vista de variables

3. Medidas de tendencia central

Moda

Es el valor que se repite más dentro de un conjunto de datos.

La moda es el valor que tiene mayor frecuencia absoluta. Se representa con $M0$

Mediana

Es un valor del conjunto de datos que mide el elemento central: La mitad de los elementos se encuentran por arriba y la otra mitad por debajo de él.

La mediana es el valor que ocupa el lugar central de todos los datos cuando éstos están ordenados de menor a mayor. Se representa con x˜.

Media

La media aritmética es el valor obtenido al sumar todos los *datos* y dividir el resultado entre el número total elementos. Se suele representar con la letra griega μμ. Si tenemos una muestra de n valores, xi, la *media aritmética*, μ, es la suma de los valores divididos por el numero de elementos; en otras palabras:

$$\mu = \frac{1}{n} \sum_i x_i$$

Desviación respecto a la media: La desviación respecto a la media es la diferencia en valor absoluto entre cada valor de la variable estadística y la media aritmética.

$$D_i = |x_i - \mu|$$

Media, mediana y moda

- **La media, la mediana y la moda son idénticas en una distribución simétrica**
- **La mediana puede ser la idónea en distribuciones sesgadas, ya que no se afecta tanto por valores extremos.**
- **Sin embargo no se cuenta con un criterio único para aplicar alguna de las tres medidas**

Varianza

La varianza es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística. La varianza intenta describir la dispersión de los *datos*. Se representa como σ2

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

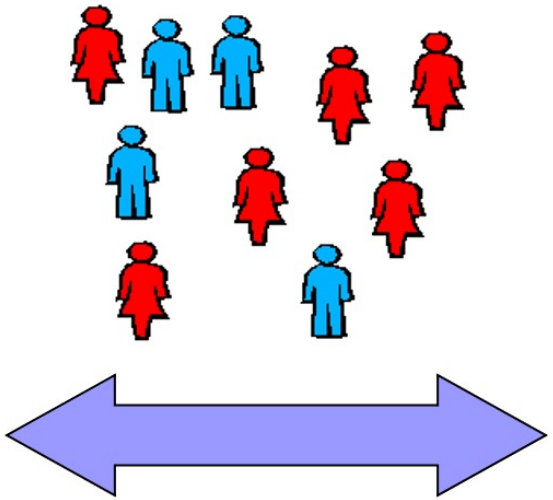
Desviación típica: La desviación típica es la raíz cuadrada de la varianza. Se representa con la letra griega σ.

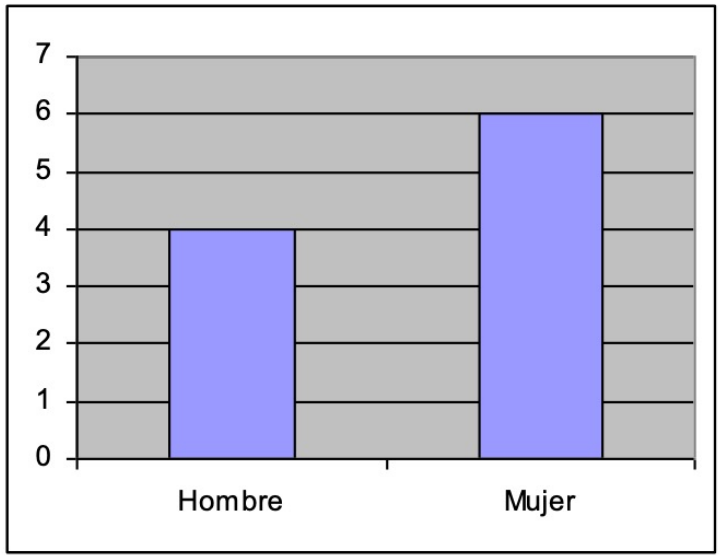
$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

4. Representación de datos

Representación ordenada de datos

Género	Frec.
Hombre	4
Mujer	6





Las tablas de frecuencias y las representaciones gráficas son dos maneras ***equivalentes*** de presentar la información. Las dos exponen ordenadamente la información recogida en una muestra.

Tablas de frecuencia

- Exponen la información recogida en la muestra, de forma que no se pierda nada de información (o poca).
  - Frecuencias absolutas: Contabilizan el número de individuos de cada modalidad
  - Frecuencias relativas (porcentajes): Idem, pero dividido por el total
  - Frecuencias acumuladas: Sólo tienen sentido para variables ordinales y numéricas
    - Muy útiles para calcular cuantiles (ver más adelante)
      - ¿Qué porcentaje de individuos tiene menos de 3 hijos? Sol: 83,8
      - ¿Entre 4 y 6 hijos? Soluc 1ª: 8,4%+3,6%+1,6%= **13,6%**. Soluc 2ª: 97,3% - 83,8% = **13,5%**

Sexo del encuestado

		Frecuencia	Porcentaje	Porcentaje válido
Válidos	Hombre	636	41,9	41,9
	Mujer	881	58,1	58,1
	Total	1517	100,0	100,0

Nivel de felicidad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Muy feliz	467	30,8	31,1	31,1
	Bastante feliz	872	57,5	58,0	89,0
	No demasiado feliz	165	10,9	11,0	100,0
	Total	1504	99,1	100,0	
Perdidos	No contesta	13	,9		
Total		1517	100,0		

Número de hijos

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	419	27,6	27,8	27,8
	1	255	16,8	16,9	44,7
	2	375	24,7	24,9	69,5
	3	215	14,2	14,2	83,8
	4	127	8,4	8,4	92,2
	5	54	3,6	3,6	95,8
	6	24	1,6	1,6	97,3
	7	23	1,5	1,5	98,9
	Ocho o más	17	1,1	1,1	100,0
	Total	1509	99,5	100,0	
Perdidos	No contesta	8	,5		
Total		1517	100,0		

Histogramas y distribuciones

Las distribuciones se pueden clasificar en dos grandes grupos:

1.Las **distribuciones continuas**, que son aquellas que presentan un número infinito de posibles soluciones. Dentro de este grupo vamos a encontrar a las distribuciones:

1. normal,
2. gamma,
3. chi cuadrado,
4. t de Student,
5. pareto,
6. entre otras

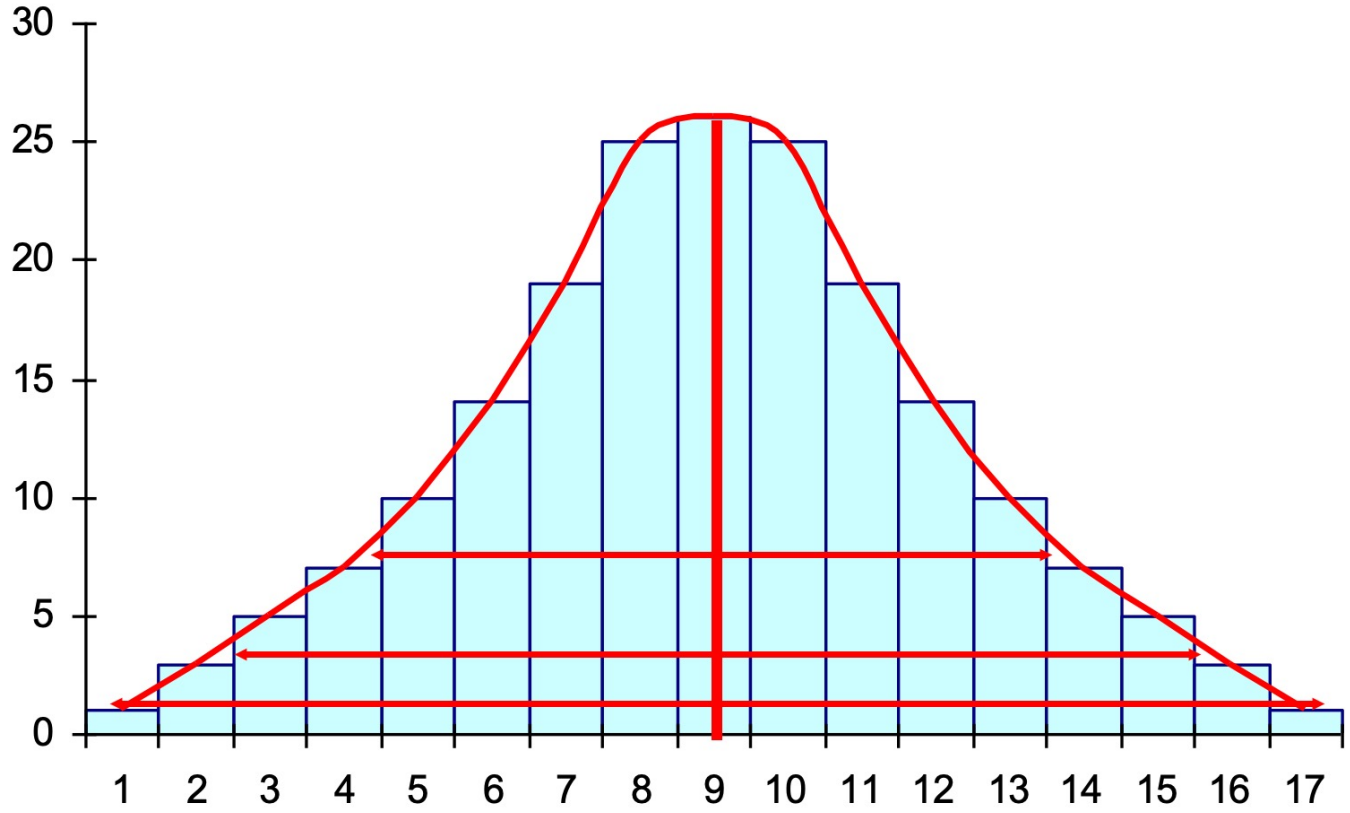
2.Las **distribuciones discretas**, que son aquellas en las que la variable puede tomar un número determinado de valores. Los principales exponentes de este grupo son las distribuciones:

1. poisson,
2. binomial,
3. hipergeométrica,
4. bernoulli
5. entre otras

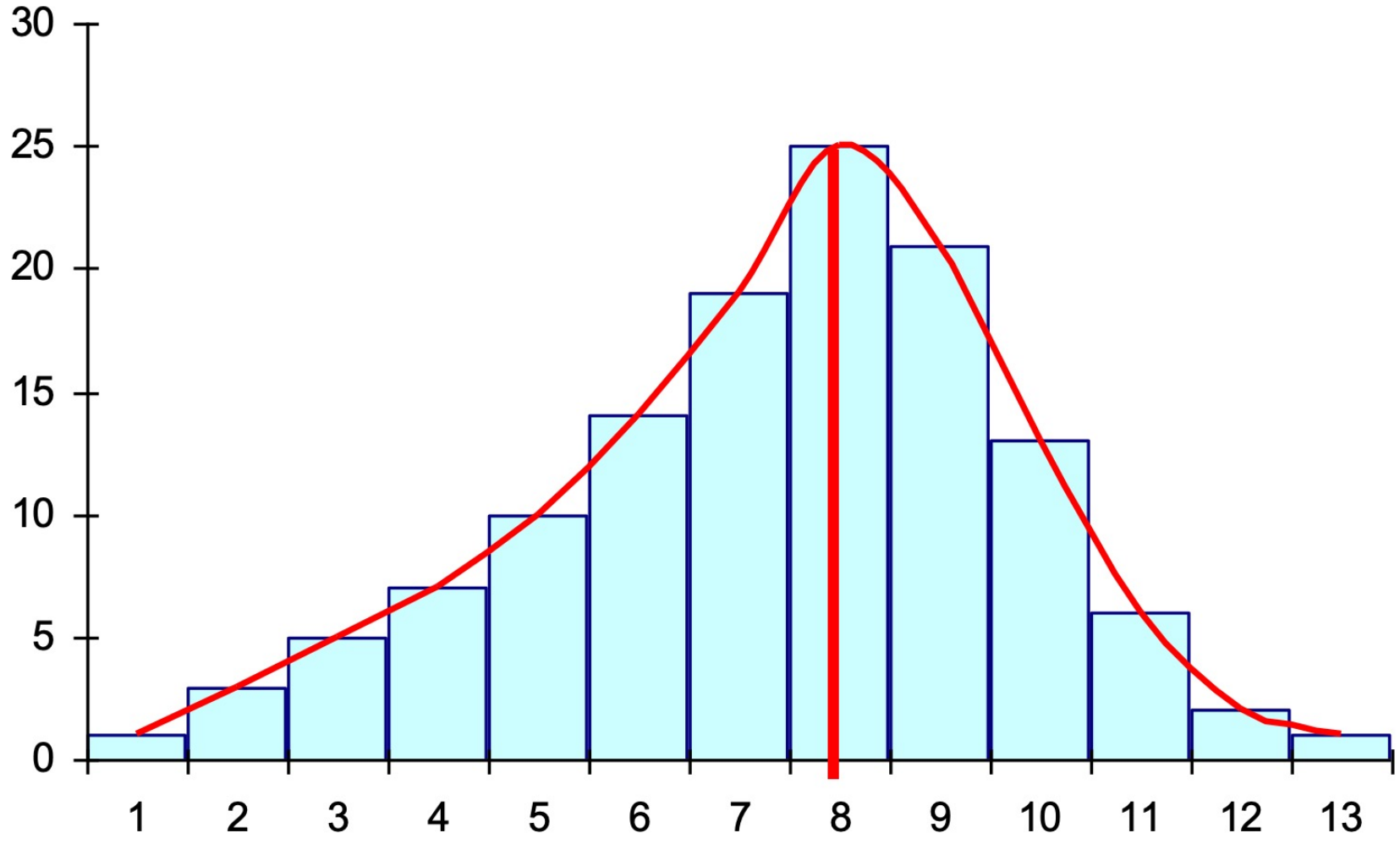
Veamos algunos ejemplos graficados con la ayuda de Python.

Distribución Normal: Curva simétrica

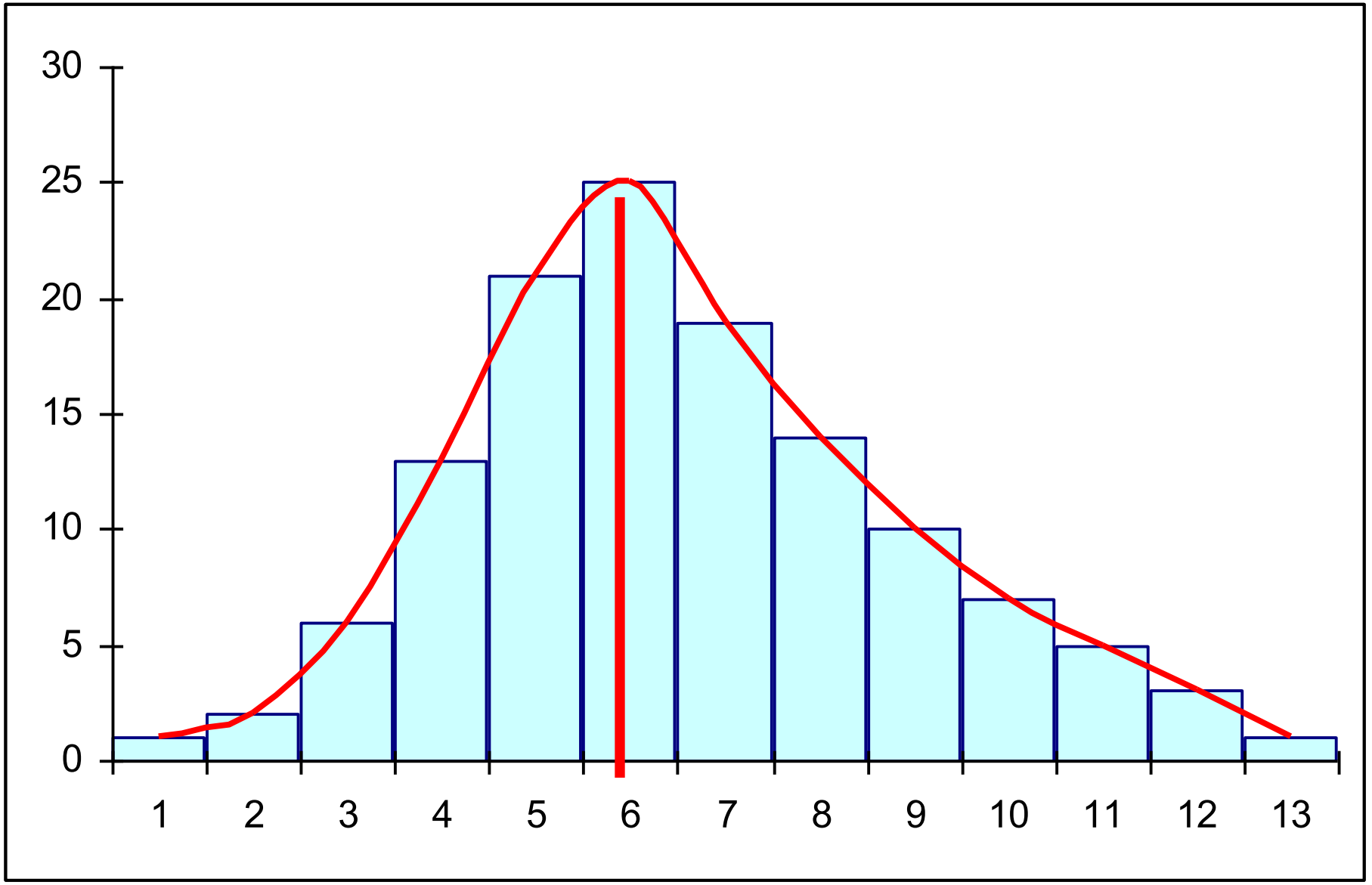
La distribución normal es una de las principales distribuciones, ya que es la que con más frecuencia aparece aproximada en los fenómenos reales. Tiene una forma acampanada y es simétrica respecto de un determinado parámetro estadístico



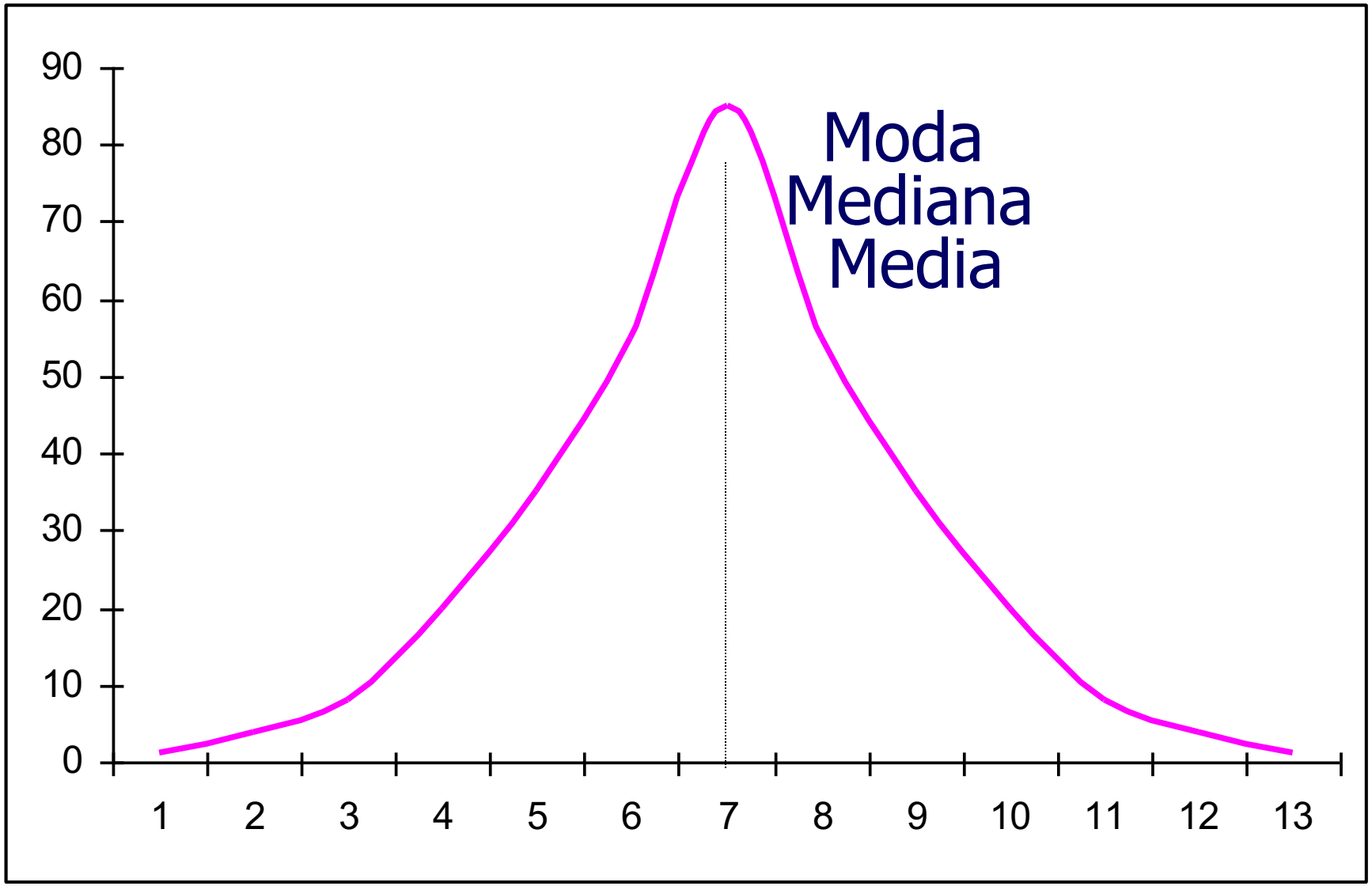
Asimetría a la izquierda



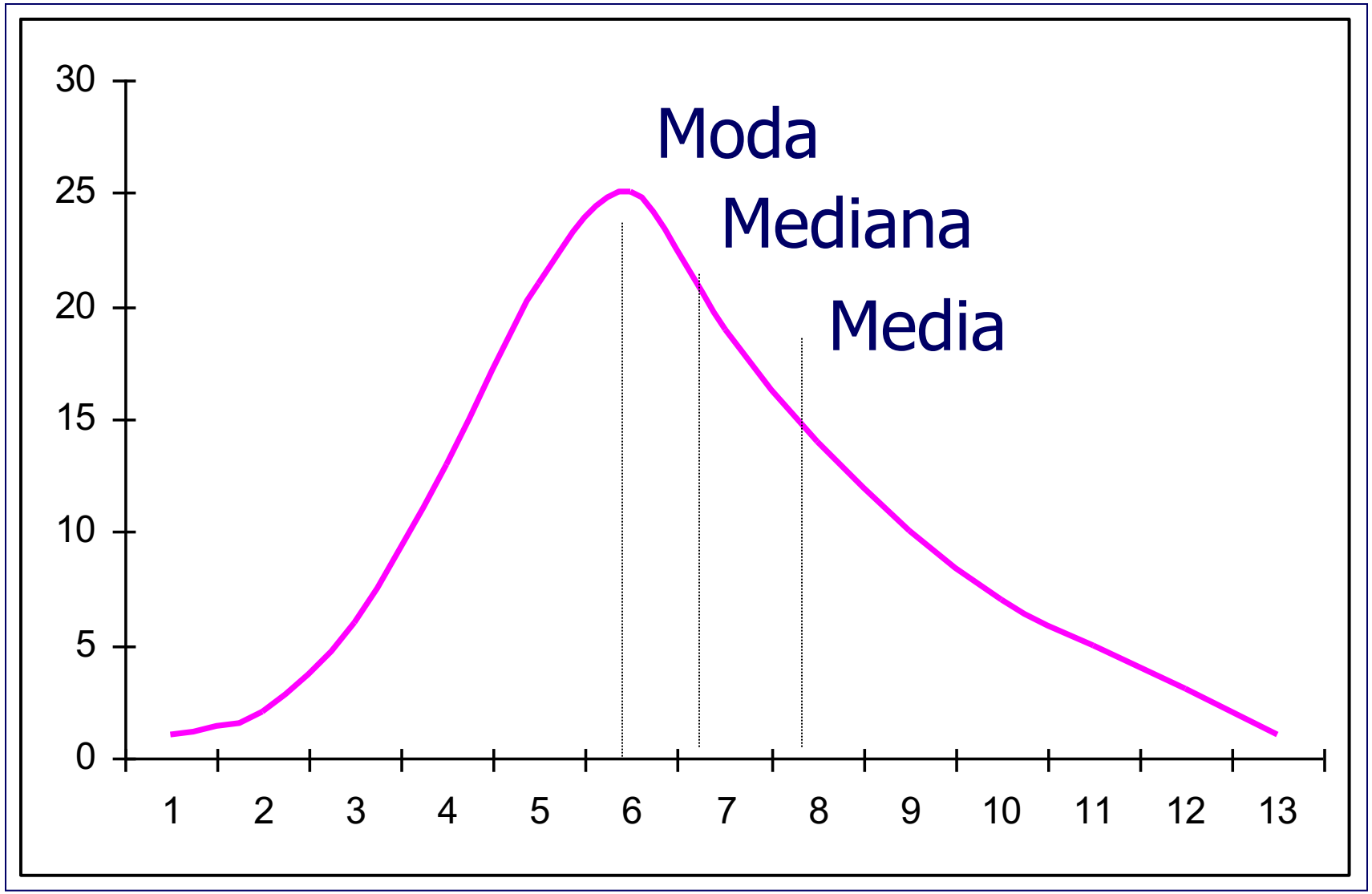
Asimetría a la derecha



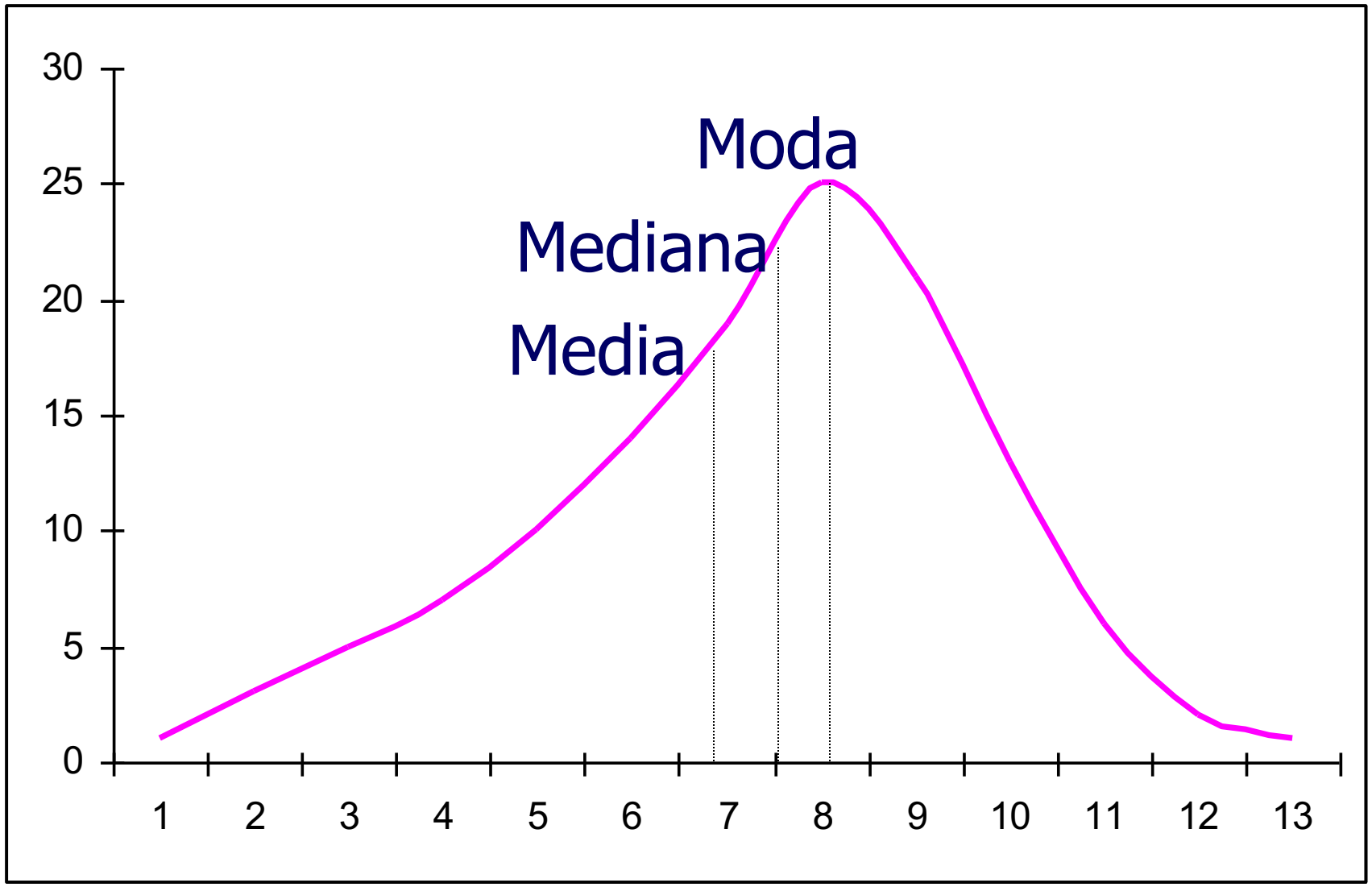
Distribución simétrica



Distribución sesgada a la derecha



Distribución sesgada a la izquierda



5. Rangos de dsipersión

Cuartiles

Los cuartiles dividen en cuatro partes las observaciones. El primer cuartil $\mathbf{Q_1}$ es un valor que deje por debajo de él 25% de las y por encima 75% de las observaciones. El $\mathbf{Q_2}$ es la mediana (50%) y $\mathbf{Q_3}$ deja por debajo 75% y por encima 25% de las observaciones

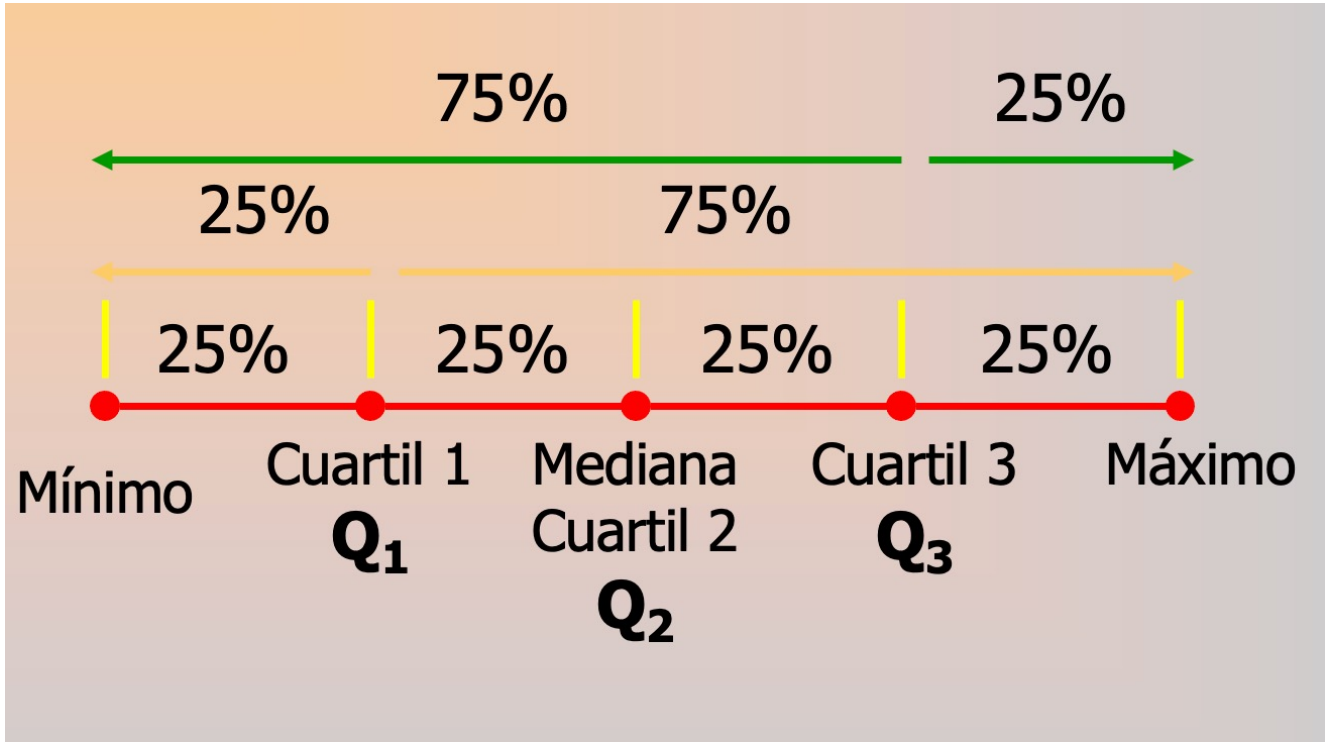
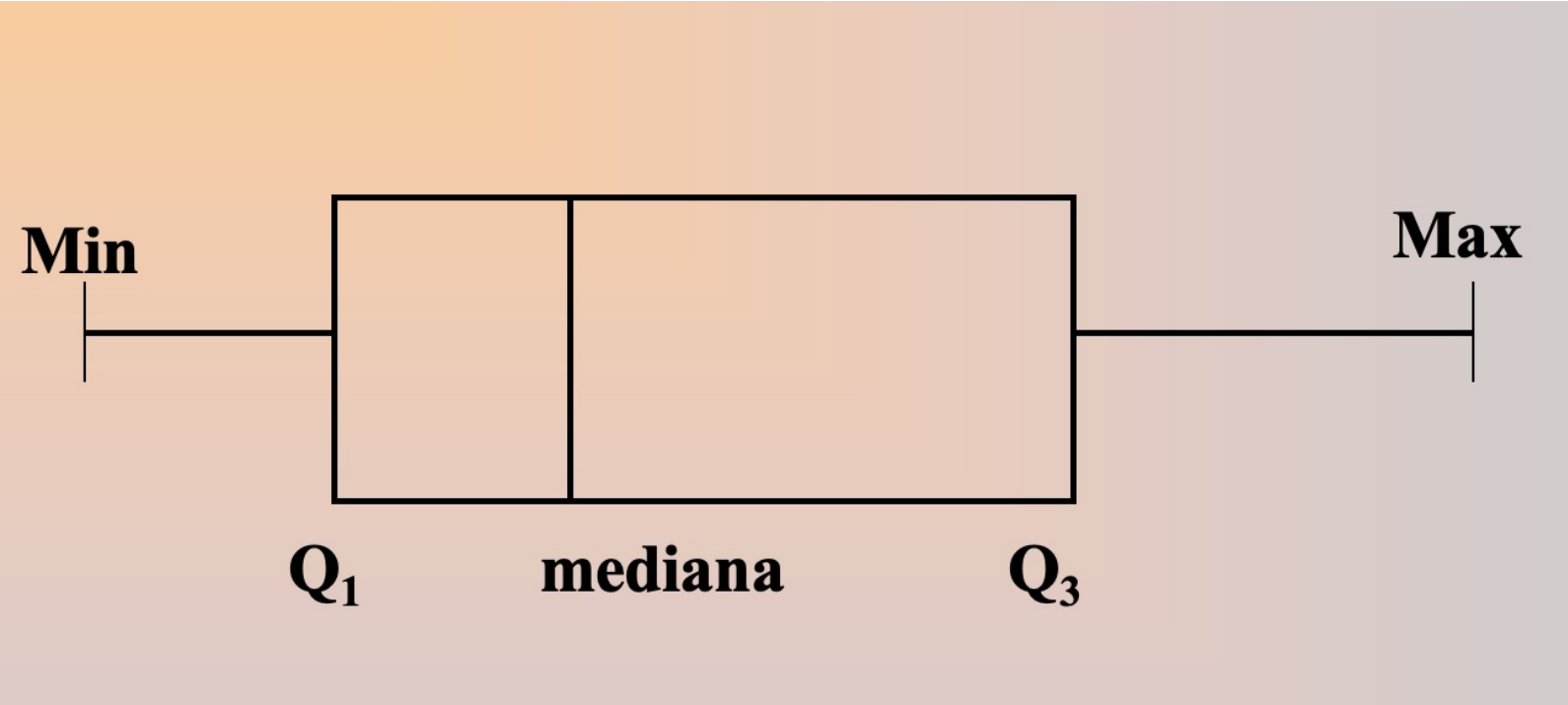
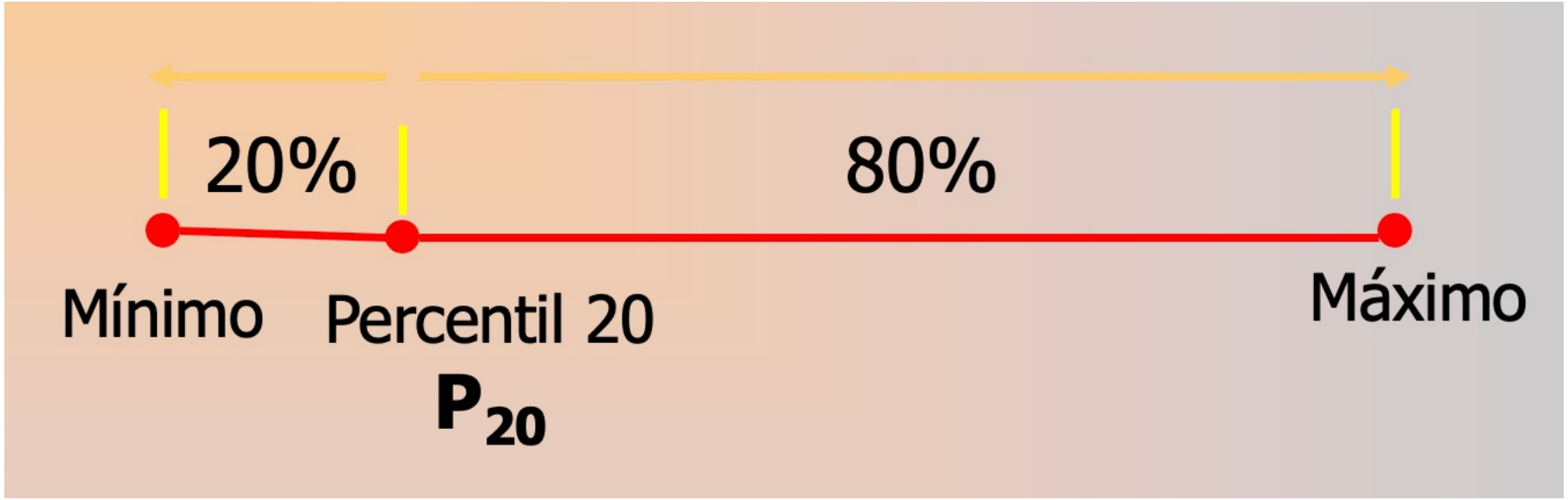


Diagrama de Caja - Boxplot



Percentiles

Los percentiles dividen en dos partes las observaciones. Por ejemplo, el percentil 20, $\mathbf{P_{20}}$, es el valor que deja por debajo un 20% y por encima un 80% de las observaciones



6. Librerías Python

Python Librerías para estadística

•**numpy**: El popular paquete matemático de Python, se utiliza tanto que mucha gente ya lo considera parte integral del lenguaje. Nos proporciona algunas funciones estadísticas que podemos aplicar fácilmente sobre los *arrays* de Numpy.

•**scipy.stats**: Este submodulo del paquete científico Scipy es el complemento perfecto para Numpy, las funciones estadísticas que no encontremos en uno, las podemos encontrar en el otro.

•**statsmodels**: Esta librería nos brinda un gran número de herramientas para explorar *datos*, estimar modelos estadísticos, realizar pruebas estadísticas y muchas cosas más.

•**matplotlib**: Es la librería más popular en Python para visualizaciones y gráficos. Ella nos va a permitir realizar los gráficos de las distintas distribuciones de datos.

•**seaborn**: Esta librería es un complemento ideal de matplotlib para realizar gráficos estadísticos.

•**pandas**: Esta es la librería más popular para análisis de *datos* y financieros. Posee algunas funciones muy útiles para realizar *estadística descriptiva* sobre nuestros datos y nos facilita sobremanera el trabajar con series de tiempo.

•**pyMC**: pyMC es un módulo de Python que implementa modelos estadísticos bayesianos, incluyendo la cadena de Markov Monte Carlo(MCMC). pyMC ofrece funcionalidades para hacer el análisis bayesiano lo mas simple posible.

A Practicar!!!!

KIBERNUM

Creando Juntos Excelencia

Software Engineering | IT Staffing | IT Academy | IT Consulting

Proyecto apoyado por CORFO