KIBERNUM
IT Academy

Módulo 3 – Análisis Exploratorio y Programación Estadística

Conceptos Básicos de Estadística Descriptiva

Ciencia de Datos

Definición

¿Cómo se define la estadística?

La estadística se define como una disciplina matemática que se utiliza para recopilar, analizar, interpretar y presentar datos numéricos. La estadística se utiliza en una amplia variedad de campos, incluyendo la investigación científica, el análisis de negocios, la economía, la psicología, la medicina y muchos otros.

La estadística implica la aplicación de métodos y técnicas para la recopilación de datos, la descripción de datos, la inferencia estadística y la toma de decisiones basada en datos. Estos métodos y técnicas incluyen la probabilidad, la estadística descriptiva, la estadística inferencial, el análisis de regresión, el diseño de experimentos, la visualización de datos y la toma de decisiones bajo incertidumbre.

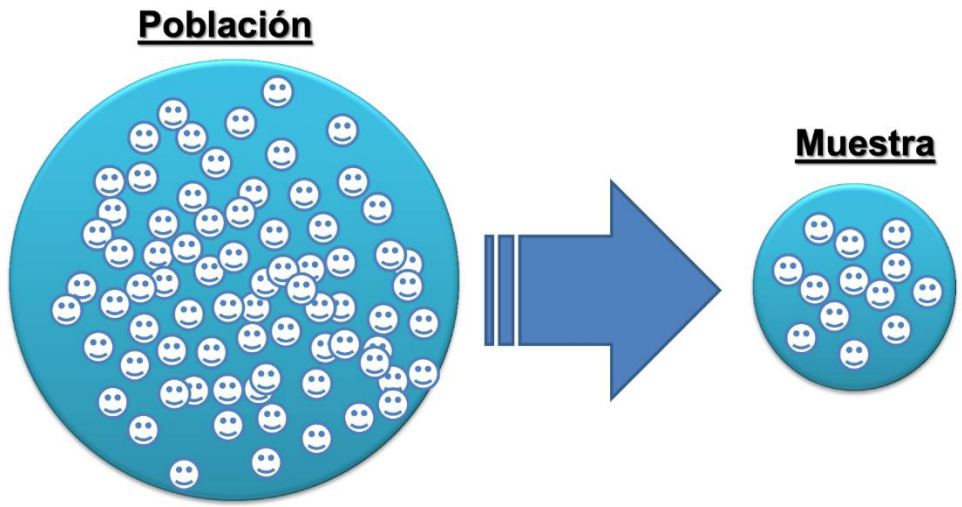
El objetivo principal de la estadística es ayudar a las personas a comprender mejor los datos, hacer inferencias y tomar decisiones informadas. Al aplicar métodos estadísticos, se pueden obtener conclusiones basadas en evidencia a partir de los datos y se pueden evaluar las hipótesis y teorías existentes.

KIBERNUM

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.

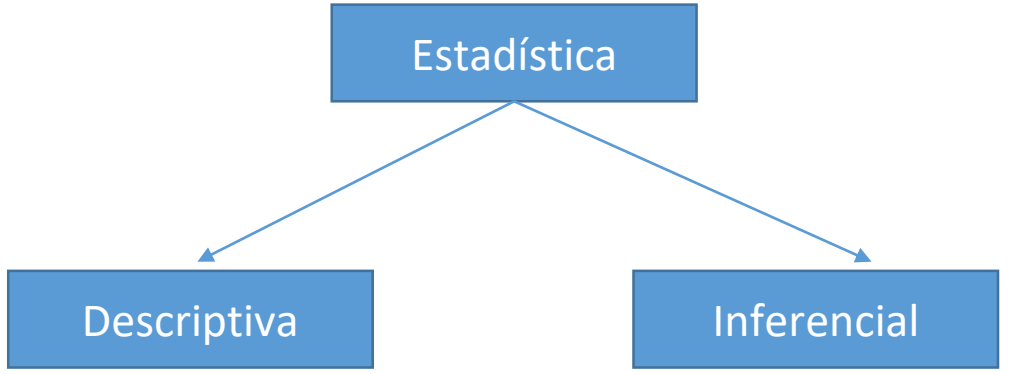
Muestra (‘sample’) es un subconjunto suyo al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones). Debería ser “representativo”. Está formado por miembros “seleccionados” de la población (individuos, unidades experimentales).



KIBERNUM

Tipos de estadística

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



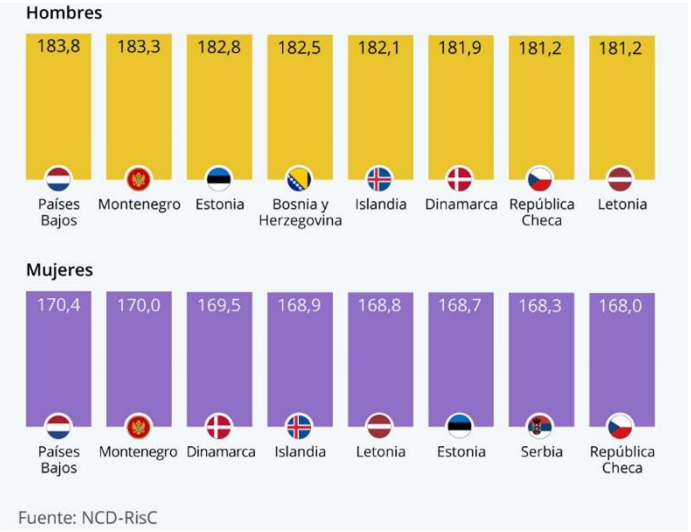
KIBERNUM

Estadística descriptiva

La estadística descriptiva es una rama de la estadística que se enfoca en resumir y describir características importantes de un conjunto de datos. Su objetivo principal es proporcionar una comprensión clara y concisa de la información presente en los datos, sin inferir conclusiones más allá de los datos mismos.

Se encarga de recoger, almacenar, ordenar, realizar tablas o gráficos, calcular parámetros básicos sobre el conjunto de datos.

Describe de forma cuantitativa un fenómeno.



Estatura media de los jóvenes de 19 años en 2019.

KIBERNUM

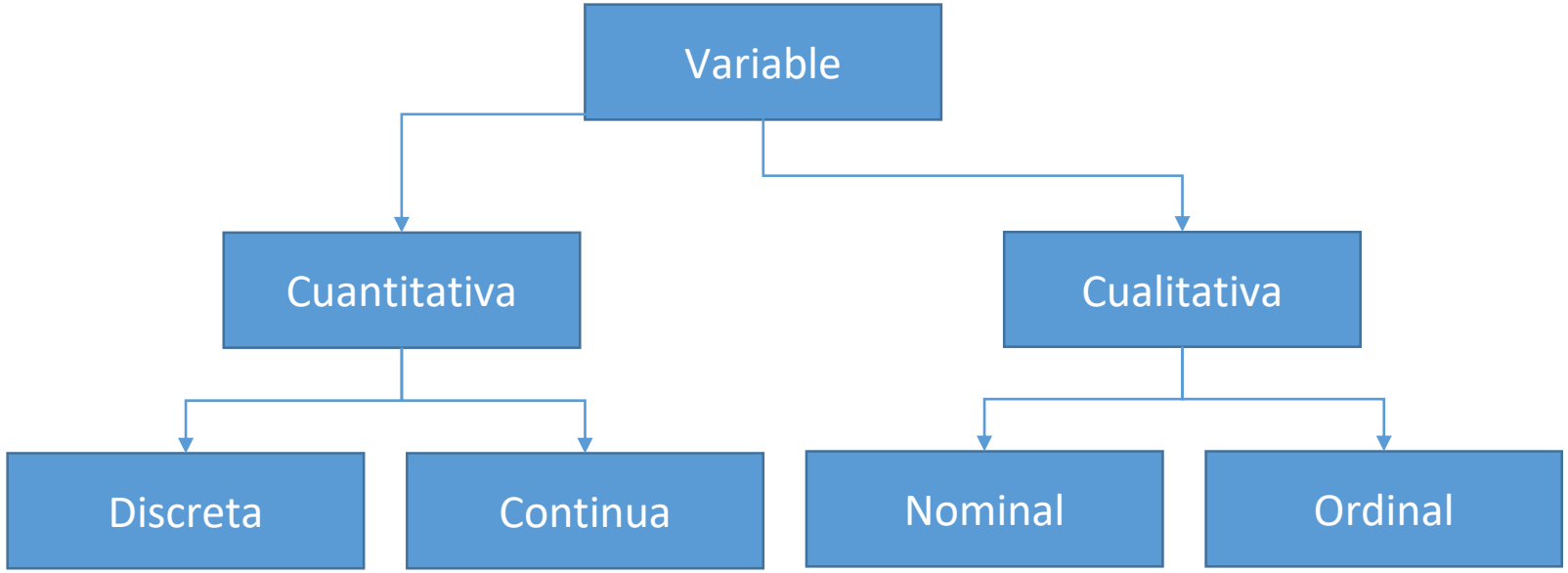
Variables

Una variable estadística es una **característica o propiedad de un fenómeno o evento que puede ser medida, cuantificada o categorizada, y que puede variar de un individuo a otro dentro de una población o muestra**. En otras palabras, una variable estadística es cualquier característica o atributo que puede ser observado y medido en un conjunto de datos.

Ejemplos de **características que varían** en una muestra:

- Edad de los estudiantes en una clase.
  {18, 19, 19 ,17, 18,…}
- Género de los encuestados en una investigación.
  {M, F, F, M, M,…}
- Nivel de ingresos de los hogares en una comunidad.
  {Alto, Medio, Medio Alto, …}
- Altura de los jugadores de un equipo de baloncesto.
  {1.94, 2.01, 1.89, …}
- Número de hermanos de una persona.
- Temperatura registrada durante el día en una ciudad.
- Puntuación obtenida en un examen.
- Color de los ojos de los habitantes de una ciudad.
- Precio de una acción en la bolsa de valores.
- Tiempo empleado por un vehículo en recorrer una distancia determinada.

Tipos de variables



Datos Cuantitativos

Estas variables representan **cantidades numéricas que se pueden medir o contar** (tiene sentido hacer operaciones aritméticas con ellas).

Se dividen en dos subtipos:

- **Discretas**: Los valores de estas variables son números enteros que representan conteos o recuentos. No pueden tener valores fraccionarios o continuos. Por ejemplo, el número de hijos en una familia o el número de estudiantes en una clase.

- **Continuas**: Los valores de estas variables son números reales que representan mediciones precisas y pueden tener un rango infinito de valores posibles. Por ejemplo, la altura, el peso, la temperatura o el tiempo.

Datos Cuantitativos

- Los posibles valores de una variable, suele denominarse **modalidades**.
- Las modalidades pueden agruparse en **clases (intervalos)**.
  - Por ejemplo:

    Edades (menor de 20 años, de 20 a 50 años, más de 50 años).

    Hijos (menos de 3 hijos, entre 3 y 5, 6 o más).
- Las modalidades/clases deben formar un sistema **exhaustivo y excluyente**.
  - Exhaustivo: No podemos olvidar ningún posible valor de la variable.
  - Excluyente: Nadie puede presentar dos valores simultáneos de la variable.

Datos Cualitativos

Un dato cualitativo es una información o característica que describe cualidades o atributos que **no pueden ser medidos numéricamente** (no pueden hacerse operaciones aritméticas con ellos). Estos datos representan categorías o clasificaciones y se utilizan para describir características cualitativas de un fenómeno o una población.

Los datos cualitativos pueden ser de dos tipos:

- **Nominales**: Los datos nominales representan categorías o nombres que no tienen un orden inherente. Por ejemplo, el género (masculino, femenino), el tipo de sangre (A, B, AB, O), el estado civil (soltero, casado, divorciado) o el color de los ojos (azul, marrón, verde).

- **Ordinales**: Los datos ordinales también representan categorías, pero tienen un orden natural o jerarquía. Aunque los valores son categorías, tienen un orden. Por ejemplo, el nivel educativo (primario, secundario, terciario), el grado de satisfacción (muy insatisfecho, insatisfecho, neutral, satisfecho, muy satisfecho) o la clasificación socioeconómica (baja, media, alta).

Datos Cualitativos

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).

Este es un ejemplo de una Tabla de Frecuencias para describir una variable cualitativa:

Tipo de transporte	Frecuencia
Automóvil	120
Transporte público	80
Bicicleta	30
A pie	50

Codificación de variables cualitativas

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	

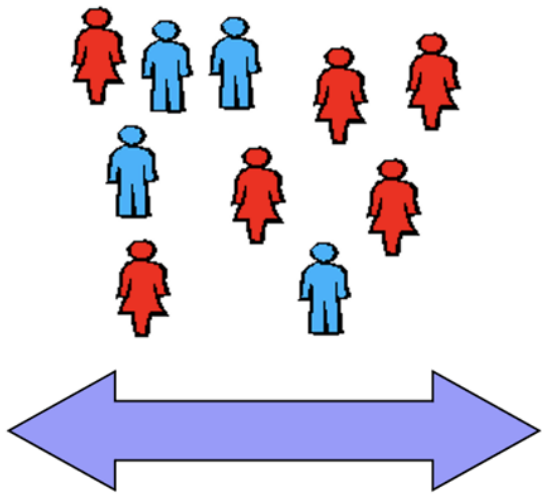
Vista de datos | Vista de variables

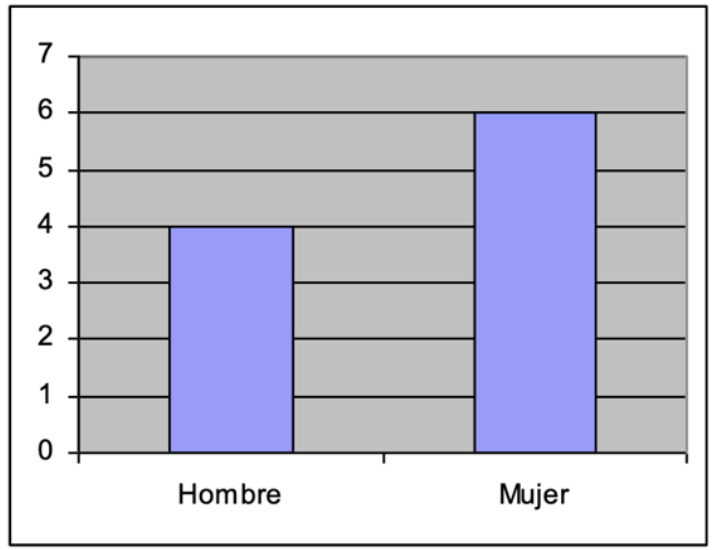
KIBERNUM

Representación ordenada de datos

Las **Tablas de Frecuencia** y las **Representaciones Gráficas** son dos maneras equivalentes de presentar la información. Las dos exponen ordenadamente la información recogida en una muestra.

Género	Frec.
Hombre	4
Mujer	6





Tablas de frecuencia

Una **tabla de frecuencia** es una herramienta utilizada en estadística para **organizar y resumir datos**, mostrando la **frecuencia con la que ocurren diferentes valores** en un conjunto de datos. Básicamente, cuenta cuántas veces aparece cada valor o rango de valores en los datos.

En una tabla de frecuencia, los valores únicos del conjunto de datos se enumeran en una columna, mientras que en la columna adyacente se muestra la frecuencia de cada valor, es decir, el número de veces que aparece en los datos.

- **Frecuencia Absoluta**: contabiliza el número de individuos de cada modalidad
- **Frecuencia Relativa**: ídem, pero dividido por el total (porcentajes)

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

KIBERNUM

Tablas de frecuencia

Algunos ejemplos más:

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		

Dudas y consultas

KIBERNUM

Fin de la Presentación

KIBERNUM