KIBERNUM
IT Academy

Módulo 1
Clase 3

Introducción a la Ciencia de Datos

Introducción a la Estadística

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.

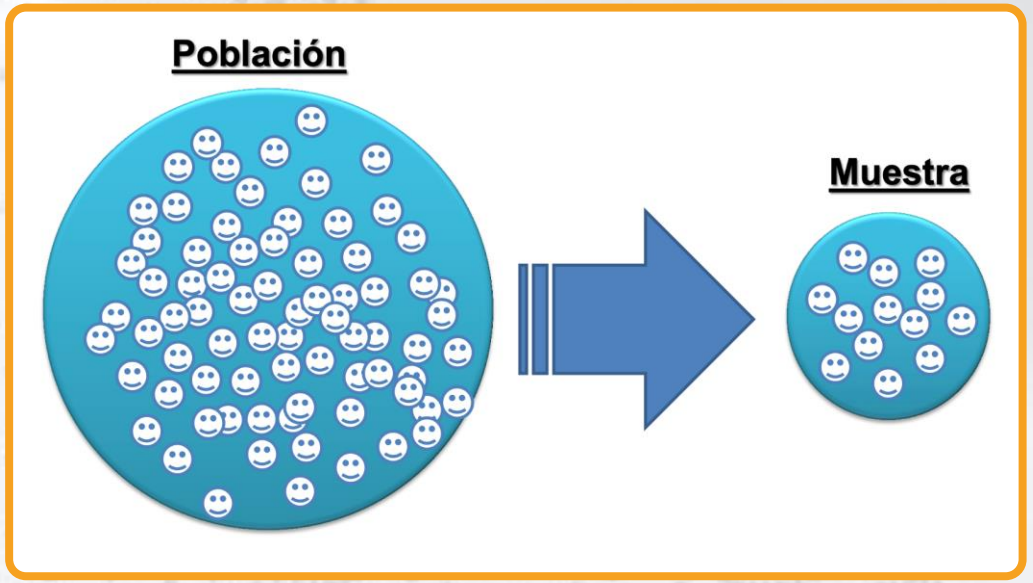
La estadística implica la aplicación de métodos y técnicas para la recopilación de datos, la descripción de datos, la inferencia estadística y la toma de decisiones basada en datos. Estos métodos y técnicas incluyen la probabilidad, la estadística descriptiva, la estadística inferencial, el análisis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

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.

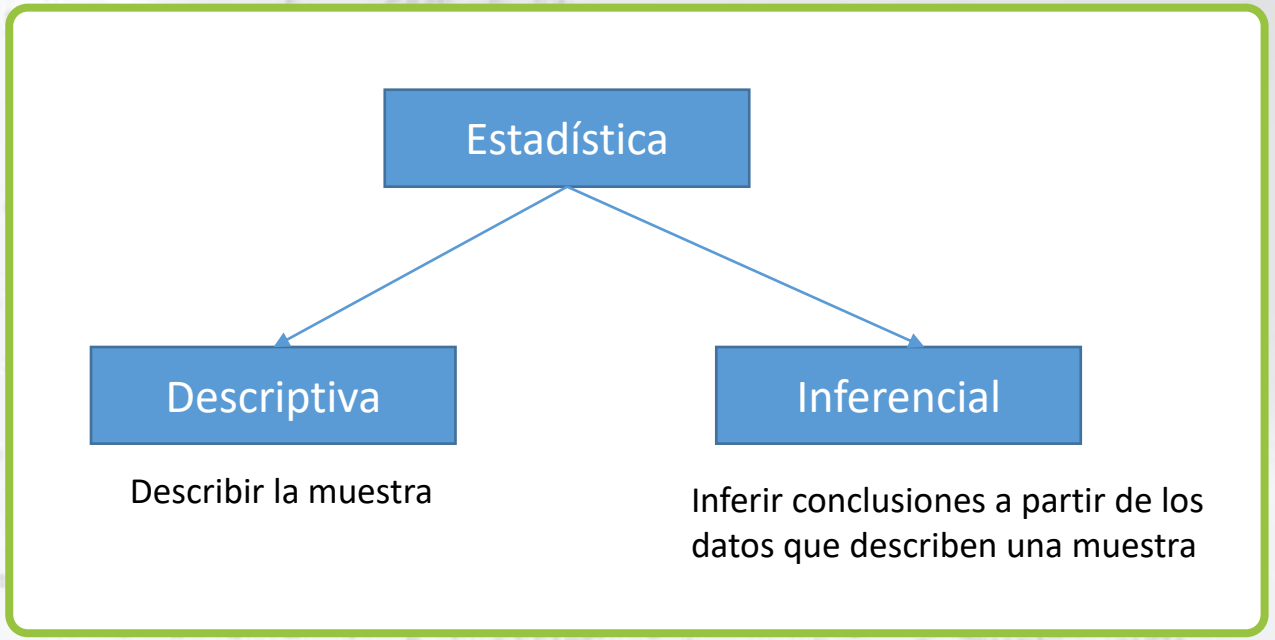
Muestra (‘sample’) es un subconjunto suyo al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones). Debería ser “representativo”. Esta formado por miembros “seleccionados” de la población (individuos, unidades experimentales).



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



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.

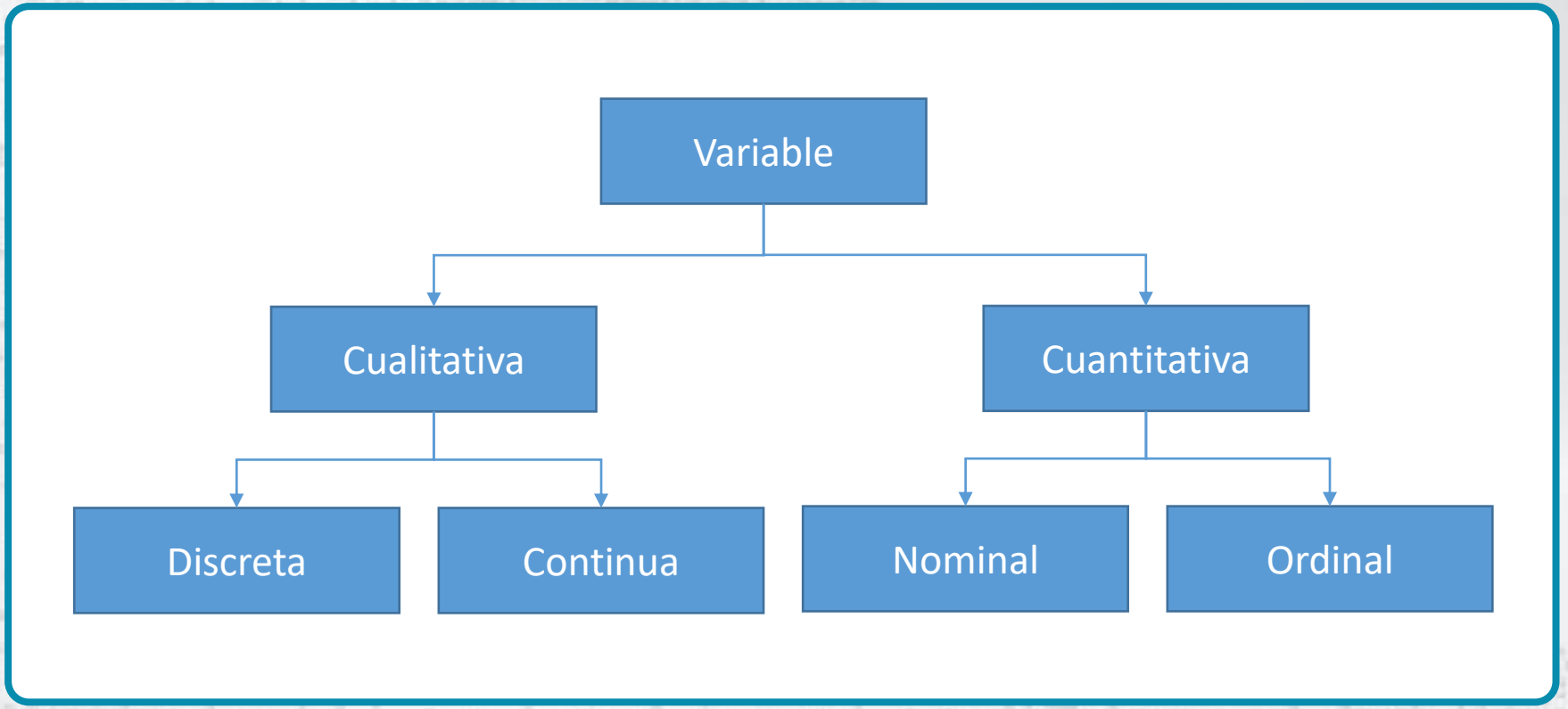
Variables

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)

Tipos de variables

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

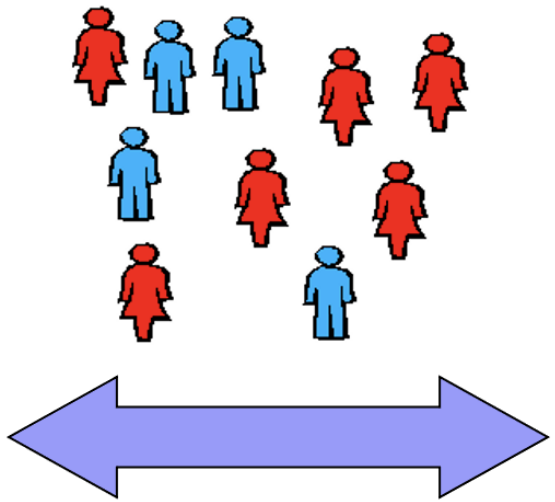
	sexo	raza	región	feliz	vida	herma	hijos	educ	edad	ed
1	2	1	1	1	1	1	2	12	61	
2	2	1	1	2	1	2	1	20	32	
3	1	1	1	1	0	2	1	20	35	
4	2	1	1	9	2	2	0	20	26	
5	2	2	1	2	1	4	0	12	25	
6	1	2	1	2	0	7	5	10	59	
7	1	2	1	1	1	7	3	10	46	
8	2	2	1	2	0	7	4	16	99	

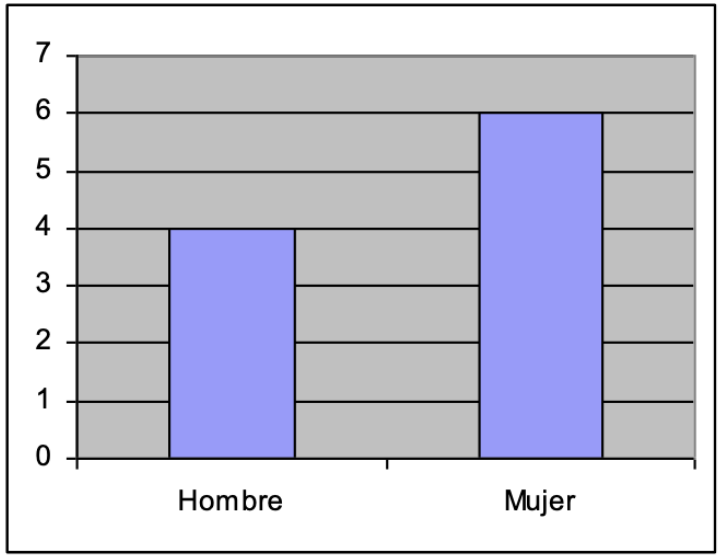
Vista de datos | Vista de variables

Representación ordenada de datos

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.

Género	Frec.
Hombre	4
Mujer	6





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		

Gracias