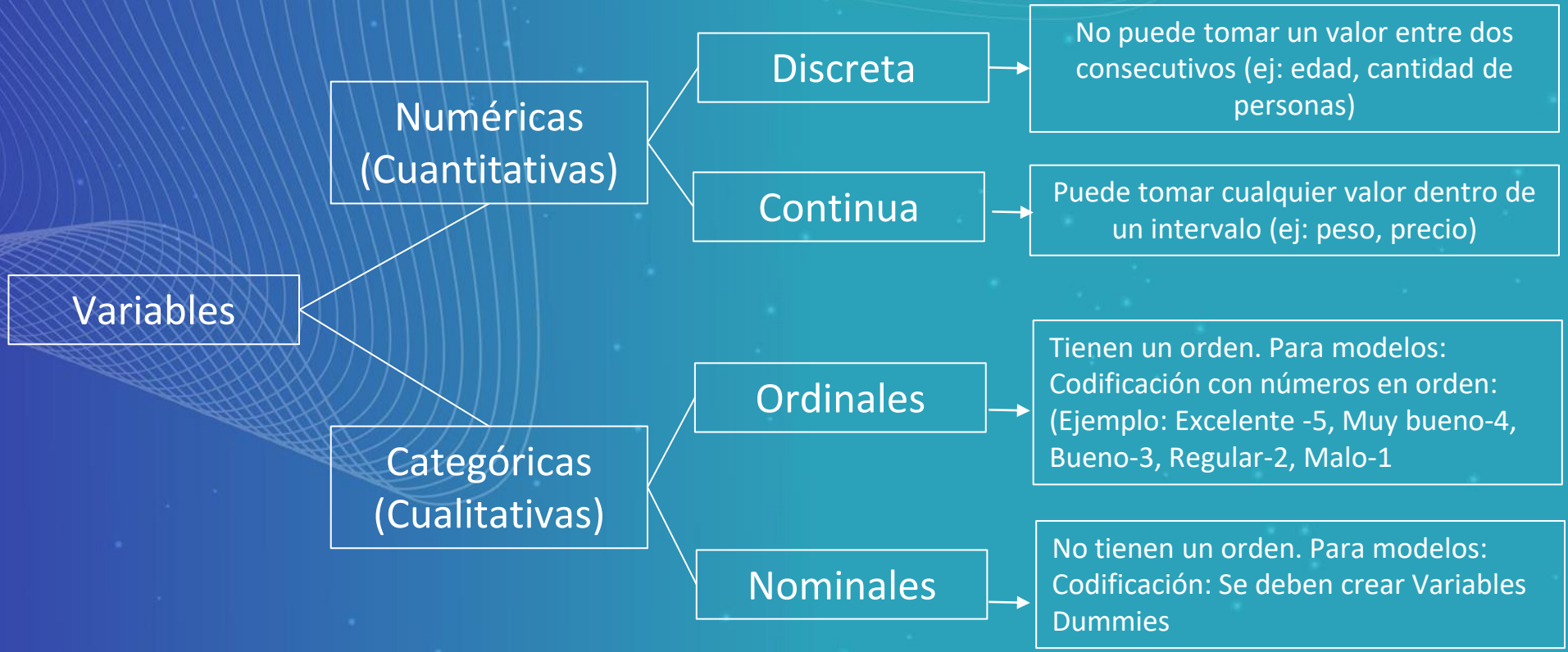
Trayecto Datos

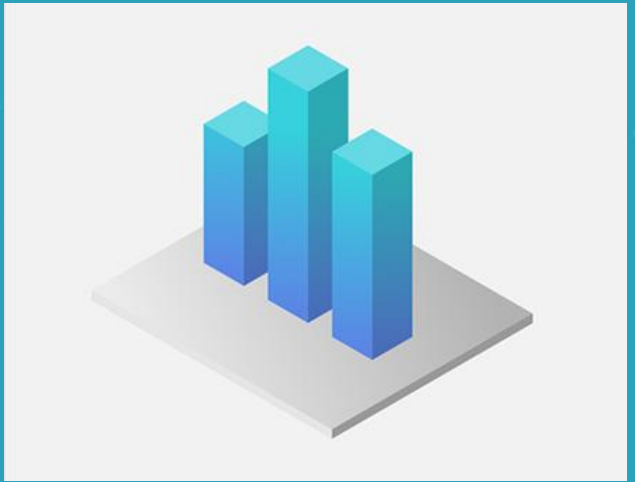
Clase 6: Estadística

Variables


Variables
Numéricas (Cuantitativas)
Discreta
No puede tomar un valor entre dos consecutivos (ej: edad, cantidad de personas)
Continua
Puede tomar cualquier valor dentro de un intervalo (ej: peso, precio)
Categóricas (Cualitativas)
Ordinales
Tienen un orden. Para modelos: Codificación con números en orden: (Ejemplo: Excelente -5, Muy bueno-4, Bueno-3, Regular-2, Malo-1
Nominales
No tienen un orden. Para modelos: Codificación: Se deben crear Variables Dummies

Estadística

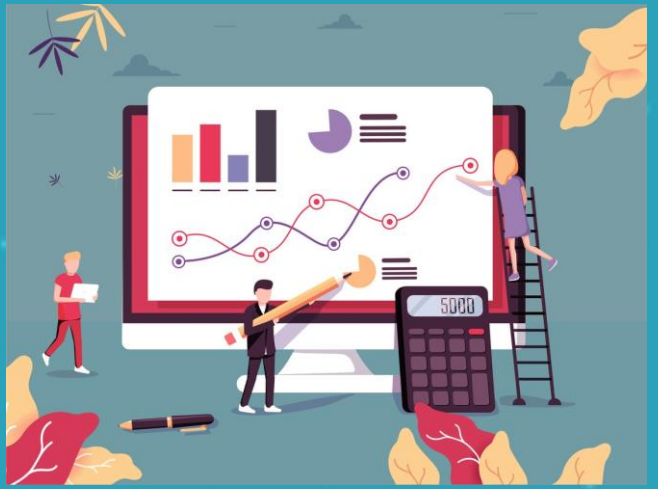
La estadística es la ciencia que se encarga de inferir valores o generar predicciones sobre algún evento del cual tenemos información previa.
Básicamente sirve para obtener información general con solo una muestra de las observaciones, permitiéndonos calcular
la confianza que podemos tener sobre nuestras predicciones.



Estadística - Tipos

Existen dos tipos de estadística:

- Descriptiva: Busca describir los datos que se tienen, analizar distribuciones y correlaciones, etc. Tiene una relación directa con el análisis exploratorio de los datos
- Inferencial: Conociendo el comportamiento de los datos recolectados, busca obtener información más general sobre el total de los datos (no solo los recolectados). De alguna forma relacionado a la parte del modelado en la ciencia de datos actual



Media

La media es el promedio de los datos numéricos

Se calcula sumando todos lo valores y dividiendo por la cantidad de registros

Conjunto de datos | **5 3 2 13 9 7 6 1 4 3 6 2 3**

Suma | **5 + 3 + 2 + 13 + 9 + 7 + 6 + 1 + 4 + 3 + 6 + 2 + 3 = 64**

Media/Promedio | **64 / 13 = 4,92**

Mediana

La mediana es el valor del medio en un conjunto de datos numéricos ordenados. Se calcula ordenando los números de menor a mayor y seleccionando el valor del medo.

Si la cantidad de valores es impar es el del medio, si es par es el promedio entre los dos del medio.

Conjunto de datos **5 3 2 13 9 7 6 1 4 3 6 2 3**

Ordenados **1 2 2 3 3 3 4 5 6 6 7 9 13**

6 6

Mediana **4**

Moda

La moda es el valor que aparece con mayor frecuencia en un conjunto de datos numéricos o categóricos, es decir el que más se repite

Conjunto de datos

5	3	2	13	9	7	6	1	4	3	6	2	3

Moda **3**

De este conjunto de datos numéricos:

Media **4,92**

Mediana **4**

Moda **3**

Cuartiles

Los cuartiles se obtienen dividiendo un conjunto de datos numéricos en 4, siendo que el primer cuartil Q1, es el valor en el cual o por debajo del cual queda el 25% del conjunto de datos, el segundo cuartil Q2 es el valor por debajo del cual queda el 50% de los datos (coincide con la Mediana), el tercer cuartil Q3 es el valor por debajo del cual quedan el 75%:

Conjunto de datos	5	3	2	13	9	7	6	1	4	3	6	2	3

Ordenados	1	2	2	3	3	3	4	5	6	6	7	9	13

Q1: 3 Q2: 4 Q3: 6

Medidas de dispersión

El Desvío estándar sirve para conocer la dispersión de los datos numéricos de una variable.

Primero vamos a ver que es el **Rango** y la **Varianza**

Rango: Diferencia entre valor más alto y más bajo

Conjunto de datos	5	3	2	13	9	7	6	1	4	3	6	2	3
Ordenados	1	2	2	3	3	3	4	5	6	6	7	9	13

Rango: **13 – 1 = 12**

Medidas de dispersión

Varianza: Es la suma de las diferencias entre cada valor y la media al cuadrado divido la cantidad de registros

$$\frac{\sum (X - \bar{X})^2}{N}$$

Conjunto de datos | 5 | 3 | 2 | 13 | 9 | 7 | 6 | 1 | 4 | 3 | 6 | 2 | 3 |

$$(5-4)^2 + (3-4)^2 + (2-4)^2 + (13-4)^2 + (9-4)^2 + (7-4)^2 + (6-4)^2 + (1-4)^2 + (4-4)^2 + (3-4)^2 + (6-4)^2 + (2-4)^2 + (3-4)^2$$

$$13$$

$$1 + 1 + 4 + 81 + 25 + 9 + 4 + 9 + 0 + 1 + 4 + 4 + 1 = 144$$

Varianza: **144 / 13 = 11.08**

Medidas de dispersión

Desvío Estándar: Es la raíz cuadrada de la varianza

$$\sqrt{\frac{\sum_i^N (X_i - \bar{X})^2}{N}}$$

Conjunto de datos	5	3	2	13	9	7	6	1	4	3	6	2	3

Varianza: **144 / 13 = 11.08**

Desvío Estándar: **3,33**

Valores atípicos

Valor atípico (Outlier) son valores que no son consistente con los demás datos. Eestadísticamente son observaciones numéricamente distantes de demás datos, es decir, son un valor demasiado alto o demasiado bajo. Esto puede indicar que hay un error en los datos.

No hay manera única de calcularlo, las visualizaciones son una manera. Una forma es con el rango intercuartil que es la diferencia entre Q3 y Q1.

Conjunto de datos	**5 3 2 13 9 7 6 1 4 3 6 2 3 // Q1 = 3 Q3 = 6**

Rango intercuartil (IRQ): 6 – 3 = 3 ⇒ 3 x 1.5 = **4.5**

En este caso un valor atípico es un valor menor a 1.5 o mayor a 10.5

En este conjunto de datos 13 es un outlier

¡MUCHAS GRACIAS!

UTN AVELLANEDA

aprendé programando

GYEE