CIENCIA
DE DATOS

Conceptos preliminares

MATEMÁTICA PARA CIENCIA DE DATOS

PROF. ESTEBAN BALLESTERO

Estructura de los Datos…

- Una variable es una propiedad o característica de un individuo
- Una colección de variables describe a un individuo
- Un individuo también se conoce como un registro, un punto, un caso, objeto, entidad, ejemplo de observación, etc.

Ind	Estado civil	Salario_mensual	Sexo	Fraude
1	Casado	732456,00	H	No
2	Soltero	543232,34	M	No
3	Casado	289045,12	H	No
4	Casado	303662,28	H	No
5	Casado	1205607,78	H	No
6	Soltero	186042,98	M	No
7	Soltero	938435,76	M	Si
8	Casado	274355,54	M	No
9	Casado	997302,37	M	No
10	Casado	452938,61	H	Si

Tipos de variables: Cualitativas vs cuantitativas

- Cualitativo (categórico): Las variables representan una cualidad del individuo en distintas categorías, en lugar de números. Operaciones matemáticas no se pueden aplicar sobre ellas.

- Ejemplos: color de ojos, sexo, estado civil, ciudad de procedencia,

- OJO: algunas variables cualitativas pueden venir disfrazadas de números. Ejemplo, si para la variables sexo se usa 1 para hombre y 0 para mujer.

Tipos de variables: Cualitativas vs cuantitativas

- Cuantitativo (numéricos): Las variables representan números y pueden ser tratados como tales, es decir, admiten operaciones matemáticas sobre ellos.

- Ejemplos: peso, fallos por hora, temperatura, número de vehículos etc.

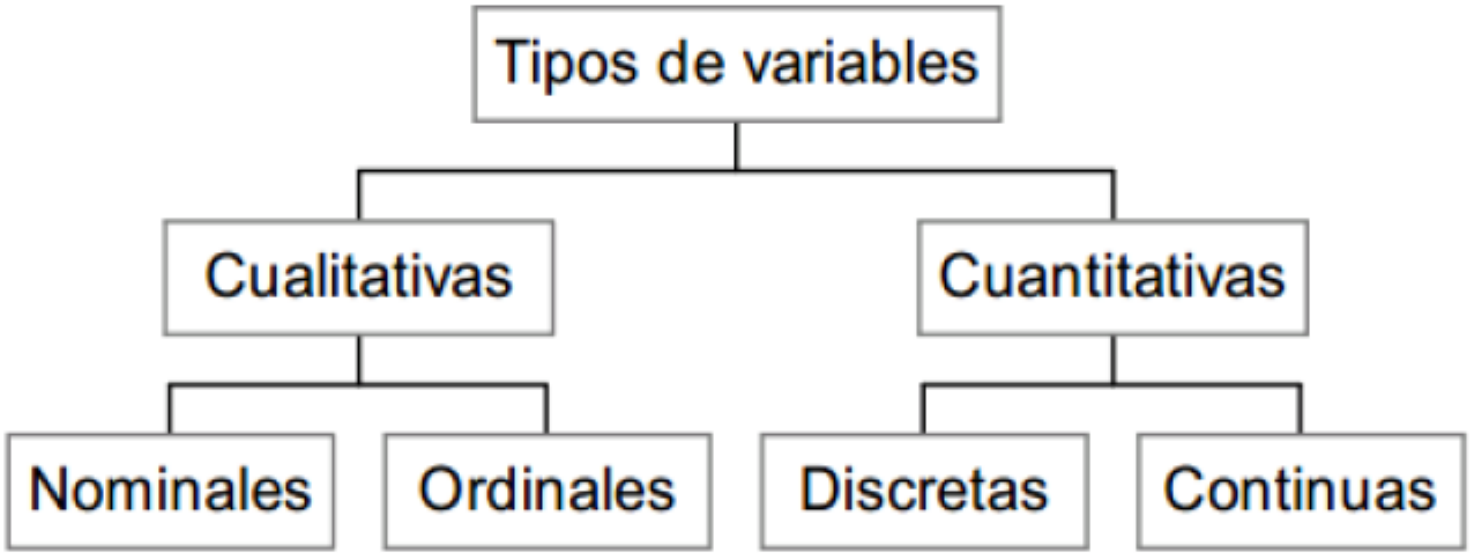
OJO: hay números que son etiquetas y por tanto, no son susceptibles de ser operados matemáticamente. Ejemplo, código postal, N° de cédula, Carné Estudiantil, número de abonado, número de factura, teléfono, etc.

Variables Cualitativas

ind	SEXO	EDAD	INGRESO
1	F	5	Medio
2	F	3	Alto
3	M	4	Bajo
4	F	1	Bajo
5	M	2	Medio
6	M	5	Alto
7	F	2	Medio
8	M	3	Bajo
9	M	1	Alto
10	F	4	Medio

Tabla 1: ejemplo de una tabla de datos de tres variables cualitativas observadas sobre 10 individuos.

Tipos de Variables



Tipo de variable cuantitativa: discreta o continua

- La variables discreta es aquella en la cuál se puede contar el número posible de valores.
- Ejemplos: número naturales, números enteros, algunas series…
- Tienen la particularidad de que es posible determinar para cada número, su antecesor y su sucesor.
- No admite la propiedad de densidad

Tipo de variable cuantitativa: discreta o continua

- La variables continua puede tomar cualquier valor en un intervalo dado.
- Ejemplos: número reales.
- Admite la propiedad de densidad

Tipo de variable cualitativas: nominales u ordinales

- Los valores de las variables ordinales, como su nombre lo indica, se rigen bajo un orden.
- No se puede realizar operaciones matemáticas con ellas
- Ejemplo: grado académico, categoría profesional, nivel de pobreza

Tipo de variable cualitativas: nominales u ordinales

- Los valores de las variables nominales, son categorías y no siguen un orden, no se puede decir quien es mayor o menor .
- No se puede realizar operaciones matemáticas con ellas
- Ejemplo: estado civil, sexo, región donde vive, partido político.

¿Se puede transformar una variable cuantitativa en una cualitativa?

- Se ordena la variable
- Se define el número de categorías que se desea.
- Se toma el rango para los datos y se divide entre el número de categorías para definir el tamaño de la clase
- Se establecen las clases y se le asigna una etiqueta a cada una de ellas
- Ejemplo: si se consideran el ingreso salarial, podría hacerse una categoría de clase baja, media o alta

Los datos…

Se parte de una tabla de datos:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix} \hookleftarrow \text{individuo } i$$

Variable j

Ejemplo de tabla de datos:

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

Ejemplo de tabla de datos:

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

Individuo o vector

Ejemplo de tabla de datos:

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

Variable, columna o vector

Nube de puntos…

INDIVIDUOS - FILAS

Luis	5.0	6.5	6.5	7.0	9.0

$\in \mathbb{R}^5$

VARIABLES - COLUMNAS

Español
9.2
7.3
8.0
6.5
7.8
7.7
8.2
7.5
6.5
8.7

$\in \mathbb{R}^{10}$

Tablas para Métodos Exploratorios…

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

Tablas para Métodos Predictivos…

	Matemáticas	Ciencias	Español	Historia	EdFísica	Tipo
Lucía	7.0	6.5	9.2	8.6	8.0	Regular
Pedro	7.5	9.4	7.3	7.0	7.0	Bueno
Inés	7.6	9.2	8.0	8.0	7.5	Bueno
Luis	5.0	6.5	6.5	7.0	9.0	Malo
Andrés	6.0	6.0	7.8	8.9	7.3	Regular
Ana	7.8	9.6	7.7	8.0	6.5	Bueno
Carlos	6.3	6.4	8.2	9.0	7.2	Regular
José	7.9	9.7	7.5	8.0	6.0	Bueno
Sonia	6.0	6.0	6.5	5.5	8.7	Regular
María	6.8	7.2	8.7	9.0	7.0	Malo

Variable Discriminante

Id	Reembolso	Estado Civil	Ingresos Anuales	Fraude
1	Sí	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Sí	Casado	120K	No
5	No	Divorciado	95K	Sí
6	No	Casado	60K	No

Tabla de Aprendizaje

Id	Reembolso	Estado Civil	Ingresos Anuales	Fraude
7	No	Soltero	80K	No
8	Si	Casado	100K	No
9	No	Soltero	70K	No

Tabla de Testing

Variable Discriminante

Descripción de una variable cuantitativa

- Una variable cuantitativa está definida para los valores que toma en el conjunto de los "n" individuos para los cuáles fue definida.

Ejemplo 1

individuo	tamaño
1	1.70
2	1.65
3	1.70
4	1.80

Descripción de una variable cuantitativa

- Para poder describir a esta variables, se utilizan algunos indicadores especiales conocidos también como parámetros (poblacional) o Estadísticos (muestral)

- Promedio:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Descripción de una variable cuantitativa

- Varianza: describe que tan lejos están los datos con respecto a su promedio.

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

- Desviación estándar: Equivale a la raíz cuadrada de la varianza

$$s = \sqrt{var(x)}$$

Descripción para la relación entre dos variables cuantitativas

- Coeficiente de Determinación: Me indica cuanto de la variación de la variable dependiente, se ve explicada o se debe a la variación de la variable independiente. Se simboliza con r^2

Ejemplo:

r^2 Si tenemos una relación entre estatura vs la edad de la persona con un =0.85 significa que, el 85% de la variación de la estatura (Variable dependiente o de respuesta) de una persona se debe a o se explica por la variación de su edad (variable independiente o predictora). El restante 15% se debe a otros factores.

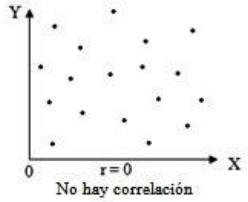
Descripción para la relación entre dos variables cuantitativas

- Coeficiente de correlación: Si se intenta modelar mediante una recta (ajuste lineal) la relación entre dos variables, el coeficiente de correlación lo que me indica es qué tan dispersos están los datos con respecto a la recta de ajuste, es decir, me indica la calidad del modelo. Se simboliza con r

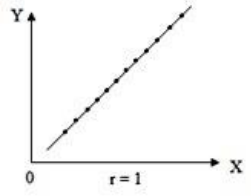
Ejemplo:

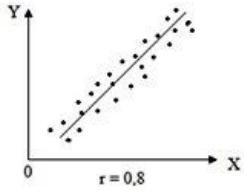
Si tenemos una relación entre estatura vs la edad de la persona con un r=0.85 significa que aproximadamente el 85% de los puntos están sobre la recta y al ser positivo, indica que la relación lineal es creciente.

Interpretación de r...

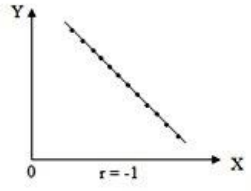


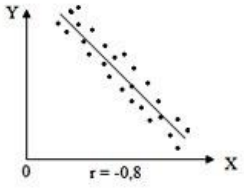
No hay correlación





Correlación Positiva





Correlación Negativa

Interpretación del coeficiente de correlación de Pearson

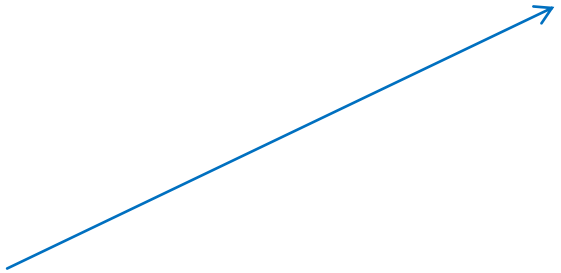
Interpretación geométrica del coeficiente de correlación…

- Una variable X, que toma "n" valores puede ser representada como un vector en $\mathbb{R}^n$, a lo que se le conoce como espacio vectorial de las variables.
- RECORDEMOS que en nuestras tablas de datos, cada columna equivale a una variable, es decir, cada columna es un vector.

VARIABLES - COLUMNAS

Español
9.2
7.3
8.0
6.5
7.8
7.7
8.2
7.5
6.5
8.7

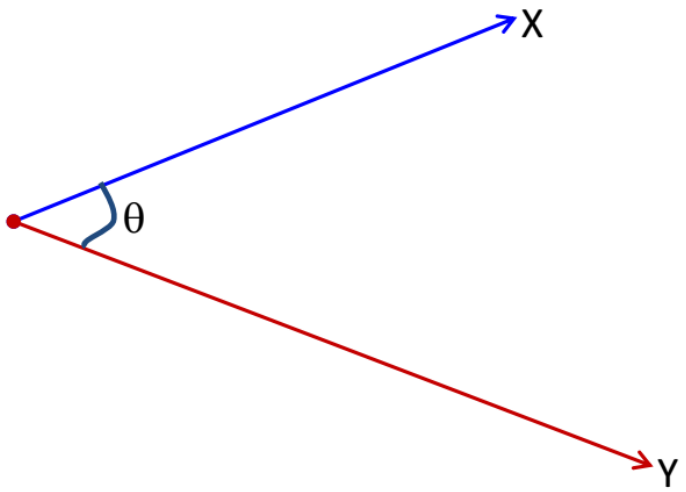
$\in \mathbb{R}^{10}$



Interpretación geométrica del coeficiente de correlación…

- Dados dos vectores, entonces, existe un ángulo entre ellos, en este caso lo llamaremos θ

> **Teorema 9** En el espacio vectorial de las variables $\mathbb{R}^n$ el coseno del ángulo entre 2 variables centradas y reducidas es igual al coeficiente de correlación entre esas 2 variables.

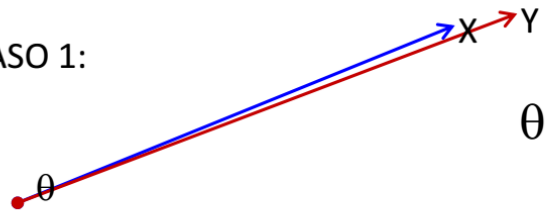


$$\text{Cos}(\theta) = R(X,Y)$$

$$\cos(\theta) = \frac{\mu \cdot v}{\|\mu\|\|v\|}$$

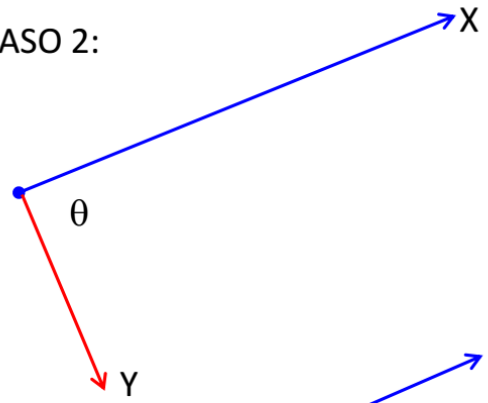
Interpretación de los cosenos…

CASO 1:



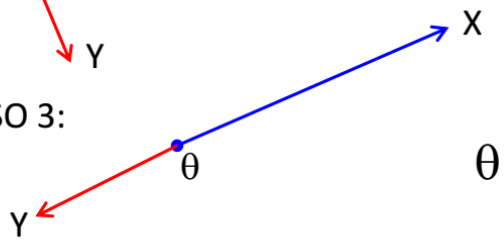
$\theta = 0°$ implica que $Cos(\theta) = R(X,Y) = 1$

CASO 2:



$\theta = 90°$ implica que $Cos(\theta) = R(X,Y) = 0$

CASO 3:



$\theta = 180°$ implica que $Cos(\theta) = R(X,Y) = -1$