

# Procesamiento de Señales II

## Ciencia de Datos II

### **Análisis y Visualización de Datos - Probabilidad**

Capítulo [1]

**Dr. José Ramón Iglesias**

DSP-ASIC BUILDER GROUP

Director Semillero TRIAC

Ingeniería Electronica

Universidad Popular del Cesar

Primero: ¿cuál es el problema?

Si me dedico a la programación...  
¿Cuánto puedo cobrar?...¿ Se podrá  
implementar un sistema que, dadas las  
características de una persona, devuelva el  
sueldo posible?

# Encuesta Sysarmy

- Encuesta personal y voluntaria que busca relevar información sobre salarios y condiciones de trabajo de programadores, que se realiza anualmente.
- Usaremos sólo los datos provenientes de un país como Argentina

# Demo con Notebook

CIED2\_Probabilidad.ipynb

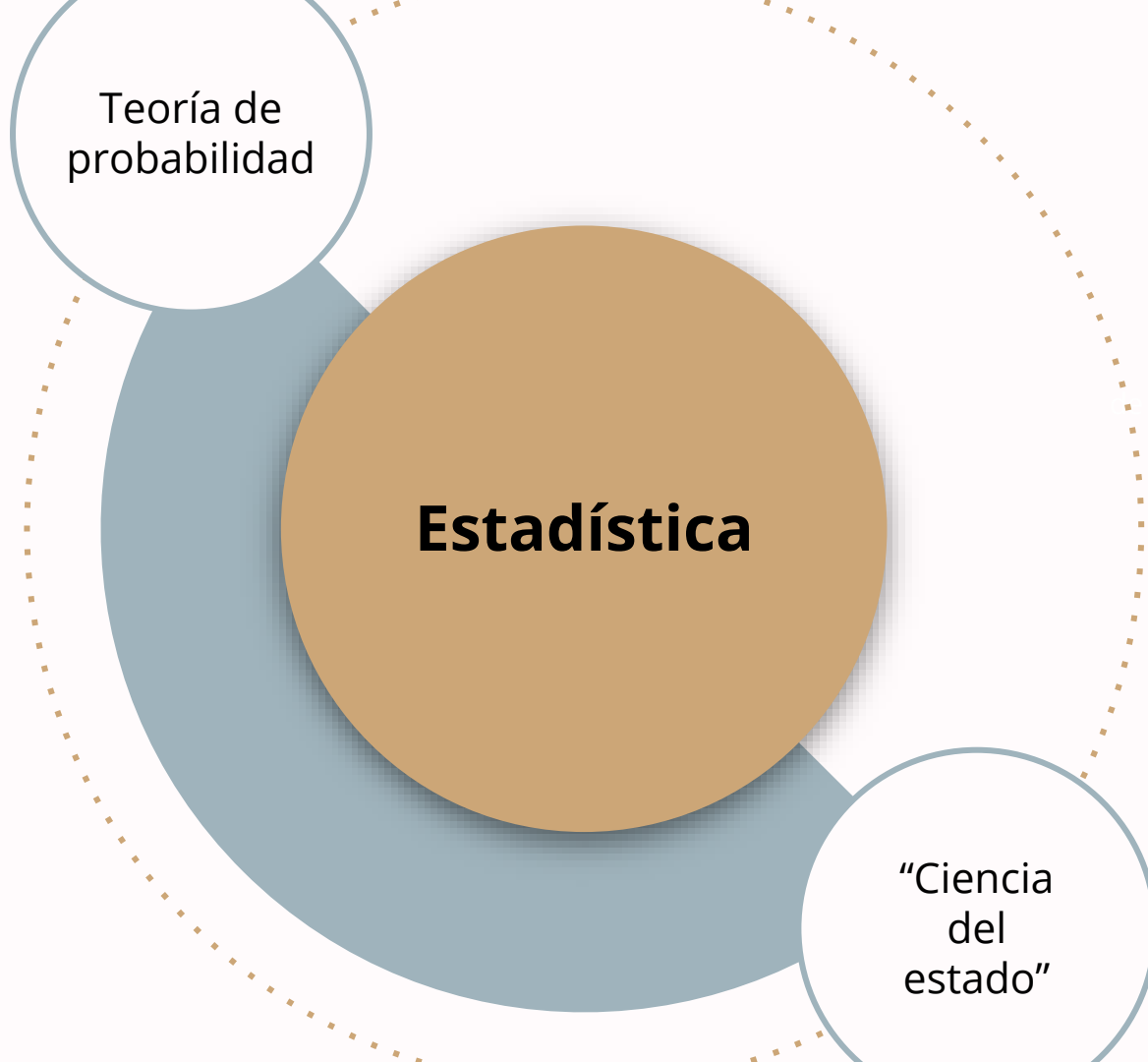
Teoría del azar

Teoría de  
probabilidad

**Estadística**

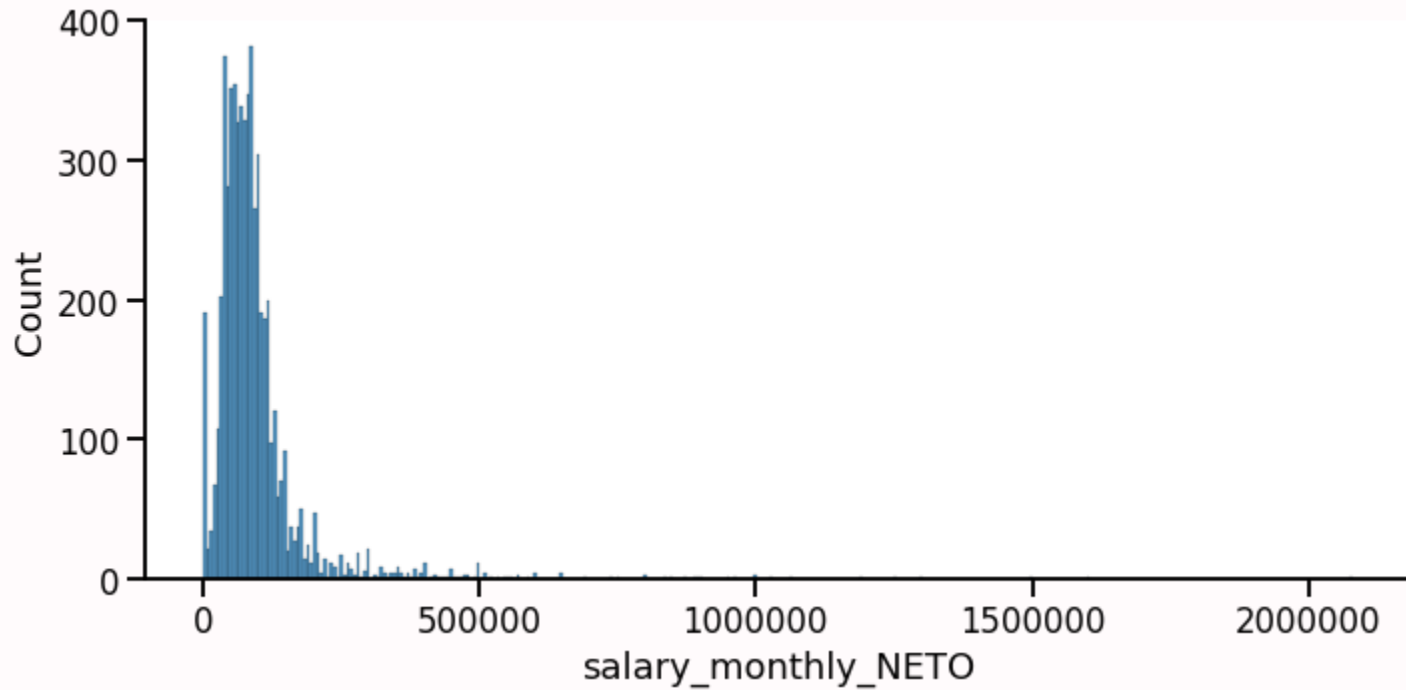
"Ciencia  
del  
estado"

Recolección y uso  
de datos en el  
gobierno de un  
estado



# Utilidad de la Estadística

- **Descripción de datos**
- **Análisis de muestras**
- Medición de relaciones
- Toma de decisiones
- Contrastación de Hipótesis
- Inferencia
- Predicción



¿Cuál es el concepto matemático que usamos para modelar la columna salary\_monthly\_NETO?

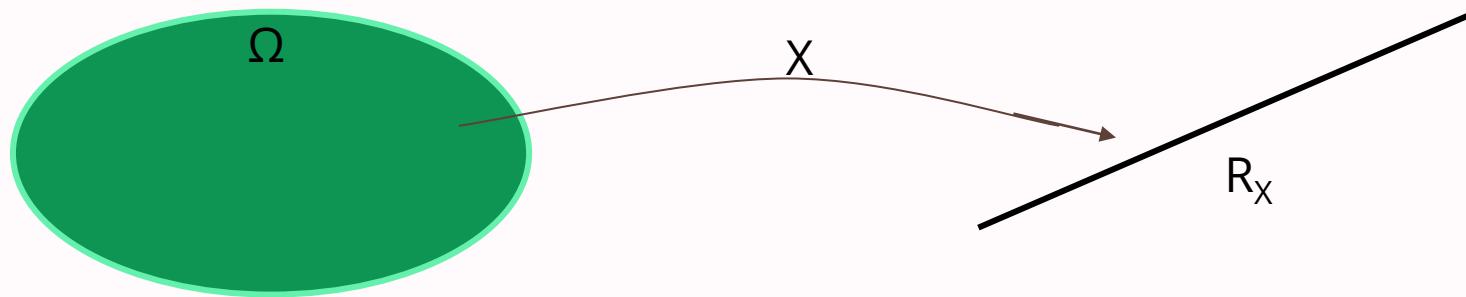


# Variable Aleatoria

Una **variable aleatoria (v.a.)**  $X$  es una función

$$X: \Omega \rightarrow R_X$$

donde  $\Omega$  es un conjunto llamado **Espacio de estados** y  $R_X$  es un conjunto de valores que toma la variable llamado **Rango**.



Una **variable aleatoria**

(v.a.)  $X$  es una función

$$X: \Omega \rightarrow R_X$$

donde  $\Omega$  es un conjunto

llamado **Espacio de**

**estados** y  $R_X$  es un

conjunto de valores que

toma la variable llamado

**Rango.**

Podemos utilizar una v.a.  $X$  que

represente la columna

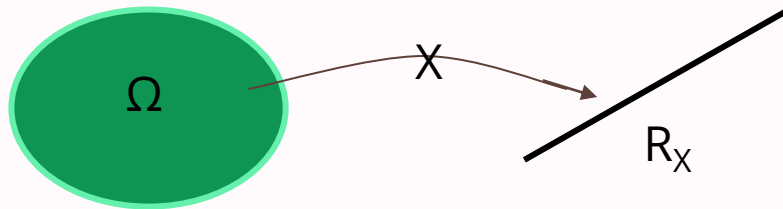
salary\_monthly\_NETO,

En este caso  $\Omega$  puede ser la población

total de programadores en un país en

2020 (o bien los que hicieron la encuesta)

y  $R_X$  el conjunto de salarios posibles.



Una **variable aleatoria**

(v.a.)  $X$  es una función

$$X: \Omega \rightarrow R_X$$

donde  $\Omega$  es un conjunto

llamado **Espacio de**

**estados** y  $R_X$  es un

conjunto de valores que

toma la variable llamado

**Rango.**

El espacio de estados  $\Omega$  es el conjunto de estados (personas) que podríamos haber encontrado en nuestra encuesta.

$\Omega = \{\omega / \omega \text{ es una persona viva que trabaja en un país}\}$

Puede tener más de una definición:

$\Omega = \{\omega / \omega \text{ es una persona viva que trabaja en un país como desarrollador/a}\}$

Una **variable aleatoria**

(v.a.)  $X$  es una función

$$X: \Omega \rightarrow R_x$$

donde  $\Omega$  es un conjunto

llamado **Espacio de**

**estados** y  $R_x$  es un

conjunto de valores que

toma la variable llamado

**Rango.**

El rango  $R_x$  es el conjunto de valores posibles de salary\_monthly\_NETO.

$R_x = \mathbb{R}$  ? (conjunto de números reales)

$R_x = \mathbb{N}$  ? (conjunto de números naturales)

¿Cómo podemos calcular el rango de  $R_x$  en la encuesta?

Una **variable aleatoria**

(v.a.)  $X$  es una función

$$X: \Omega \rightarrow R_X$$

donde  $\Omega$  es un conjunto

llamado **Espacio de**

**estados** y  $R_X$  es un

conjunto de valores que

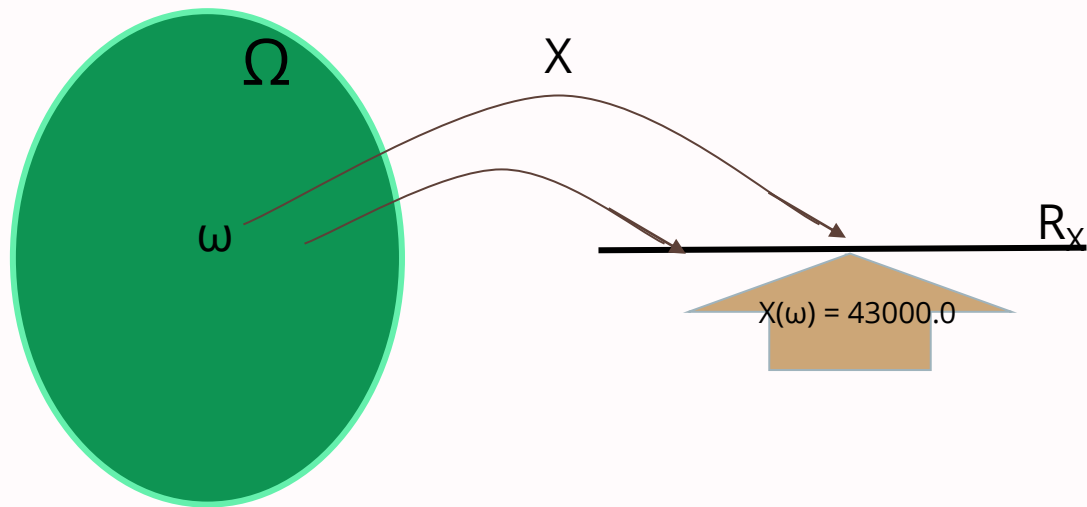
toma la variable llamado

**Rango.**

$\omega$  = Persona que respondió primero

$$X(\omega) = 43000.0$$

$X(\omega)$  se denomina **realización** de la v.a.  $X$



# Variable Aleatoria - Otros ejemplos

$X$	$\Omega$ (espacio de estados posibles, universo, población que vamos a estudiar,...)	$R_X$
horas diarias que trabaja	personas que son programadores ...	1 - 24
cantidad de glóbulos rojos en sangre	personas	valores en números enteros positivos.
nivel de estudio	asistentes al curso de Ciencia de Datos 2022-2	{estud., grado, posgrado, estud crónico}
altura al nivel del mar	globo terráqueo	$\mathbb{R}$

# Tipos de variables aleatorias

Las variables aleatorias pueden ser de distinto tipo, de acuerdo a los valores presentes en el Rango y su interpretación.

- Numéricas
  - Continuas
  - Discretas (un conjunto finito o infinito numerable de valores posibles)
- Categóricas
- Ordinales

Determinar los tipos de datos/variable  
que estamos usando nos permite  
seleccionar las herramientas adecuadas  
para obtener información a partir de ellos



# Demo con Notebook

[CIED2\\_Probabilidad.ipynb](#)

Hagamos una pregunta interesante:  
¿Tener más años de experiencia  
significa que se cobra más?

# ¿Cómo hacer este análisis?

Plantear una hipótesis

Si no planteamos una hipótesis primero, es difícil determinar qué pasos hay que seguir para poder hacer el análisis

Identificar las variables

Una vez que la hipótesis está definida, hay que determinar QUÉ hay que medir para poder comprobarla.

Diseñar el experimento

Una vez que está definido qué medir, se seleccionan las herramientas para medirlo.

# ¿Cómo hacer este análisis?

Plantear una hipótesis

Tener más años de  
experiencia significa  
que se cobra más

Identificar las  
variables

salary\_monthly\_NETO  
profile\_years\_experience

Diseñar el  
experimento

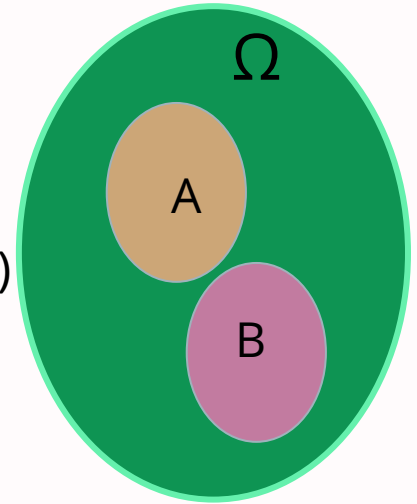
????

# Teoría de probabilidad

# ¿Probabilidad? - Interpretación axiomática

**P** es una **medida de Probabilidad** en el **espacio  $\Omega$**  si para cada subconjunto **A** de  **$\Omega$** , **P(A)** es un número tal que:

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- $P(A \cup B) = P(A) + P(B)$ , para A y B disjuntos (o excluyente)
- $P(\cup_i A_i) = \sum_i P(A_i)$  para  $A_1, A_2, \dots$  disjuntos



# ¿Cómo se calcula?

Si  $\Omega$  tiene  $k$  elementos equiprobables (i.e. si  $\omega_i$  es un elemento de  $\Omega$ ,  $P(\{\omega_i\}) = 1/k$ )

Si el conjunto  $A$  son los elementos en los que el fenómeno ocurre.

Entonces la probabilidad de un conjunto  $A \subset \Omega$  es la proporción de eventos en  $A$ .

$$P(\{\omega_i\}) = 1/k \implies |A|/k$$

# Situaciones más complejas

Si hay dos situaciones a estudiar, entonces se modela el problema usando las columnas salary\_monthly\_NETO y profile\_years\_experience para crear conjuntos de eventos y comprobar si existe una relación entre ellos.

Los conjuntos que se eligen son los que determinan el **experimento**

- $A = \{ \omega_i : \text{salary\_monthly\_NETO} > \text{avg}(\text{salary\_monthly\_NETO}) \}$
- $B = \{ \omega_i : \text{profile\_years\_experience} > 5 \}$



# Situaciones más complejas

$A = \{ \omega_i : \text{salary\_monthly\_NETO} > \text{avg} \}$

$B = \{ \omega_i : \text{profile\_years\_experience} > 5 \}$

intersección: A & B, A y B

La **probabilidad conjunta** de que ocurran ambos eventos al mismo tiempo se modela usando la intersección de los conjuntos:

$$P(A \cap B)$$

# Situaciones más complejas

$A = \{ \omega_i :$   
salary\_monthly\_NETO > avg  
 $\}$

$B = \{ \omega_i :$   
profile\_years\_experience > 5  
 $\}$

La **probabilidad condicional** de que el salario esté por encima del promedio, suponiendo que ocurre el evento de tener más de 5 años de experiencia, se calcula como:

$$P(B) \neq 0 \implies P(A|B) = \frac{P(A \cap B)}{P(B)}$$

# Situaciones más complejas

$A = \{ \omega_i :$   
salary\_monthly\_NETO > avg  
 $\}$

$B = \{ \omega_i :$   
profile\_years\_experience > 5  
 $\}$

A y B se dicen conjuntos **independientes** si

$$P(A \cap B) = P(A)P(B)$$

$$P(B) \neq 0 \implies P(A|B) = P(A)$$

¿Si uno tiene más de 5 años de experiencia, la probabilidad de cobrar más que el promedio aumenta? ¿Estos eventos, son independientes?

Ejercicio en la Ntb.

¿Son independientes o no?

# Teorema de Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Tiene muchas aplicaciones en la ciencia de datos, incluyendo el aprendizaje bayesiano, pero no profundizamos en este tema porque lo van a ver con mucho más detalle en materias siguientes, cuando vean el clasificador Naive Bayes.