Análisis y Visualización de Datos

Diplomatura CDAAyA 2024

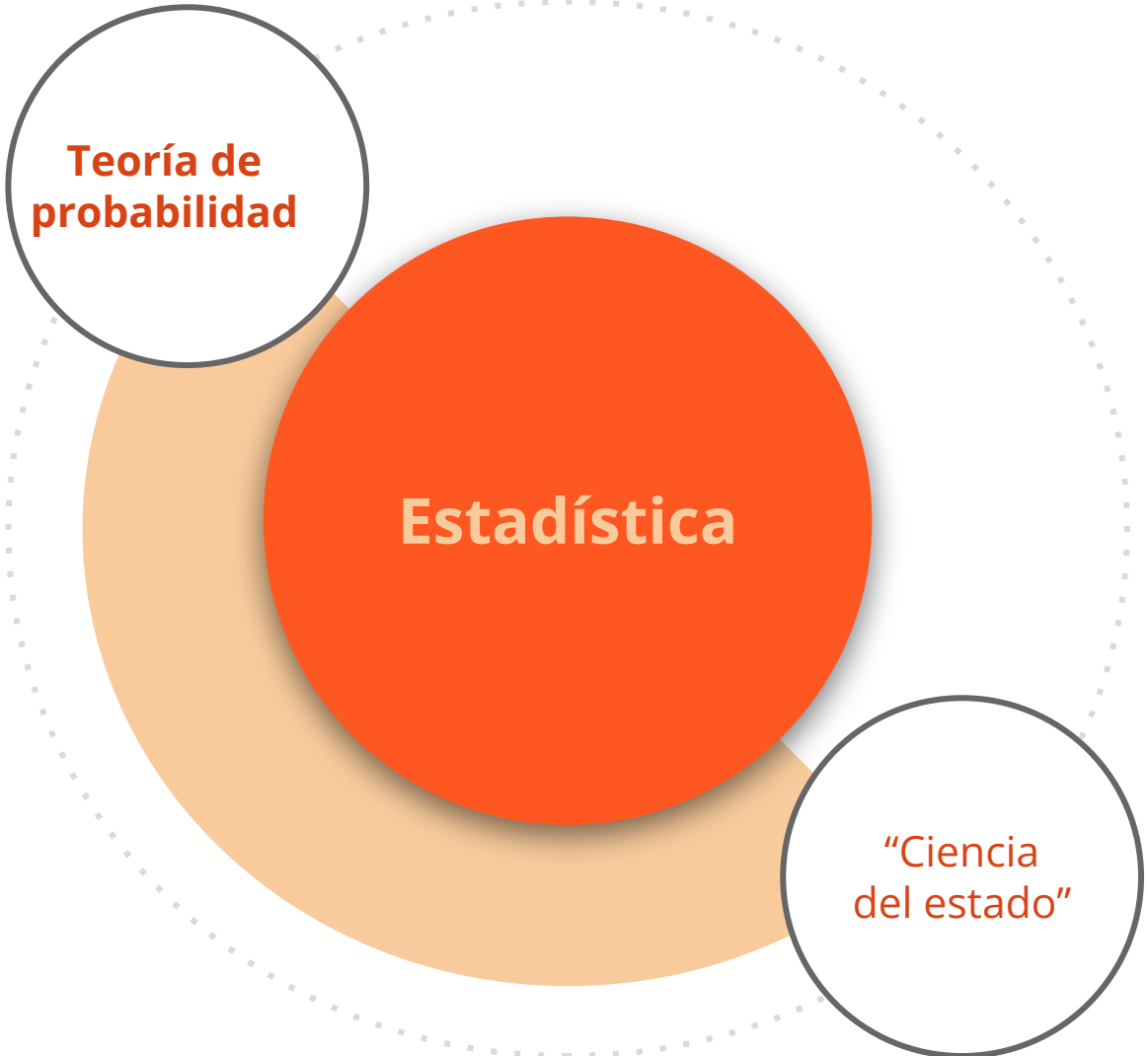
Primero: ¿cuál es el problema?

Si me dedico a la programación...
¿Cuánto puedo cobrar?

¿ Se podrá implementar un sistema que, dadas las características de una persona, devuelva el sueldo más probable?

Encuesta Sysarmy

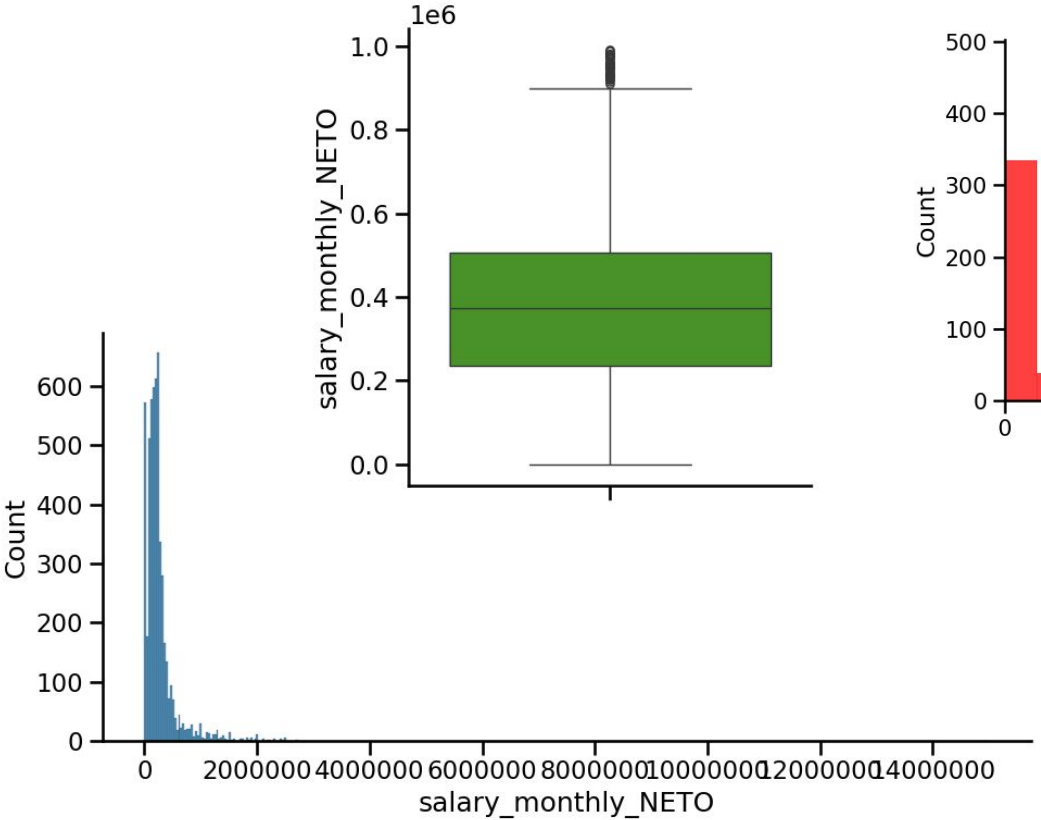
- Encuesta personal y voluntaria que busca relevar información sobre salarios y condiciones de trabajo de programadores, que se realiza anualmente.
- Usaremos sólo los datos provenientes de Argentina 2023
- Link a la página de los datos

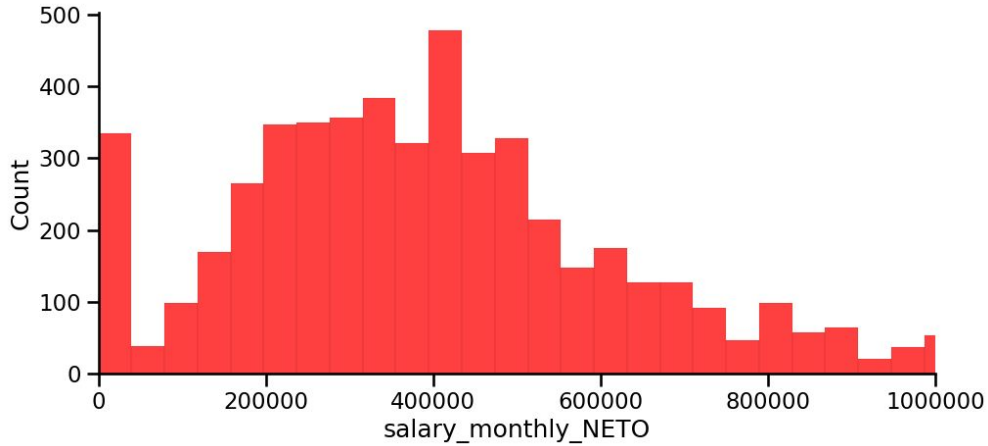


Recolección y uso de datos en el gobierno de un estado

Utilidad de la Estadística

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





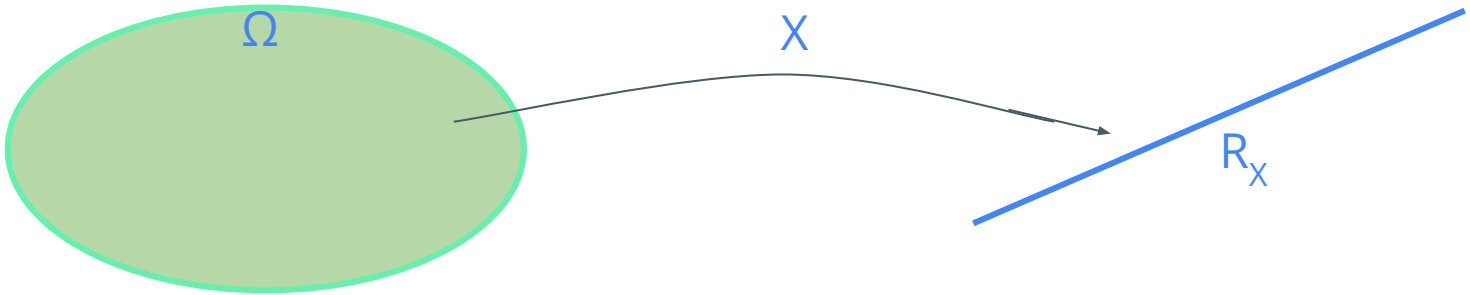
¿Cuál es el concepto matemático que usamos para modelar la columna del sueldo?

Variable Aleatoria

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

$$X: \Omega \rightarrow R_X$$

donde Ω es el universo (de posibilidades) y R_X es un conjunto de valores que toma la variable.



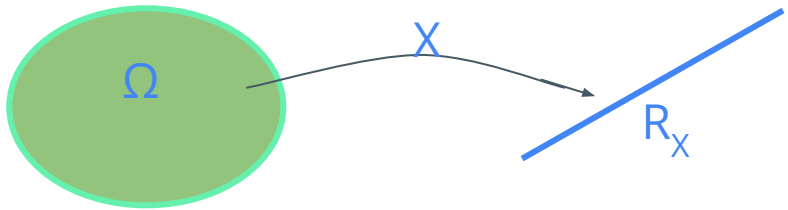
Una **variable aleatoria (v.a.)** X es una función

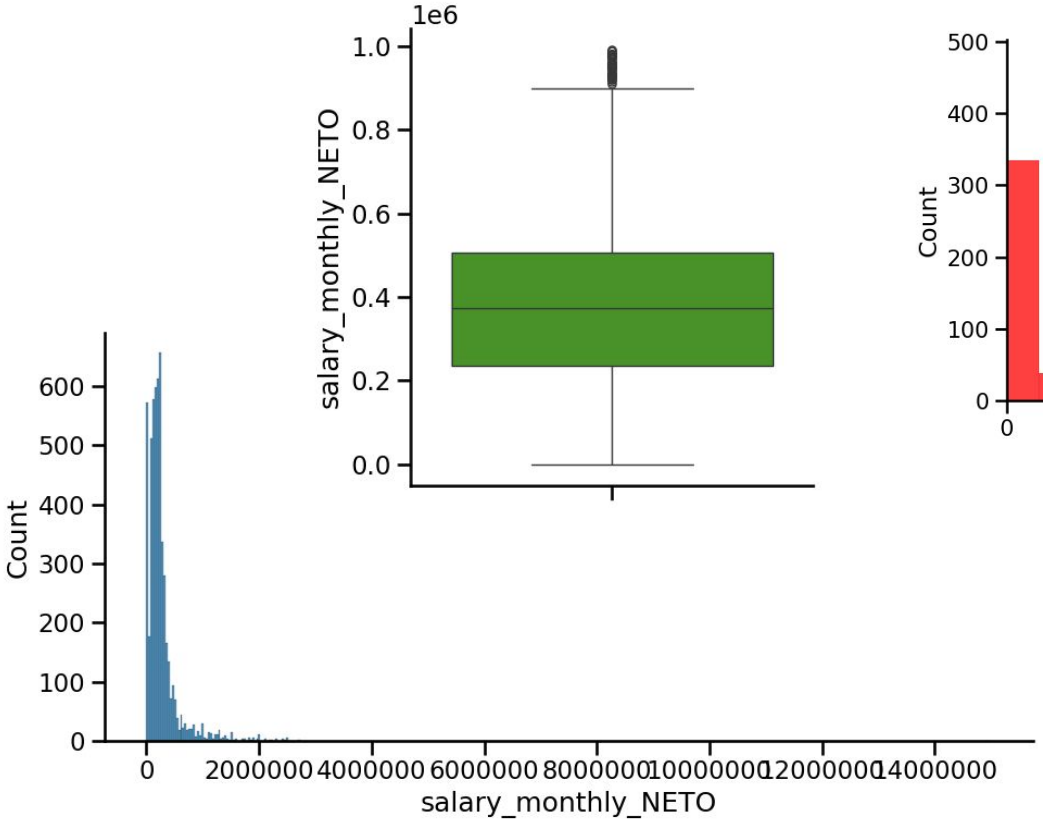
$$X: \Omega \rightarrow R_X$$

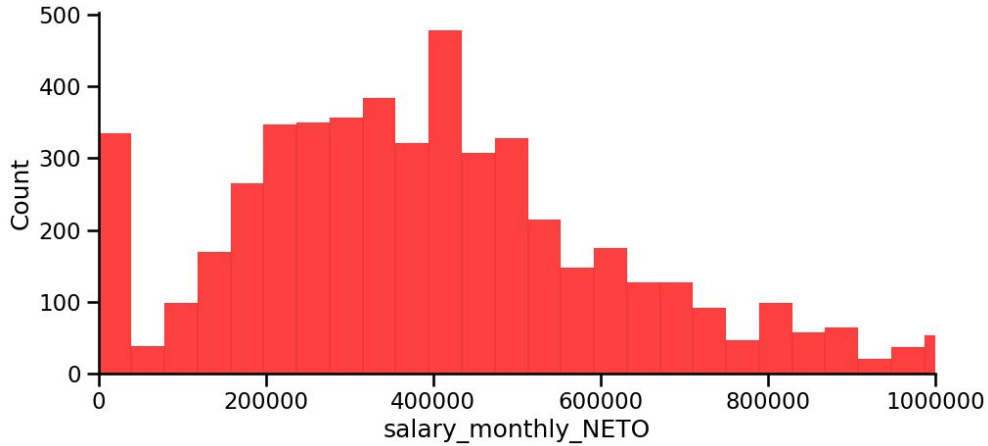
- Ω es el universo
- R_X conjunto de valores que puede tomar la variable.

Consideremos una v.a. X que modele la columna del sueldo,

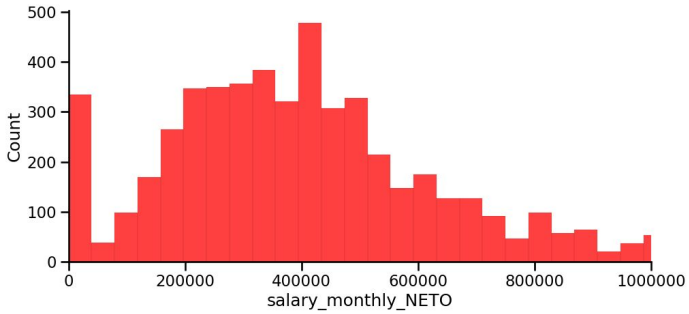
En este caso Ω puede ser la población total de programadores en Argentina en 2023 y R_X el conjunto de salarios posibles.







Variable salary_monthly_NETO



X v.a.

$X: \Omega \rightarrow R_X$

- Ω es el universo
- R_X (Rango) conj. de valores que puede tomar la variable .

El Ω es el conjunto de estados posibles, concepto a veces ideal, universo (de posibilidades), puede pensarse como la población que vamos a estudiar.

Por ej.

Ω = {ω / ω es una persona viva que trabaja en Argentina}

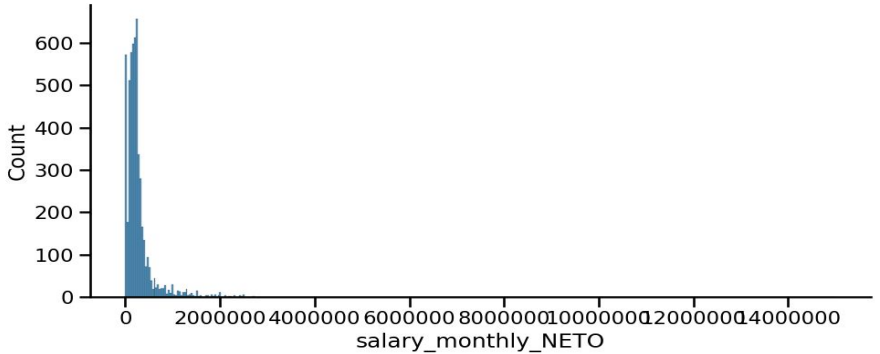
Puede tener más de una definición:

Ω = {ω / ω es una persona viva que trabaja en Argentina como desarrollador/a}

X v.a.

$X: \Omega \rightarrow R_X$

- Ω es el universo
- R_X (Rango) conj. de valores que puede tomar la variable .



El rango R_x es el conjunto de valores posibles del sueldo.

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

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

¿Cómo podemos calcular el rango R_x en la encuesta?

X v.a. salario mensual

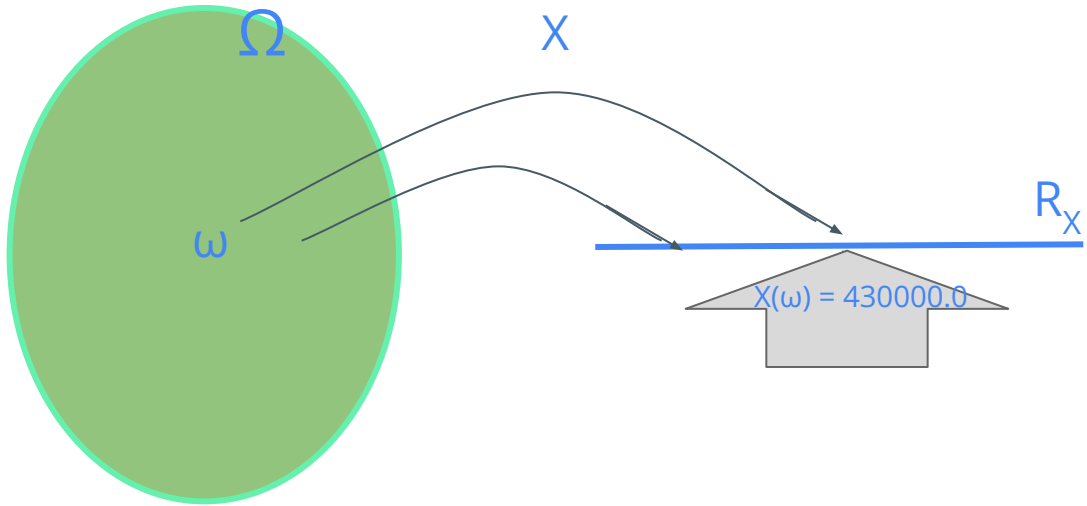
$X: \Omega \rightarrow R_X$

- Ω es el universo
- R_X (Rango) conj. de valores que puede tomar la variable .

ω = desarrollador

$X(\omega) = 430000.0$

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



Variable Aleatoria - Otros ejemplos simples

Ω (espacio de estados, universo (de posibilidades), población que vamos a estudiar)	X	R_X
personas QUE VIVEN EN ARGENTINA	horas diarias que trabaja	[0, 24)
personas QUE VIVEN EN ARGENTINA	edades	0… 130
superficie del globo terráqueo	altura al nivel del mar	[-4000, 4000] o R
personas	provincia	Córdoba, ….

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

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

Demo con Notebook

01 Probabilidad.ipynb

Sección A

- Datos
- Distribución de los datos

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

¿Cómo hacer este análisis?

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

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

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

¿Cómo hacer este análisis?

Tener más años de experiencia significa que se cobra más

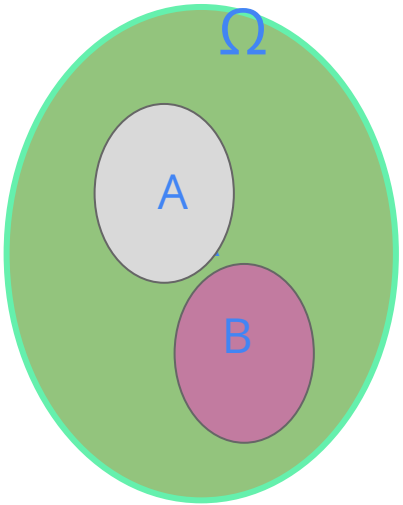
????

Teoría de probabilidad

¿Probabilidad? - Interpretación axiomática

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

- **0** ≤ **P**(A) ≤ 1
- **P**(**Ω**) = 1
- **P**(A U B) = **P**(A) + **P**(B), para A y B disjuntos (o excluyente)
- **P**(U_i A_i) = $\sum_i$ **P**(A_i) para A_1, A_2,... disjuntos



¿Cómo se calcula?

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

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 A en **Ω**.

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

Situaciones más complejas

Si hay dos características a estudiar, entonces se modela el problema usando dos v.a. (columnas) por ej 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 = { ω_i : salary_monthly_NETO > avg }

B = { ω_i : 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 = { ω_i : salary_monthly_NETO > avg }

B = { ω_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 = { ω_i : salary_monthly_NETO > avg }

B = { ω_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?

¿Si uno ... (completar), la probabilidad de cobrar más que el promedio aumenta? ¿Estos eventos, son independientes?

Ejercicio en la Ntb.

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.

Población y muestra

Cuando recogemos los datos muchas veces es imposible relevar la característica de interés de todo el grupo entero (población) o universo, se examina una pequeña parte del grupo, llamada muestra.

Se denotan los n datos de una muestra: $x_1,...x_n$ (observaciones/realizaciones de la v.a. X)

Muestras sesgadas?

Al estimar la medida de probabilidad como una proporción, estamos asumiendo una muestra representativa del campo de aplicación.

El proceso de selección de los eventos para un experimento determina las características de la muestra obtenida.

- Muestras convenientes (los que “estaban a mano”)
- Muestras de respuestas voluntarias

Población

El grupo completo de estados Ω
que se busca estudiar

¿Cuál es nuestra población?

Muestra

Un subconjunto de Ω elegido
para un experimento particular

¿Cuál es nuestra muestra?

Muestras sesgadas

¿Qué sesgos tenemos en esta muestra?

- ...

¿Influyen en nuestra característica de estudio (el salario)?

¿Cómo hacer un análisis?

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

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

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

¿Cómo hacer un análisis?

Tener más años de experiencia significa que se cobra más

????