Para estudiantes de Educación Básica y Media
UNIVERSIDAD DE CHILE

# Clase 3: Estadística descriptiva

FM849 - Programación Científica para Proyectos de Inteligencia Artificial (IA)

6 de enero de 2026

# Motivación

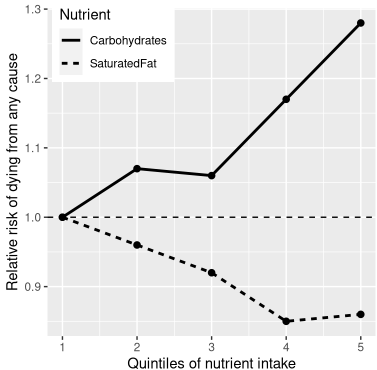
La estadística nos ayuda con las siguientes tareas:

- Describir el comportamiento de fenómenos complejos utilizando conceptos sencillos (p. ej., el promedio).
- Capturar aspectos esenciales de los datos como su estructura y saber qué tan incierto es el comportamiento generado en éstos.
- Tomar decisiones mediante el análisis de datos, identificando tendencias y variabilidad en éstos.
- Predecir resultados futuros a partir de datos pasados, sirviendo como base para modelos de aprendizaje automático.

En inteligencia artificial, los datos y su tratamiento son una parte fundamental para construir modelos precisos. La estadística entrega herramientas y métodos para analizar y generar conocimiento sobre estos datos.

# Ejemplo: grasas saturadas y carbohidratos

En el estudio "PURE", realizado con $\approx$ 135.000 personas, se evaluó el riesgo de muerte por cualquier causa según el consumo de carbohidratos y grasas saturadas.




- Los resultados muestran que un mayor consumo de carbohidratos está ligado a una mayor tasa de muerte.
- Lo inverso ocurre para las grasas saturadas.

Considerando el gráfico, ¿podemos asegurar que el consumo de grasas saturadas reduce la tasa de muerte?

No, porque la existencia de correlaciones no implica que haya causalidad.

# Un resumen de los principios de la estadística

La estadística se basa en los siguientes principios.

- **Aprendizaje de los datos**. La estadística es un conjunto de herramientas para postular hipótesis sobre los datos.
- **Agregación**. En estadística, resumimos un conjunto de datos completo calculando valores que los describen.
- **Incertidumbre**. En el mundo estadístico existen variables exógenas que no podemos medir o controlar.
- **Muestreo**. Queremos sacar conclusiones sobre una población sacando datos de una muestra.
- **Correlación y causalidad**. La correlación entre dos variables no implica causalidad.

# ¿Cómo aprendemos de los datos?

El análisis exploratorio de datos (en inglés, abreviado EDA) es un conjunto de técnicas que ayudan a entender los datos que se están trabajando.

- El objetivo central es intentar encontrar patrones en los datos que nos permitan postular hipótesis sobre ellos y/o el proyecto que contextualiza su análisis (p. ej., evaluación de factibilidad).
- Puede ser dividido en descripción (cálculos estadísticos) y visualización de los datos.

En esta clase, hablaremos sobre estadística descriptiva. Mañana ahondaremos en la visualización de información.

# Medidas de tendencia central: media

Las medidas de tendencia central intentan resumir datos de un conjunto en un valor único. La primera que veremos es la media o promedio. Dado un conjunto de datos $X = \{x_1, x_2, \ldots, x_n\}$, ésta se calcula como:

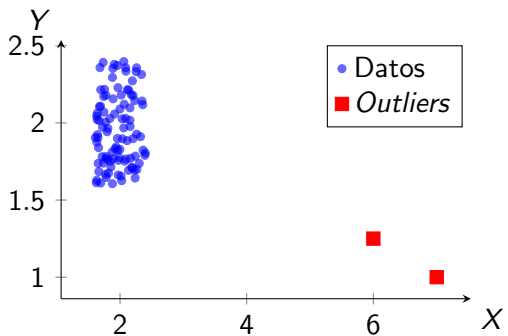
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

El problema del promedio es que es muy sensitivo a *outliers* (valores atípicos).

```
>>> import numpy as np
>>> sample = [1, 2, 3, 100]
>>> np.mean(sample)
```

# Hay que tener cuidado con los valores atípicos (*outliers*)

Los *outliers* son valores atípicos en un conjunto de datos. Éstos valores se encuentran alejados de la distribución de los datos, y se pueden presentar en una variable o en interacciones entre varias variables.




- Pueden deberse a errores de medición.
- Pueden representar eventos reales pero poco frecuentes.
- No deben ser eliminados ni imputados sin un previo análisis de por qué están presentes.
- Tienen un fuerte impacto en medidas como la media.

# Medidas de tendencia central: mediana

La mediana es el valor en la posición central de un conjunto de datos ordenado de menor a mayor. Para un conjunto de datos **ordenado** $X = \{x_1, x_2, \ldots, x_n\}$, se calcula como:

$$\text{mediana}(X) = \begin{cases} x_{(n+1)/2} & \text{si } n \text{ es impar} \\ \frac{x_{n/2} + x_{n/2+1}}{2} & \text{si } n \text{ es par} \end{cases}$$

En simple, es la observación que separa el conjunto de datos en una mitad de valores menores que él y otra mitad de valores mayores a él.

```
>>> import numpy as np
>>> sample = [1, 2, 33, 100, 4]
>>> np.median(sample)
```

# Medidas de tendencia central: moda

La moda es el valor que aparece con mayor frecuencia en un conjunto de datos. Para un conjunto de datos $X = \{x_1, x_2, \ldots, x_n\}$, se define como:

$$\text{moda}(X) = \underset{x_i \in X}{\arg\max}\ f(x_i)$$

donde $f(x_i)$ es la frecuencia de aparición del valor $x_i$ en el conjunto $X$.

```
>>> from scipy import stats
>>> sample = [1, 2, 2, 3, 4]
>>> stats.mode(sample)
```

# Medidas de tendencia central: moda

La moda es especialmente útil para datos categóricos, donde la media y la mediana no son aplicables. Por ejemplo, en un conjunto de datos que representa colores favoritos, la moda sería el color que más personas prefieren.

- Un conjunto de datos puede tener más de una moda (bimodal, multimodal).
- La moda es menos sensible a valores atípicos en comparación con la media. ¿Se les ocurre por qué?

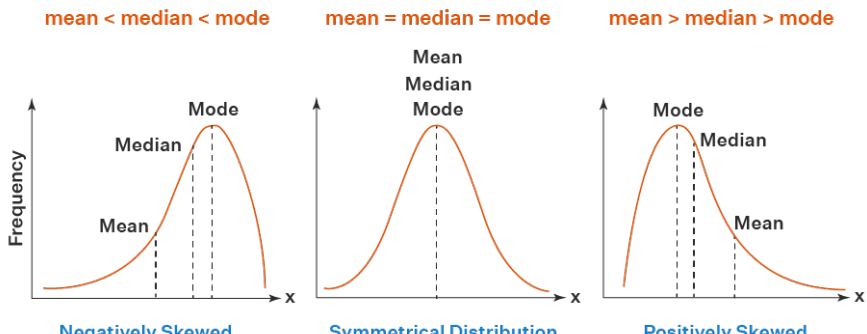
## Ejemplo

Consideremos el conjunto de datos simplificado que muestra las edades y alturas de tres personas. ¿Cuál es la moda de las alturas? ¿Por qué? ¿Y la moda de las edades?

| $X$ | Edad | Altura $(\mathrm{m})$ |
|---|---|---|
| $x_1$ | 17 | 1.76 |
| $x_2$ | 29 | 1.76 |
| $x_3$ | 10 | 1.58 |

# Comparación entre media, mediana y moda

Es interesante ver como cambian estas medidas de tendencia central según el tipo de distribución:

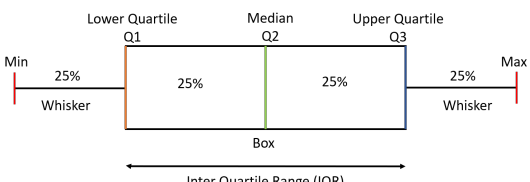


# Medidas de tendencia central (percentiles)

La mediana puede ser generalizada con el concepto de percentil. Los percentiles dividen los datos ordenados de menor a mayor en 100 partes iguales.

- El percentil 50 (denotado $P_{50}$) corresponde a la **mediana**.
- Los percentiles 25 ($P_{25}$) y 75 ($P_{75}$) definen el **rango intercuartílico**. También son conocidos como cuartil 1 ($Q_1$) y cuartil 3 ($Q_3$) respectivamente.
- $P_{25}, P_{50}$ y $P_{75}$ dividen la muestra en 4 partes que son descritas mediante los *boxplots*.

```
>>> import numpy as np
>>> sample = [1, 2, 3, 4, 100]
>>> np.percentile(sample, 25)
>>> np.percentile(sample, 75)
```



# La interpretación estadística **SÍ** importa

Consideremos un caso típico de análisis estadístico. Tomemos una muestra con 5 chilenos, cada uno representado por una entrada en $X = \{x_1, x_2, x_3, x_4, x_5\}$, y sus respectivos sueldos:

| $X$ | Sueldo (CLP) |
|---|---|
| $x_1$ | \$ 300 000 |
| $x_2$ | \$ 10 000 000 |
| $x_3$ | \$ 75 000 |
| $x_4$ | \$ 510 000 |
| $x_5$ | \$ 7 500 000 |

Tabla 1: Sueldo mensual de una muestra de personas chilenas.

- ¿Cuál es el valor del promedio $\overline{X}$? ¿Y de mediana$(X)$?
- ¿Qué pasa si sólo nos enfocamos en informar sobre el promedio?

# Medidas de dispersión

## Definición

Las medidas de dispersión se vinculan con estadísticos que permiten describir la variabilidad de los datos en una muestra con respecto a una medida de tendencia central (usualmente la media).

La figura 1 muestra la dispersión en un conjunto de puntos. **¿Cómo la resumimos en un número real?**

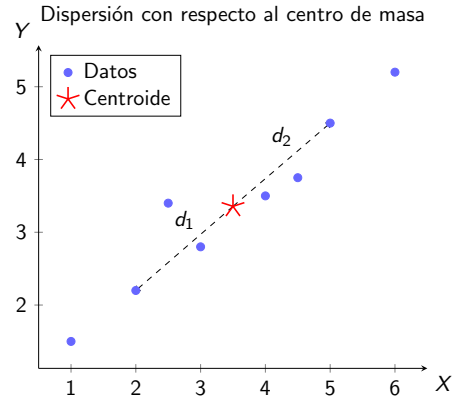



Figura 1: Dispersión de puntos en el espacio $\mathbb{R}^2$.

# Medidas de dispersión: varianza

La varianza mide qué tan dispersos están los datos con respecto a la media aritmética. Para un conjunto $X = \{x_1, x_2, \ldots, x_n\}$, se puede calcular como:

$$\mathbb{V}\text{ar}(X) = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{X}\right)^2$$

Lo que es análogo a un promedio de distancias considerando $n - 1$ grados de libertad.

```
>>> import numpy as np
>>> sample = [8, 3, 5, 10]
>>> np.var(sample, ddof=1) # ddof=1 permite dividir por n-1
```

## Pregunta

¿Cuál es la varianza de la muestra $X = \{1, 1, \ldots, 1\}$ (sólo unos)?

# Medidas de variabilidad: covarianza

Para cuantificar cuánto varia una variable con respecto a otra utilizamos metricas multivariadas.

La covarianza $\text{cov}(X, Y)$ mide el grado de cambio lineal en conjunto de dos variables $X$ e $Y$, y se calcula como:

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y}) \in (-\infty, +\infty)$$

Una covarianza positiva indica que ambas variables tienden a aumentar juntas, mientras que una covarianza negativa indica que una variable tiende a aumentar cuando la otra disminuye. Si es cero, no hay una relación lineal entre las variables, pero puede ser no lineal.

```
>>> import numpy as np
>>> X = [42, 9, 21, 31]
>>> Y = [3, 2, 1, 11]
>>> print(np.cov(X,Y))
```

# Medidas de variabilidad (covarianza y correlación)

- Si dos variables $X$ e $Y$ son independientes, entonces su covarianza será 0 (¡al revés esta afirmación no necesariamente es cierta!).
- Si la dependencia entre $X$ e $Y$ se puede aproximar mediante una recta, entonces diremos que existe una **correlación lineal** entre ellas.

Una métrica bastante importante es el coeficiente de correlación de Pearson $\rho(x, y)$, que estima la relación entre dos variables independiente de su magnitud, lo que permite que la interpretación sea más intuitiva:

$$\rho(x, y) = \frac{cov(x, y)}{sd(x)sd(y)} \in [-1, 1]$$

Si es 1, existe una correlación lineal positiva perfecta. Si es $-1$, existe una correlación lineal negativa perfecta. Si es 0, no existe correlación lineal entre las variables.