المحاضرة الثانية

# الإحصاء في Machine Learning

الدليل الشامل لفهم البيانات وتحليلها كمهندس ذكاء اصطناعي

# مقدمة: ما هي Statistics؟

## التعريف

هي علم جمع البيانات (Data)، تلخيصها، فهمها، ومن ثم استخلاص النتائج منها.

> **Statistics = Data Collection + Analysis + Interpretation**

## لماذا في AI/ML؟

- لفهم شكل البيانات من خلال **EDA**.
- لاكتشاف القيم الشاذة (**Outliers**) والالتواء (**Skewness**).
- للمقارنة بين النماذج واختيار الأفضل.
- لإجراء اختبارات **A/B Testing** على المنتجات.

# أنواع الإحصاء: Descriptive vs Inferential

## Descriptive Statistics

تهدف لوصف وتلخيص البيانات المتاحة فقط دون تعميم.

- Mean, Median, Mode
- Standard Deviation
- Histogram

## Inferential Statistics

تهدف لاستخلاص نتائج عن مجتمع (**Population**) بناءً على عينة.

- Hypothesis Testing
- Confidence Intervals
- Estimation


# Population vs Sample

## Population

كل العناصر التي نهتم بدراستها (مثل كل العملاء، كل الصور الممكنة).

## Sample

جزء صغير من الـ Population نجمع منه البيانات لتدريب النموذج.

**تحدي الـ AI Engineer:**
هل العينة (**Training Data**) تمثل المجتمع الحقيقي جيداً؟ انتبه من الـ **Sampling Bias**.

# مقاييس المركز: Means

## Mean

المتوسط الحسابي للقيم.

$$\bar{x} = \frac{1}{n} \sum x_i$$

حساس جداً للـ Outliers.

## Weighted Mean

إعطاء وزن (Weight) لكل قيمة حسب أهميتها.

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

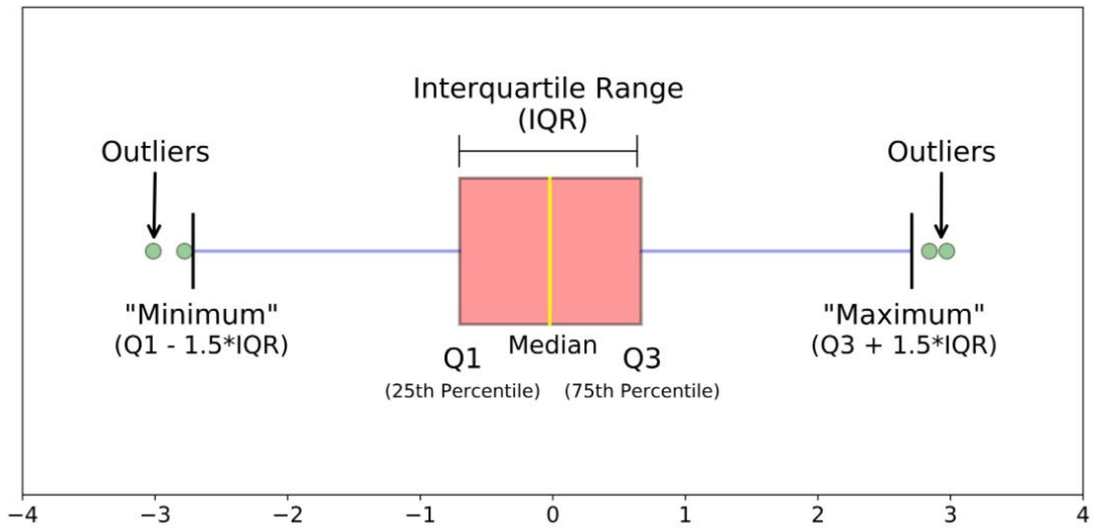
مفيد في Imbalanced Data.

## Trimmed Mean

حذف نسبة (مثلاً 5%) من القيم الصغرى والكبرى.

يقلل تأثير الـ Outliers.

# Median, Quartiles & Outliers




- **Median (Q2):** القيمة التي تقسم البيانات لنصفين. (Robust to Outliers).
- **IQR (Interquartile Range):** الفرق بين Q3 و Q1.

$$IQR = Q3 - Q1$$

**اكتشاف القيم الشاذة (Outliers):**

أي قيمة خارج النطاق:

[Q1 - 1.5*IQR , Q3 + 1.5*IQR]

# التشتت: Variance & Standard Deviation

## Variance

متوسط مربعات انحراف القيم عن المتوسط.

$$\sigma^2 = \frac{1}{N} \sum ( x_i - \mu )^2$$

## Standard Deviation (σ)

الجذر التربيعي للتباين (يعبر عن تباعد القيم).

## الأهمية في Machine Learning

أساس عملية الـ **Standardization (Z-Score)**:

$$z = \frac{x - \mu}{\sigma}$$

ضروري لخوارزميات مثل **K-Means**, **Gradient Descent**, و **SVM** لتسريع التعلم ومنع سيطرة ميزة (Feature) على أخرى.

# Correlation & Covariance

## Covariance

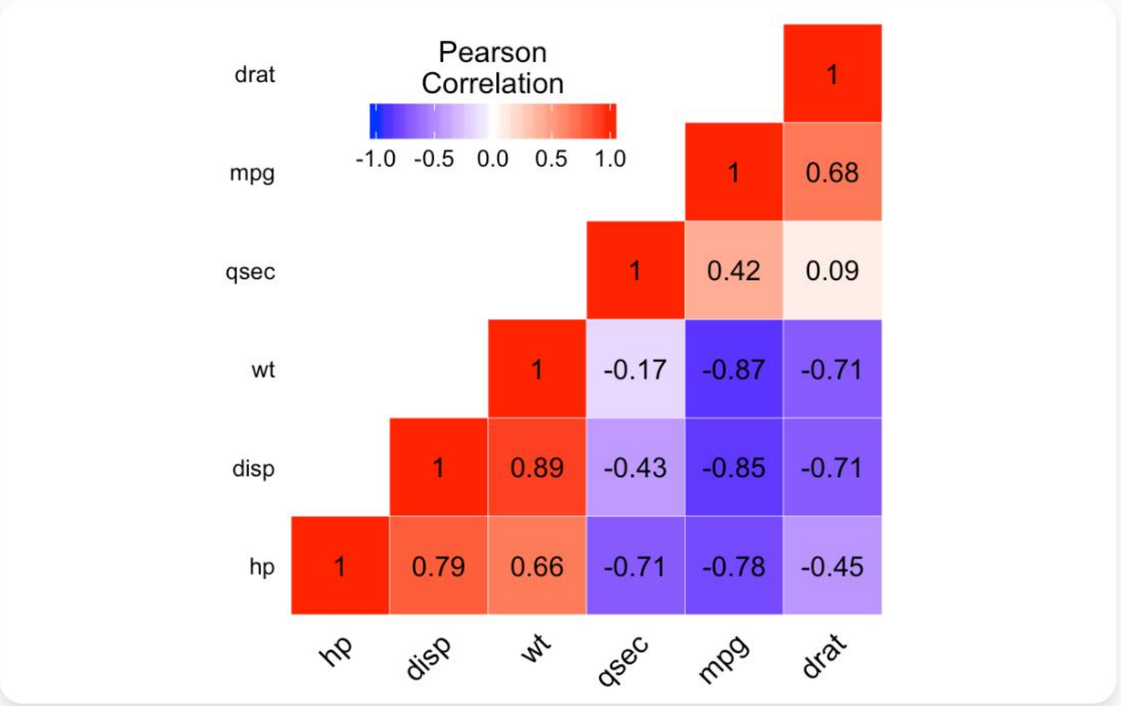
يقيس اتجاه العلاقة بين متغيرين (موجبة أو سالبة)، لكنه يعتمد على الوحدات.

## Correlation (Pearson)

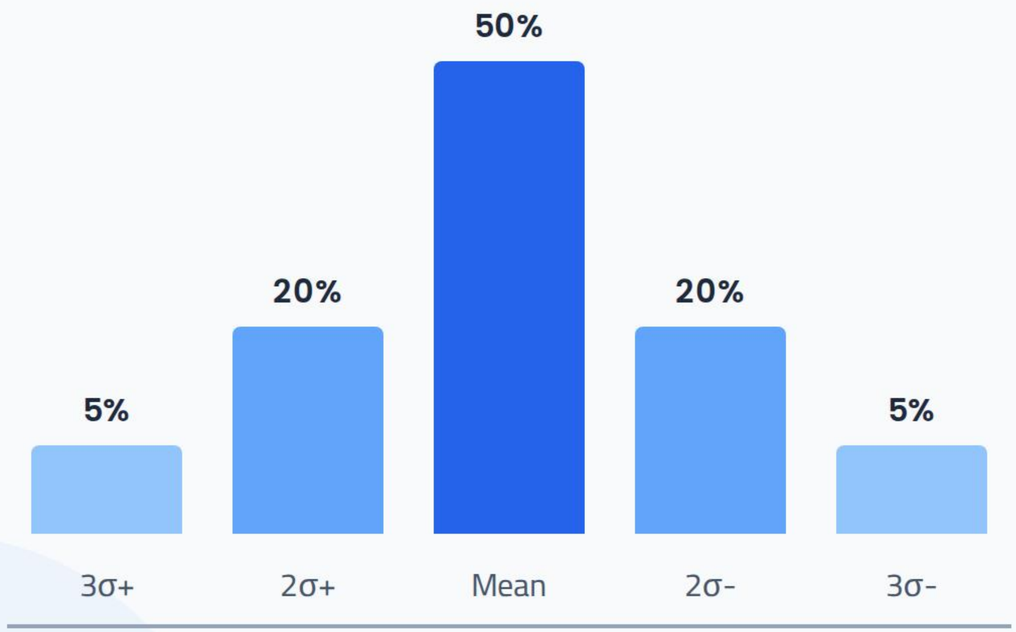
نسخة معيارية من Covariance تتراوح بين [-1 , 1].

- **1:** علاقة طردية قوية.
- **-1:** علاقة عكسية قوية.
- **0:** لا توجد علاقة خطية.

**في ML:** نستخدمه لحذف الـ Features المكررة (Multicollinearity).



# التوزيع الطبيعي: Normal Distribution




أهم توزيع في الإحصاء، يظهر بشكل "الجرس" (**Bell Curve**).

- معرف بالـ Mean (المركز) و Std Dev (العرض).
- أساسي لنماذج مثل **Gaussian Naive Bayes**.
- يستخدم **Histogram** (الرسم البياني المقابل) لفهم توزيع البيانات.

الرسم يوضح توزيع تكراري (Frequency) يأخذ شكل الجرس.

# Parametric vs Non-Parametric

| وجه المقارنة | Parametric | Non-Parametric |
| --- | --- | --- |
| **الافتراضات** | تفترض توزيعاً محدداً (مثل Normal). | لا تفترض توزيعاً محدداً (Distribution-free). |
| **التعقيد** | عدد Parameters ثابت ومحدود. | يزداد التعقيد مع زيادة البيانات. |
| **أمثلة** | Linear Regression, Logistic Regression | KNN, Decision Trees, Random Forest |
| **المميزات** | سريعة، تحتاج بيانات أقل. | مرنة، تمثل علاقات معقدة. |

# ملخص لمهندس الذكاء الاصطناعي

استخدم **Descriptive Stats** (Histogram, Boxplot) دائماً كخطوة أولى لفهم بياناتك قبل التدريب.

تأكد أن الـ **Sample** تمثل الـ **Population** وتجنب التحيز (Bias).

عالج الـ **Outliers** باستخدام IQR أو Trimmed Mean لأنها تؤثر بشدة على نماذج مثل Linear Regression.

استخدم **Variance & Std Dev** لعمل Standardization للبيانات (مهم جداً للـ Deep Learning).

# أسئلة؟

شكراً لاستماعكم

Presented by: AI Engineering Team

# Image Sources

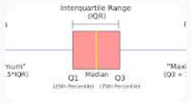
https://www.simplilearn.com/ice9/free_resources_article_thumb/population_vs_sample.jpg

Source: www.simplilearn.com


---

https://miro.medium.com/max/9000/1*2c21SkzJMf3frPXPAR_gZA.png

Source: www.kdnuggets.com




---

https://www.sthda.com/sthda/RDoc/figure/ggplot2/ggplot2-correlation-matrix-heatmap-add-correlation-coefficients-1.png

Source: www.sthda.com