

# Lecture: Data Analysis and Machine Learning Theory

KTH AI Student

January 31, 2025

# About me



- ▶ Jag heter Martín! I am from Chile, did my Bachelor's Degree at Universidad de Chile, doing a year-long exchange at KTH.
- ▶ I am currently interning at Hopsworks, an AI Lakehouse in Stockholm. Working on ML pipelines and LLMs.
- ▶ I enjoy running, hiking, and trying to improve my awful Swedish.



# In this lecture



- ▶ We will review key concepts in data analysis and machine learning theory.
  - ▶ Descriptive Statistics and Data Visualization.
  - ▶ Probability Theory and Simulation.
  - ▶ Correlation and Regression Analysis.
  - ▶ A/B Testing and Hypothesis Testing.

# Question 1



Imagine we have a big dataset, and we want to summarize it.  
What are some ways we can do this?

# Example: Student Test Scores



- ▶ **Dataset:** Contains scores of students.
- ▶ **Goals:**
  - ▶ Compute key descriptive statistics to summarize performance.
  - ▶ Visualize score distributions to identify trends or outliers.
  - ▶ Provide actionable insights to improve teaching methods.

**Descriptive Statistics:** Summarize and describe the main features of a dataset.

- ▶ **Mean:** The average value of a dataset.

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

- ▶ **Median:** The middle value when data is sorted.

$$x_{\text{median}} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if } n \text{ is even} \end{cases}$$

- ▶ **Mode:** The most frequently occurring value.

$$x_{\text{mode}} = \text{value with highest frequency}$$

- ▶ **Variance:** Measures the spread of data points from the mean.

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

- ▶ **Standard Deviation:** Square root of variance, represents data dispersion.

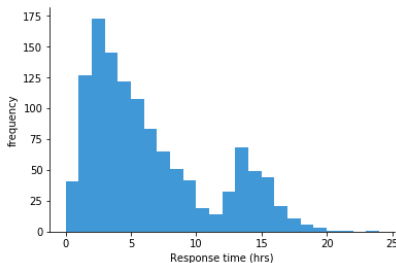
$$SD = \sqrt{\sigma^2}$$

- ▶ **Range:** Difference between the maximum and minimum values.

$$\text{Range} = \max(x) - \min(x)$$

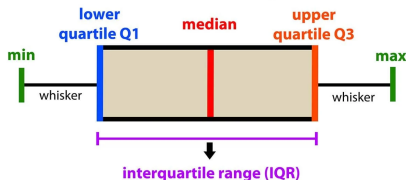
**Data Visualization:** Graphical representation of data.

- **Histograms:** Show frequency distribution of data.



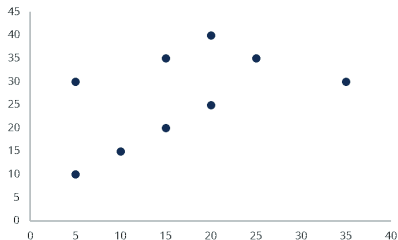
- **Box Plots:** Visualize data spread and identify outliers.

**introduction to data analysis: Box Plot**

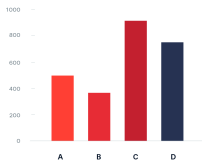




- **Scatter Plots:** Display relationships between two variables.



- **Bar Charts:** Compare categorical data.



## Question 2



We have a dataset, but does this dataset represent the real world?  
How can we estimate the probability of events?

# Example: Simulation Tasks



- ▶ Simulate 1000 coin tosses to calculate the probability of heads and compare with theoretical value.
- ▶ Simulate 1000 dice rolls to calculate:
  - ▶ Probability of rolling a prime number.
  - ▶ Conditional probability of a prime given the number is odd.
- ▶ Use Monte Carlo simulation to estimate  $\pi$ .

- ▶ **Probability:** Study of the likelihood of events.
  - ▶ **Theoretical Probability:** Based on known outcomes (e.g., coin toss).

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

- ▶ **Simulated Probability:** Estimated by running experiments or simulations.
- ▶ **Bayes' Theorem:** Describes conditional probability, updates beliefs based on evidence.

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

- ▶ **Probability Distributions:** Represent how probabilities are distributed over values.

- ▶ **Uniform Distribution:** All outcomes are equally likely.

$$P(x) = \frac{1}{n} \quad \text{for } x \in \{1, 2, \dots, n\}$$

- ▶ **Binomial Distribution:** Number of successes in fixed trials.

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

- ▶ **Normal Distribution:** Bell-shaped curve, common in natural data.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- ▶ **Monte Carlo Simulation:** Uses random sampling to estimate mathematical results.
  - ▶ Example: Estimate  $\pi$  by generating random points in a square and calculating the ratio inside a quarter circle.

$$\pi \approx 4 \times \frac{\text{Number of points inside circle}}{\text{Total number of points}}$$

## Question 3



We have two variables, how can we determine if they are related?  
How can we predict one variable based on the other?

# Example: Car Prices and Mileage



- ▶ **Dataset:** Contains car prices and mileage.
- ▶ **Tasks:**
  - ▶ Compute the correlation coefficient to assess the strength and direction of the relationship.
  - ▶ Build a simple linear regression model to predict prices based on mileage.
  - ▶ Visualize the data and regression line to interpret the results.



- ▶ **Correlation:** Measures the strength and direction of the linear relationship between two variables.
  - ▶ **Range:** Values range from  $-1$  to  $1$ .
  - ▶ **Interpretation:**
    - ▶  $1$ : Perfect positive correlation.
    - ▶  $-1$ : Perfect negative correlation.
    - ▶  $0$ : No linear correlation.

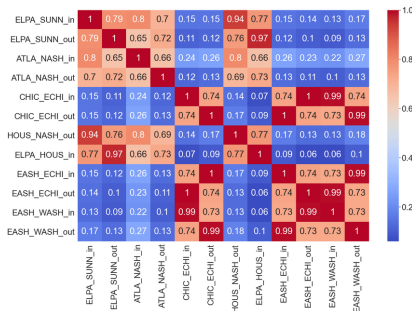
# Key Concepts: Correlation



- **Correlation Coefficient:** Denoted by  $r$ .

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

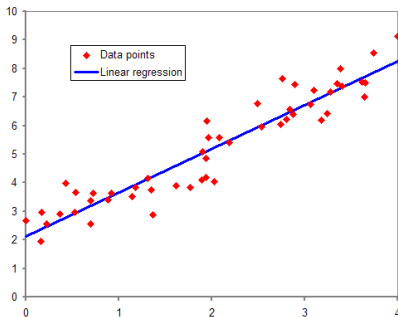
- **Correlation Matrix:** Displays pairwise correlations between variables.



# Key Concepts: Regression Analysis



- ▶ **Regression Analysis:** Models the relationship between a dependent variable and one or more independent variables.
  - ▶ **Simple Linear Regression:**  $y = \beta_0 + \beta_1 x + \epsilon$
  - ▶ **Goals:**
    - ▶ Estimate the coefficients  $(\beta_0, \beta_1)$ .
    - ▶ Minimize prediction error  $(\epsilon)$ .
  - ▶ **Evaluation Metrics:** Assess model fit using metrics such as Mean Squared Error (MSE).



## Question 4



We have two groups, how can we determine if they are significantly different? How can we validate our assumptions?

# Example: Website Redesign A/B Test



- ▶ **Dataset:** User engagement metrics for old and new designs.
- ▶ **Tasks:**
  - ▶ Perform a t-test to compare engagement levels.
  - ▶ Calculate and interpret the p-value.
  - ▶ Determine whether the new design significantly improves engagement.

- ▶ **Hypothesis Testing:** Framework to evaluate whether observed data provides sufficient evidence to reject a null hypothesis ( $H_0$ ).
  - ▶ **Null Hypothesis ( $H_0$ ):** Assumes no effect or difference.
  - ▶ **Alternative Hypothesis ( $H_a$ ):** Suggests a significant effect or difference.
- ▶ **t-Test:** Compares means of two groups.
  - ▶ **t-statistic:** Quantifies the difference relative to variability.
  - ▶ **p-value:** Probability of observing results as extreme as the data, assuming  $H_0$  is true.
- ▶ **Significance Level:** Common threshold  $\alpha = 0.05$ .

- ▶ **Interpretation:**

- ▶ **p-value**  $< \alpha$ : Reject  $H_0$ , evidence supports  $H_a$ .
- ▶ **p-value**  $\geq \alpha$ : Fail to reject  $H_0$ , insufficient evidence.

- ▶ **Type I Error:** Incorrectly reject  $H_0$  (false positive).

- ▶ **Type II Error:** Incorrectly fail to reject  $H_0$  (false negative).

- ▶ **Power:** Probability of correctly rejecting  $H_0$ .

# Key Concepts: Hypothesis Testing



How to perform a t-test:

1. Define null and alternative hypotheses.
2. Choose a significance level  $\alpha$ .
3. Calculate the t-statistic.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

4. Calculate the degrees of freedom.

$$\text{df} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

5. Calculate the p-value.
6. Make a decision based on the p-value.



- ▶ Reviewed essential concepts in data analysis and machine learning:
  - ▶ Descriptive statistics and visualization to summarize and understand data.
  - ▶ Probability and simulation to estimate theoretical and practical outcomes.
  - ▶ Regression analysis to model relationships and make predictions.
  - ▶ Hypothesis testing to assess differences and validate assumptions.
- ▶ Emphasized critical thinking and interpretation of results for data-driven decisions.

# Lecture: Data Analysis and Machine Learning Theory

KTH AI Student

January 31, 2025