# Lecture 10: Introduction to Statistics

Dr. Pooja Sarin

January 27-28, 2025

## 1 Introduction to Statistics

Statistics is the study of collecting, analyzing, interpreting, and presenting data. It plays a crucial role in pattern recognition (PR) by enabling:

- Data summarization and visualization.
- Identification of trends and patterns.
- Decision-making under uncertainty.
- Evaluation of model performance.

## 2 Role of Statistics in Pattern Recognition (PR)

In pattern recognition, statistics helps:

- Quantify uncertainty and variability in data.
- Develop models to identify and classify patterns.
- Evaluate the performance of algorithms using statistical measures.
- Support inferential processes, such as hypothesis testing and parameter estimation.

**Example:** In spam detection, statistical analysis of word frequencies helps identify spam messages.

## 3 Types of Statistics

Statistics is broadly divided into two categories:

- **Descriptive Statistics:** Summarizes and describes the main features of a dataset.
- **Inferential Statistics:** Makes predictions or inferences about a population based on a sample.

# 4 Descriptive Statistics

Descriptive statistics provides a summary of data through measures of central tendency and dispersion.

## 4.1 Measures of Central Tendency

Central tendency measures indicate the center or typical value of a dataset.

- **Mean (Average):** The sum of all values divided by the number of values.

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

  **Example:** For scores $[70, 80, 90]$, the mean is:

$$\frac{70 + 80 + 90}{3} = 80 \tag{2}$$

- **Median:** The middle value in a sorted dataset. **Example:** For scores $[60, 75, 80, 90, 95]$, the median is 80.

- **Mode:** The most frequently occurring value in a dataset. **Example:** For scores $[60, 70, 70, 80]$, the mode is 70.

## 4.2 Measures of Dispersion

Dispersion measures indicate the spread or variability of data.

- **Range:** The difference between the maximum and minimum values.

$$\text{Range} = \text{Max} - \text{Min} \tag{3}$$

  **Example:** For scores $[60, 70, 90]$, the range is $90 - 60 = 30$.

- **Variance:** The average of the squared differences from the mean.

$$\text{Variance} = \frac{\sum_{i=1}^{n} (x_i - \text{Mean})^2}{n} \tag{4}$$

  **Example:** For scores $[60, 70, 80]$, variance is:

$$\frac{(60 - 70)^2 + (70 - 70)^2 + (80 - 70)^2}{3} = 66.67 \tag{5}$$

- **Standard Deviation:** The square root of the variance.

$$\text{Standard Deviation} = \sqrt{\text{Variance}} \tag{6}$$

  **Example:** For variance 66.67, standard deviation is:

$$\sqrt{66.67} \approx 8.16 \tag{7}$$

# 5 Inferential Statistics

Inferential statistics draws conclusions about populations based on samples.

- **Hypothesis Testing:** Tests assumptions about a population parameter.
- **Confidence Intervals:** Provides a range of values for estimating a population parameter.
- **Regression Analysis:** Models relationships between variables.

**Example:** In A/B testing for a website, inferential statistics determine if a new design increases user engagement.

# 6 Use Cases of Descriptive Statistics in Pattern Recognition

- **Feature Analysis:** Understand the distribution of features in a dataset.
- **Data Cleaning:** Identify outliers and missing values.
- **Model Input:** Standardize features for machine learning models.

# 7 Conclusion

Descriptive and inferential statistics are foundational tools in pattern recognition. Measures of central tendency and dispersion summarize data, while inferential methods enable data-driven decisions. Understanding these concepts is essential for preprocessing, analyzing, and interpreting data in real-world applications.