Lecture Title: Introduction to Statistics

Dr. Amit Ranjan

School of Computer Science

What is Statistics?

- Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data.
- It encompasses techniques for summarizing data, making inferences, and drawing conclusions.
- Statistical methods help measure uncertainty and variability in datasets.

Role of Statistics in Machine Learning

- Statistics plays a pivotal role in extracting meaningful insights from data to make informed decisions.
- Provides the foundation for various ML algorithms, enabling analysis, interpretation, and prediction of complex patterns.
- Helps quantify uncertainty and variability in data, allowing for confident data-driven decisions.

Applications of Statistics in Machine Learning

- **Feature Engineering**: Converts raw data into meaningful predictors for ML models.
- **Image Processing**: Supports object recognition and segmentation.
- **Anomaly Detection & Quality Control**: Identifies deviations from norms in industrial and security settings.
- **Environmental Observation**: Monitors land cover patterns and ecological trends.

Types of Statistics

Descriptive Statistics

- Helps simplify and organize large datasets for better understanding.
- Summarizes and describes key features of data.

Inferential Statistics

- Uses sample data to make predictions or inferences about a larger population.
- Helps in hypothesis testing, estimation, and model validation.

Descriptive Statistics

- **Measures of Central Tendency**

Mean (Arithmetic Average)

- Formula: $\mu = \frac{\sum X_i}{N}$ (for population) $\bar{X} = \frac{\sum X_i}{n}$ (for sample)
- Example:
  - Given data: $5, 10, 15, 20, 25$
  - Mean = $(5 + 10 + 15 + 20 + 25)/5 = 15$

Median (Middle Value)

- If n is odd: Median = $\left(\frac{n+1}{2}\right)^{th}$ value.
- If n is even: Median = Average of $\left(\frac{n}{2}\right)^{th}$ value and next value.
- Example:
  - Given data: $5, 10, 15, 20, 25$ → Median = 15
  - Given data: $5, 10, 15, 20$ → Median = $(10 + 15)/2 = 12.5$

Descriptive Statistics

- **Measures of Central Tendency**

Mode (Most Frequent Value)

- The value that appears most frequently in a dataset.
- Example:
  - Given data: $2, 3, 3, 3, 4, 5, 6, 6$
  - Mode = 3

Descriptive Statistics

- **Measures of Dispersion**

Range

- Formula: $Range = X_{max} - X_{min}$
- Example:
  - Given data: $5, 10, 15, 20, 25$
  - Range = $25 - 5 = 20$

Variance

- Measures the average squared deviation from the mean.
- Formula (Population variance): $\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$
- Formula (Sample variance): $s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$
- Example:
  - Given data: $2, 4, 6, 8, 10$
  - Variance = 8

Descriptive Statistics

- **Measures of Dispersion**

Standard Deviation

- Square root of variance, indicating data spread relative to the mean.
- Formula: $\sigma = \sqrt{\sigma^2}$ (Population) $s = \sqrt{s^2}$ (Sample)
- Example:
  - Given data: $2, 4, 6, 8, 10$, variance = 8
  - Standard deviation = $\sqrt{8} = 2.83$

Interquartile Range (IQR)

- Difference between the first (Q1) and third (Q3) quartiles.
- Measures data spread around the median.

Inferential Statistics

Population vs. Sample

- **Population**: The entire group under study.
- **Sample**: A subset of the population used for analysis.

Estimation Methods

Point Estimation

- Provides a single value estimate of a population parameter.

Interval Estimation

- Offers a range of values (confidence interval) where the parameter likely falls.

Confidence Intervals

- Indicate reliability of an estimate.
- Example: A CI of 95% suggests that the true parameter lies within the range 95% of the time.