Data Science & Statistics: An Overview

Exploring the Universe of Data

Thomas Torku, Ph.D.

February 10, 2024

Agenda

Introduction to Data Science

Scope of Data Science

Applications of Data Science

Basic Statistics in Data Science

Measures of Central Tendency

Measures of Dispersion

Probability Theory

Introduction to Data Science

- Definition: An interdisciplinary field using scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
- Interdisciplinary nature: Combines aspects of mathematics, statistics, computer science, and domain expertise.
- Components:
  - Data: The raw material for data analysis.
  - Algorithms: Procedures or formulas for solving problems.
  - Inference: Concluding data.

Scope of Data Science

- Industry landscape
- Demand for skills
- Future trends

Applications of Data Science

- Healthcare: Predictive analytics in patient care
- Finance: Fraud detection and risk management
- E-Commerce: Personalized experiences
- Other industry examples

Basic Statistics in Data Science

- Definition: Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data.
- Descriptive vs. Inferential:
  - Descriptive Statistics: Summarize or describe the characteristics of a data set.
  - Inferential Statistics: Make predictions or inferences about a population based on a sample of data.

Measures of Central Tendency

- Mean: The sum of all measurements divided by the number of observations in the data set.
- Median: The middle value when the data set is ordered from least to greatest.
- Mode: The value that appears most frequently in a data set.
- Usage: These measures provide a single value that describes the center of the data.
- Examples: Give numerical examples for a hypothetical data set.

Measures of Dispersion

- Range: The difference between the highest and lowest values in the data set.
- Variance: Measures how far a set of numbers is spread out from their mean.

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

- Standard Deviation: The square root of the variance, representing the average amount of variability in the data set.

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

- Implications: Dispersion gives insight into the variability of the data.
- Examples: Provide examples calculating the range, variance, and standard deviation for a data set.

Probability Theory

- Definition: Probability theory is the branch of mathematics concerned with analysis of random phenomena.
- Events & Outcomes: An event is a set of outcomes of an experiment to which a probability is assigned.
- Probability Scale: A measure between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.
- Basic concepts:
  - Independent Events: Two events are independent if the occurrence of one does not affect the occurrence of the other.
  - Dependent Events: Two events are dependent if the occurrence of one event affects the occurrence of another.
  - Mutually Exclusive Events: Two events are mutually exclusive if they cannot occur at the same time.
- Examples: Illustrate with a coin toss, dice roll, or drawing cards from a deck.