TOPIC-03: INTRODUCTION TO STATISTICS & PROBABILITY

Statistics is all about data.

- Descriptive statistics
- Inferential statistics

Average: Give me ‘middle’ or ‘typical number’ or ‘measure of central tendency’.

Arithmetic Mean = Sum of all data points / Total number of data points.

Median = Middle Number.

Mode = Most common number in the dataset.

PROBABILITY AND RANDOM VARIABLES

Probability measures how likely an event is to occur, ranging from 0 to 1, calculated by dividing the number of favorable outcomes by the total number of possible outcomes.

Random variables represent outcomes of random events, with discrete variables having countable outcomes (like rolling a dice) and continuous variables having an infinite range of values (like arrival times of a bus).

PROBABILITY DISTRIBUTION

For discrete random variables, probability distributions assign a probability to each possible outcome in the sample space, such as the uniform distribution where each outcome has an equal probability. For continuous variables, probability distributions are described by probability density functions, where probabilities are determined for intervals of values rather than specific outcomes.

VARIANCE, AND STANDARD DEVIATION

Variance measures the spread of values around the mean, calculated by averaging the squared deviations from the mean, while the standard deviation is the square root of the variance, representing the average distance from the mean.

QUARTILES

Quartiles split the data into four parts, with the first quartile (Q1) being the value below which 25% of the data falls and the third quartile (Q3) below which 75% falls. The inter-quartile range (IQR) measures the spread of the middle 50% of the data.

REAL-WORLD DATA

Real-world data often resemble samples from a probability distribution, allowing the use of statistical methods to analyze and model the data. Histograms can visualize the distribution, showing the frequency of data points within specified intervals, and box plots can illustrate the median, quartiles, and potential outliers of the data.

NORMAL DISTRIBUTION

Normal distribution, characterized by a bell-shaped curve, is common in real-world data. It is defined by its mean and standard deviation, with most values clustering around the mean. It is significant because of the central limit theorem, which states that the mean of a large sample of any distribution approximates a normal distribution.

HYPOTHESIS TESTING

Hypothesis testing evaluates whether differences between groups or samples are statistically significant. It involves comparing sample means and calculating confidence intervals to determine if observed differences are due to random variation or if they support a specific hypothesis, using tests like the student t-test.

LAW OF LARGE NUMBERS AND CENTRAL LIMIT THEOREM

The law of large numbers of states that as the sample size increases, the sample mean approaches the population mean. The central limit theorem asserts that the distribution of the sample mean of any independent, random variable will approximate a normal distribution as the sample size grows.

COVARIANCE AND CORRELATION

Covariance measures how two variables change together, indicating the direction of their relationship. Correlation normalizes covariance to a range of -1 to 1, where 1 means a perfect positive relationship, -1 a perfect negative relationship, and 0 no relationship. Correlation helps identify the strength and direction of the linear relationship between variables.