Tasks Lec 5

# Task 1:

Gaussian Distribution (Normal Distribution): The curve has a symmetric bell shape with a well-defined mean and standard deviation.

Because of its frequency in natural events, it is used in a variety of fields. It is frequently used to model continuous variables like heights, weights, and test scores.

Binomial Distribution: The number of successes in a fixed number of independent Bernoulli trials is described by a discrete distribution.

Applications: When dealing with a binary outcome (e.g., success/failure) and counting the number of successes, such as determining the chance of success in a certain number of trials.

The Poisson Distribution:

Characteristics: Discrete distribution that models the probability of a certain number of events occurring within a fixed interval of time or space.

Applications: Used to model rare events or the occurrence of events in a fixed period, such as the number of phone calls received per hour or the number of accidents per day.

Exponential Distribution:

Characteristics: Continuous distribution that describes the time between events in a Poisson process.

Applications: Used to model the time between consecutive events occurring at a constant average rate, such as the time between customer arrivals at a service counter.

Uniform Distribution: Continuous distribution with a constant probability for all outcomes within a certain interval.

Applications: When all outcomes within a range are equally likely, for as when rolling a fair die or generating random numbers, this function is used.

Gamma Distribution: In a Poisson process, a continuous distribution is frequently used to simulate waiting times until a particular number of events occur.

Applications: Reliability analysis, queuing theory, and modeling continuous positive variables such as time to failure or repair are all examples of applications.

Chi-Square Distribution:

Characteristics: A continuous distribution that appears in a variety of statistical tests and calculates the sum of squared standard normal deviations.

Applications: Used to test hypotheses, perform goodness-of-fit tests, and calculate confidence intervals for population variances.

Student's t-Distribution:

Characteristics: A continuous distribution that appears in a variety of statistical tests and calculates the sum of squared standard normal deviations.

Applications: Used to test hypotheses, perform goodness-of-fit tests, and calculate confidence intervals for population variances.

F-Distribution: Continuous distribution used in statistical tests to compare variances of two or more samples.

Applications: Used in ANOVA and regression analysis to test variance equality among groups or to compare different regression models.

# Task 2:

Transform the data: In some circumstances, adding a mathematical adjustment to the data can help it resemble a normal distribution more closely. Transformations that are commonly used include logarithmic, square root, and reciprocal transformations. You may typically stabilize the variance and make the data more acceptable for statistical studies by converting the data.

Non-parametric testing: non-parametric tests can be used instead of typical parametric tests if your data does not match the assumptions of normality. Non-parametric tests are less dependent on specific distributional assumptions and therefore more resistant to deviations from normality. The Mann-Whitney U test, Wilcoxon signed-rank test, and Kruskal-Wallis test are examples of non-parametric tests.

Resampling methods: Bootstrapping and permutation testing, for example, can give robust statistical inference without relying on distributional assumptions. These methods entail repeatedly sampling your data to estimate the sampling distribution of a statistic or to test a hypothesis. They are especially useful when the underlying distribution is unknown or does not conform to a particular distribution.

Robust statistics: robust statistical approaches are those that are less sensitive to departures from normality or the presence of outliers. When the data is not normally distributed, robust estimators such as the median or trimmed mean can be employed instead of the mean. When the relationship between variables is not linear or there are influential observations, robust regression methods such as robust regression or quantile regression can be utilized.

Data stratification or grouping: Dividing your data into subgroups or categories depending on important characteristics can sometimes aid in the discovery of underlying patterns or linkages. Even if the aggregate data does not follow a normal distribution, analyzing each subgroup separately may yield more meaningful insights.

Data transformation should be done with caution because it can bring complications and interpretation issues. It is critical to comprehend the ramifications of the transformation and determine whether it is consistent with the underlying theory or context of your data.

# Task 3:

1. Parametric Tests:
   * Student's t-test: Used to compare means between two groups.
   * Analysis of Variance (ANOVA): Used to compare means across multiple groups.
   * Paired t-test: Used to compare means within the same group, typically before and after an intervention or treatment.
2. Non-parametric Tests:
   * Mann-Whitney U test: Used to compare the distributions of two independent groups when assumptions for parametric tests are violated.
   * Kruskal-Wallis test: Used to compare the distributions of more than two independent groups.
   * Wilcoxon signed-rank test: Used to compare paired observations when assumptions for parametric tests are violated.
3. Chi-Square Tests:
   * Chi-Square goodness-of-fit test: Used to test whether observed data fits an expected distribution.
   * Chi-Square test of independence: Used to assess the association between two categorical variables.
4. Regression Analysis:
   * Simple Linear Regression: Used to model and analyze the relationship between a dependent variable and one independent variable.
   * Multiple Linear Regression: Used to model and analyze the relationship between a dependent variable and multiple independent variables.
   * Logistic Regression: Used to model and analyze the relationship between a binary dependent variable and one or more independent variables.
5. Correlation Tests:
   * Pearson's correlation coefficient: Measures the linear association between two continuous variables.
   * Spearman's rank correlation coefficient: Measures the monotonic association between two variables, regardless of linearity.
6. Survival Analysis:
   * Kaplan-Meier estimator: Estimates survival probabilities over time in the presence of censored data.
   * Log-rank test: Compares survival curves between two or more groups.
7. Multivariate Analysis:
   * Principal Component Analysis (PCA): Reduces the dimensionality of a dataset while preserving the most important information.
   * Factor Analysis: Examines underlying factors or latent variables within a dataset.
   * Cluster Analysis: Groups similar objects or observations into clusters based on their characteristics.