**STATISTICS WORKSHEET-1**

1 (a) True

2 (a) Central Limit Theorem

3 (b) Modeling bounded count data

4 (d) All of the mentioned

5 (c) Poisson

6 (b) False

7 (b) Hypothesis

8 (a) 0

9 (d) None of the mentioned

**Ans.10**

**Normal Distribution** is a fundamental concept in statistics that represents a continuous probability distribution of a random variable. It is also known as a **Gaussian distribution** or **bell curve** due to its characteristic bell-shaped graph.

**Ans.11**

Handling missing data involves several strategies, depending on the type of data and the extent of missingness. Some common techniques include:

* **Mean/Median Imputation**: Replace missing values with the mean or median of the column. Simple but can distort the variance.
* **Mode Imputation**: Replace missing values with the mode for categorical data.
* **K-Nearest Neighbors (KNN)**: Impute missing values using the values of the nearest neighbors based on similarity in other features.
* **Regression Imputation**: Predict the missing value using a regression model based on other variables.
* **Multiple Imputation**: Generate several imputed datasets and average the results to account for the uncertainty in the imputation process.
* **Forward/Backward Filling**: For time-series data, fill missing values using the previous or next available value.

**Ans.12**

**A/B testing** (also known as split testing) is an experimental technique used to compare two versions of something to determine which one performs better. It is commonly used in marketing, web design, and product optimization.

* **Process**:
  + **Version A** (control) is the current version.
  + **Version B** (treatment) is the new version or variant.
  + Users are randomly assigned to one of the two groups, and the performance metrics (e.g., click-through rates, conversion rates) are measured.
  + **Hypothesis testing** is then applied to see if there’s a statistically significant difference between the two versions.

**Ans.13**

**Mean imputation** is a simple but generally not recommended method for handling missing data. While it can be used in specific cases, it has several drawbacks:

* **Pros**:
  + Simple to implement.
  + Preserves the sample size.
* **Cons**:
  + Reduces variability in the dataset, leading to biased estimates of variance and standard deviation.
  + May distort relationships between variables, particularly in regression models.
  + Does not account for the uncertainty in missing data.

In most cases, more sophisticated imputation techniques, such as **multiple imputation** or **KNN**, are preferred.

**Ans.14**

**Linear regression** is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal is to find the best-fitting linear equation to predict the target variable.

**Simple Linear Regression**: Models the relationship between one independent variable and one dependent variable.

**Multiple Linear Regression**: Extends the model to include multiple independent variables

**Ans.15**

Statistics can be broadly divided into two main branches:

1. **Descriptive Statistics**:
   * Summarizes and describes the features of a dataset using measures such as mean, median, mode, variance, standard deviation, and graphical representations (e.g., histograms, bar charts).
   * **Examples**: Central tendency measures, variability, frequency distributions.
2. **Inferential Statistics**:
   * Draws conclusions and makes predictions about a population based on a sample of data.
   * It involves techniques such as **hypothesis testing**, **confidence intervals**, and **regression analysis**.
   * **Examples**: t-tests, chi-square tests, ANOVA, regression models.

**MACHINE LEARNING**

1. A) Least Square Error
2. A) Linear regression is sensitive to outliers
3. B) Negative
4. B) Correlation
5. C) Low bias and high variance
6. B) Predictive model
7. D) Regularization
8. D) SMOTE
9. A) TPR and FPR
10. B) False
11. B) Apply PCA to project high dimensional data
12. A) We don’t have to choose the learning rate

B) It becomes slow when the number of features is very large

**Ans.13**

**Regularization** is a technique used in machine learning and statistics to prevent **overfitting** by introducing additional constraints or penalties to the model, especially when working with complex models or datasets with many features.

Two common types of regularization are:

* **Lasso (L1 Regularization)**: Adds a penalty equal to the absolute value of the magnitude of coefficients, promoting sparsity (i.e., some coefficients are shrunk to zero, effectively selecting a subset of features).
* **Ridge (L2 Regularization)**: Adds a penalty equal to the square of the magnitude of coefficients, which reduces large coefficients but does not shrink them to zero.

**Ans.14**

The following algorithms incorporate regularization:

* **Ridge Regression (L2 Regularization)**: Adds a penalty proportional to the sum of the squared coefficients.
* **Lasso Regression (L1 Regularization)**: Adds a penalty proportional to the sum of the absolute values of the coefficients, promoting sparsity.
* **Elastic Net**: Combines both L1 and L2 regularization terms, balancing between Ridge and Lasso.
* **Logistic Regression with regularization**: In classification problems, logistic regression can incorporate L1 or L2 regularization to avoid overfitting.
* **Support Vector Machines (SVM)**: The regularization parameter CCC controls the trade-off between maximizing the margin and minimizing classification error.
* **Neural Networks (with L2 Regularization or Dropout)**: L2 regularization (also called weight decay) is often used to constrain the weights of neural networks, and **Dropout** acts as a form of regularization by randomly dropping neurons during training.

**Ans.15**

In linear regression, the **error** (or **residual**) refers to the difference between the actual observed value and the predicted value from the regression model.

The goal of linear regression is to minimize the sum of squared residuals (errors) to find the best-fitting line. This process is called **Ordinary Least Squares (OLS)**

**PYTHON**

1 (c) %

2 (b) 0

3 (c) 24

4 (a) 2

5 (d) 6

6 (c) The finally block will be executed no matter if the try block raises an error or not

7 (a) It is used to raise an exception

8 (c) in defining a generator

9 (a) \_abc (c) abc2

10 (a) yield (b) raise