Statistics Worksheet-1 Answers

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.
(c) Outliers cannot conform to the regression relationship

10.
Answer:

Normal Distribution

The normal distribution is a continuous probability distribution. It looks like a bell-shaped curve, with most of the data concentrated in the middle and tapering off towards the sides. Its shape remains symmetrical on both sides of the mean. It is also known as Gaussian distribution. The shape of the normal distribution is determined by two important parameters-

Mean (μ)

The mean (μ) represents the center of the distribution.

Standard deviation (σ)

The standard deviation (σ) determines the spread or dispersion of the data around the mean.

Characteristics of the normal distribution:

Symmetry: The normal distribution is symmetric around its mean. It means, if we divide the shape in two equal parts then the left area will be equal to right area.

Bell-shaped curve: The graph of the normal distribution forms a smooth, bell-shaped curve. The highest point of the curve corresponds to the mean, and the curve gradually decreases in height as you move away from the mean in both directions.

Mean, median and mode: The mean, median, and mode of a normal distribution are all equal, and they all lie at the center of the distribution. That is why the normal distribution is also reffered as a measure of central tendency.

Standard deviation: The standard deviation determines the spread or dispersion of the data around the mean. If standard deviation is smaller then it indicates date is less spreaded around the mean. And if standard deviation is larger then it indicates date is more spreaded around the mean

Empirical Rule: The normal distribution follows the Empirical Rule.

68% of the data falls within one standard deviation of the mean.

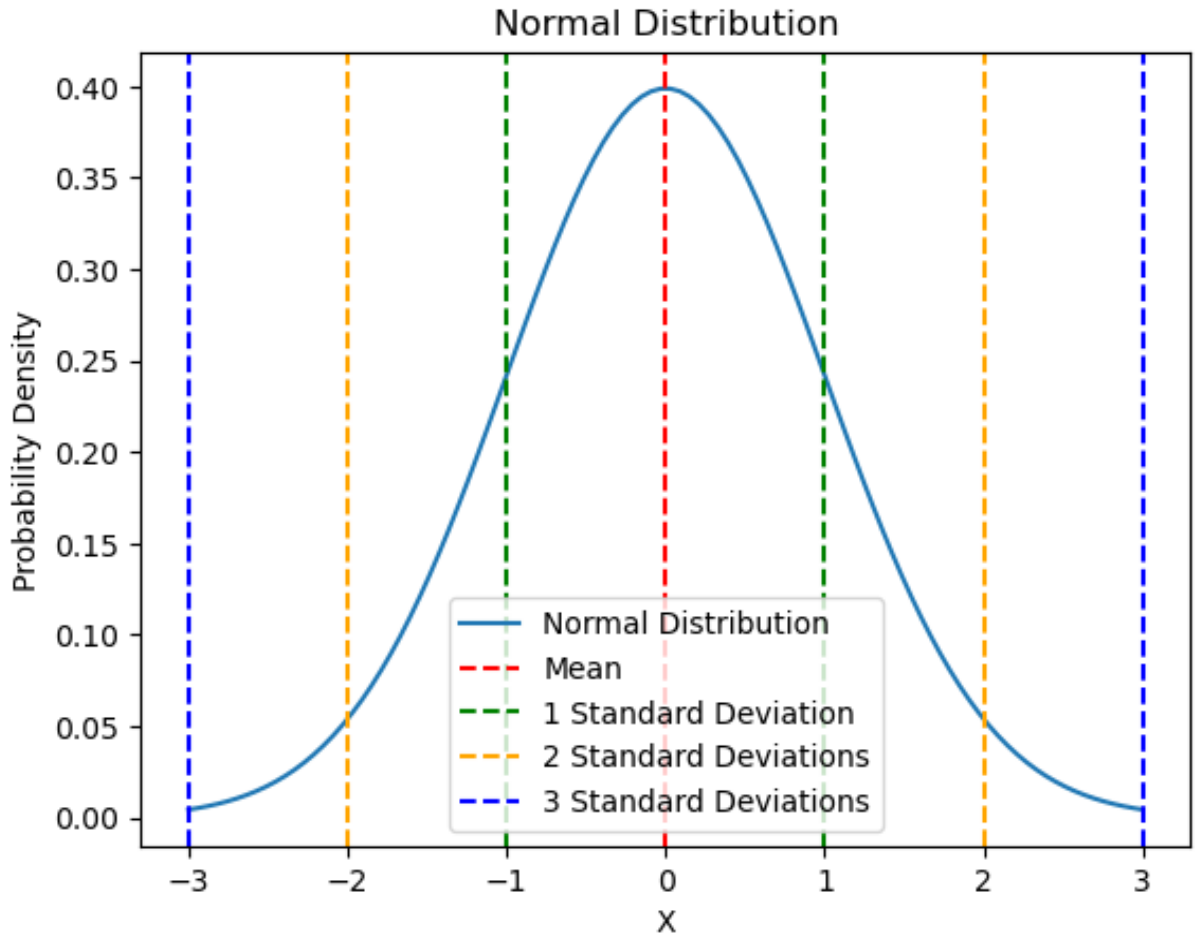
95% of the data falls within two standard deviations of the mean.

99.7% of the data falls within three standard deviations of the mean.

Z-score: The Z-score is a measure that indicates how many standard deviations a given data point is away from the mean in the context of a normal distribution.

Central Limit Theorem (CLT): Central Limit Theorem is also very important property in normal distribution. It states that when we have a large enough sample size from any population with a finite mean and standard deviation, the distribution of the sample means will be approximately normally distributed, regardless of the shape of the original population.

Graphical representation of Normal Distribution



Uses of Normal Distribution:

The normal distribution is used for several reasons in statistics and data analysis-

- It is used to understand patterns present in a dataset.
- It is used for parameter estimation
- It is used for hypothesis testing
- It is used for data transformation
- It is used for predictive modeling

11.

Answer:

Handling Missing Data

Handling missing data is an important step in data analysis and modeling. Dealing with missing values is important because they can impact the quality and reliability of data analysis and modeling. It's essential to handle missing values appropriately by applying imputation techniques or considering their effects during analysis to avoid biased or inaccurate results. Missing values can occur for various reasons, such as data entry errors, equipment malfunction, survey non-response, or simply the absence of information.

Missing values are typically represented by special markers or codes, such as "NA" (not available), "NaN" (not a number), or "null". These markers indicate the absence of a value or the lack of information.

Imputation Techniques

There are various imputation techniques dealing with missing values. The choice of technique depends on the specific dataset, nature of the variables and analysis requirements. It's important to consider the limitations and potential biases of each method and to assess the impact of missing data on the analysis.

Various imputation techniques are:

Mean/Median/Mode imputation: Fill in missing values with the average (mean), middle value (median), or most frequent value (mode) of the available data for that variable.

Forward/Backward fill: If the data has a time or sequence order, use the last observed value (forward fill) or the next observed value (backward fill) to fill in missing values.

Hot Deck imputation: Randomly pick a value from a similar record to fill in the missing value. This can be based on similarity measures like distance or other characteristics.

Regression imputation: Predict the missing values using a regression model based on the other available variables. The model learns from the observed data and estimates the missing values.

Multiple Imputation: Generate multiple possible values for missing data, create

multiple completed datasets. Each dataset is analyzed separately, and the results are combined for an overall estimate.

K-nearest neighbors imputation: Find the most similar observations based on available features and use their values to impute the missing data.

12.

Answer:

A/B Testing

In machine learning, A/B testing refers to the process of comparing the performance of different machine learning models or algorithms to determine which one produces better results. It is a technique used to evaluate and select the most effective model for a given task or problem.

"A" typically represents the control group or the baseline model. It is the current or existing model that is already in use or considered as the standard approach.

"B" represents the experimental group or the alternative model. It is the new or modified model being tested to determine if it performs better than the baseline model.

By comparing the performance of the "A" and "B" models on the same dataset, practitioners can assess which model produces better results based on predefined metrics or objectives. This comparison allows for data-driven decision-making and selecting the most effective model for a specific task or problem.

Steps to apply A/B testing in machine learning:

Objective and Hypothesis: Clearly define the objective of the A/B test, such as improving prediction accuracy or optimizing a specific metric. Formulate a hypothesis regarding which model or algorithm is expected to perform better.

Model Selection: Choose two or more models or algorithms that are candidates for the task. These models can vary in complexity, parameters, or underlying algorithms.

Data Splitting: Split the available data into separate training and testing sets. The training set is used to train each model, while the testing set is used to evaluate their performance.

Model Training: Train each model on the training set using appropriate techniques and algorithms. Ensure that the training process is consistent across all models.

Model Evaluation: Apply each trained model to the testing set and measure their performance using relevant metrics, such as accuracy, precision, recall, or F1 score.

Statistical Analysis: Perform statistical analysis to compare the performance of the models. Common techniques include hypothesis testing, confidence intervals, or effect size calculations.

Conclusion: Based on the statistical analysis, draw conclusions about the performance of the models and determine which model performs better according to the predefined objective.

13.

Answer:

Mean Imputation Technique

Mean imputation technique is a simple technique for handling missing data where the missing values are replaced with the mean value of the given data.

The process of mean imputation involves the following steps:

- Calculate the mean of the available data for the variable with missing values.
- Replace the missing values with the calculated mean.

Mean imputation is a commonly used imputation technique due to its simplicity and ease of implementation. It is acceptable but we cannot ignore its limitations.

Limitations of Mean Imputation Technique

Here are the limitations of mean imputation technique:

Loss of variability: Mean imputation replaces missing values with a single value, the mean. This means that the imputed values don't consider the range or spread of the original data. It can make the data look less variable or diverse than it actually is.

Bias introduction: Mean imputation assumes that the missing values are randomly distributed. However, if the missing values are related to the actual values, the imputed values may introduce bias into the analysis. It might distort the true relationships or patterns in the data.

Ignoring data patterns: Mean imputation doesn't take into account any specific patterns or relationships in the data. It assumes that the missing values are similar to the observed values. This assumption may not always hold true and can lead to inaccurate results.

Validity of assumptions: Mean imputation assumes that the missing values have no additional meaning and that they don't carry any important information. However, in some cases, missing values might actually have a different significance that could affect the analysis.

Before using mean imputation technique one need to consider following points-

Check assumptions: Make sure the missing data is randomly distributed. If it's related to certain factors, mean imputation may not work well.

Consider the characteristics of data: Consider the characteristics of the dataset, like how much missing data there is and the overall pattern. This helps determine if mean imputation is suitable.

Motive of analysis: It is important to understand the motive of the analysis. If accuracy is important, we might need to explore other imputation methods.

14.

Answer:

Linear Regression

Linear regression is a statistical method used to model and analyze the relationship between a dependent variable and one or more independent variables. It aims to find a linear equation that best fits the data points and allows us to make predictions or understand the relationship between the variables.

We have one variable that we want to predict or explain (the dependent variable) and another variable that we use to make predictions (the independent variable). Regression line, is a straight line that represents the overall trend or

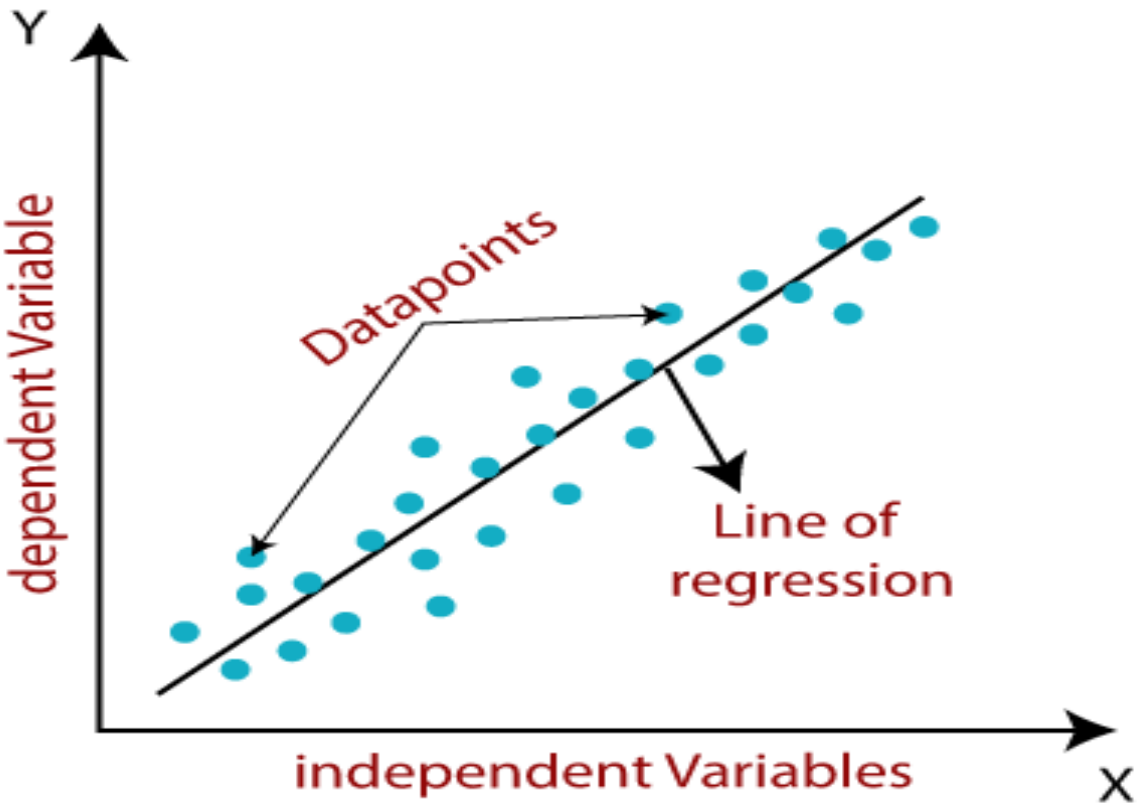
relationship between two variables. It is used in linear regression analysis to find the line that best fits the data points.

The regression line is defined by an equation of the form:

Y = a + bX

where Y is the dependent variable, X is the independent variable, a is the intercept, and b is the slope (the change in Y for a unit change in X).

Graphical representation of Linear Regression



The process of linear regression involves estimating the values of the intercept and slope based on the available data. This estimation is often done using a technique called least squares, which minimizes the sum of the squared differences between the observed and predicted values.

Linear regression is widely used in various fields, including economics, social sciences, finance, and machine learning. It helps in understanding relationships, making predictions, and identifying the strength and significance of the relationship between variables.

Types of Linear Regression

There are mainly two types of linear regression, each with its own specific characteristics and applications. Here are some common types:

- Simple Linear Regression
- Multiple Linear Regression

Simple Linear Regression

This is the most basic form of linear regression, involving only one independent variable and one dependent variable. It assumes a linear relationship between the variables and aims to find the best-fitting straight line.

Equation: $\mathbf{Y = a + bX + \varepsilon}$

Where

Y: Dependent variable

X: Independent variable

a: Intercept

b: Slope (change in Y for a unit change in X)

ε: error

Multiple Linear Regression

Multiple linear regression involves more than one independent variable. It allows for modeling the relationship between a dependent variable and multiple predictors. The goal is to find a linear equation that best fits the data, considering the effects of all the independent variables.

Equation: $\mathbf{Y = a + b_1X_1 + b_2X_2 + ... + b_nX_n + \varepsilon}$

Where

Y: Dependent variable

$\mathbf{X_1, X_2, ..., X_n}$**:** Independent variables

a: Intercept

$\mathbf{b_1, b_2, ..., b_n}$**:** Coefficients (slopes) for each independent variable

ε: error

14.

Answer:

Statistics

Statistics is the science of dealing with data. Statistics involves the collection, analysis, interpretation, presentation, and organization of data.

Statistics is used in various fields, including science, business, economics, social sciences, healthcare, engineering, and many others. It plays a crucial role in research, decision-making, quality control, policy development, and understanding patterns and trends in data.

Branches of Statistics

Statistics is a broad field that encompasses various branches or subfields. Some of the main branches of statistics are-

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics

Descriptive statistics is a branch of statistics. It is important for understanding and summarizing the characteristics of datasets used in machine learning. It involves calculating measures such as mean, median, mode, variance, and standard deviation to describe the central tendency, spread, and distribution of the data.

There are several types of descriptive statistics used to summarize different aspects of the data. Here are the main types:

- Measure of Central Tendency
- Measure of Dispersion

Measures of Central Tendency

These statistics provide information about the central or typical value of a dataset. The common measures of central tendency include:

Mean: The arithmetic average of all the values in the dataset.

Median: The middle value that separates the dataset into two equal halves when arranged in ascending or descending order.

Mode: The most frequently occurring value in the dataset.

Measures of Dispersion

These statistics indicate how spread out or dispersed the data points are. They provide information about the variability or spread of the dataset. The common measures of dispersion include:

Variance: The average of the squared differences between each data point and the mean.

Standard Deviation: The square root of the variance, representing the average distance between each data point and the mean.

Range: The difference between the maximum and minimum values in the dataset.

Interquartile Range (IQR): The difference between the third quartile (Q3) and the first quartile (Q1), representing the spread of the middle 50% of the data.

Measures of Shape: These statistics describe the shape or distribution of the dataset. They provide information about how the data points are distributed across the range. Some measures of shape include:

Skewness: Indicates the asymmetry of the dataset distribution. Positive skewness means a longer tail on the right, while negative skewness means a longer tail on the left.

Kurtosis: Measures the degree of peakedness or flatness of the dataset distribution. High kurtosis indicates a sharper peak and heavier tails, while low kurtosis indicates a flatter distribution.

Frequency Distribution: This type of descriptive statistic summarizes the dataset by counting the frequency or number of occurrences of each value or range of values. It is commonly displayed as a table or histogram, providing an overview of how often each value appears in the dataset.

Inferential statistics

Inferential statistics is a branch of statistics that involves using sample data to make inferences or draw conclusions about a larger population. It extends beyond the observed data and aims to generalize findings from the sample to the entire population. Inferential statistics uses probability theory and statistical

techniques to make these inferences and quantify the level of uncertainty associated with the conclusions.

Inferential statistics are divided into two categories:

- Hypothesis testing
- Regression analysis

Hypothesis Testing

Hypothesis testing is a method used in statistics to make conclusions about a population based on sample data. Hypothesis testing is a common technique in inferential statistics. It involves formulating a null hypothesis and an alternative hypothesis, collecting sample data, and using statistical tests to determine whether there is sufficient evidence to support rejecting the null hypothesis in favour of the alternative hypothesis. Hypothesis testing helps make decisions and draw conclusions about population parameters based on sample data.

Regression Analysis

Regression analysis is a statistical technique used to understand how one variable changes in relation to another variable. There are different types of regression models that can be used. It allows for the prediction of one variable based on the values of other variables. Inferential regression analysis involves assessing the statistical significance of the relationship between variables and making inferences about the population based on the regression model. Numerous regression models can be used, including simple linear, multiple linear, nominal, logistic, and ordinal regression.