Answers

Question no. 1

Answer: A) True

Question no. 2

Answer: A) Central Limit Theorem

Question no. 3

Answer: B) Modeling bounded count data

Question no. 4

Answer: D) All of the mentioned

Question no. 5

Answer: C) Poisson

Question no. 6

Answer: B) False

Question no. 7

Answer: B) Hypothesis

Question no. 8

Answer: A) 0

Question no. 9

Answer: C) Outliers cannot conform to the regression relationship.

Question no. 10

Answer: **A Normal distribution,** also known as a Gaussian distribution, is a continuous probability distribution that is symmetric about its mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical terms, the normal distribution will appear as a bell curve.

Characteristics of Normal Distribution:

1.Symmetry:

The distribution is symmetric around the mean, meaning the left and right halves of the graph are mirror images.

2.Mean, Median, and Mode:

For a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.

3.Bell-shaped Curve:

The shape of the distribution is bell-shaped, where the highest point is at the mean, and it tapers off equally on both sides.

4.Standard Deviation:

The spread of the data is determined by the standard deviation. A smaller standard deviation indicates that the data points are closer to the mean, while a larger standard deviation indicates that the data points are spread out over a wider range of values.

5. 68-95-99.7 Rule:

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

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

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

Understanding the normal distribution is fundamental in the field of statistics and probability, as it provides a basis for various statistical tests and real-world applications.

Question no. 11

Answer:

Handling missing data: is a crucial aspect of data preprocessing in any data analysis or machine learning pipeline.

Strategies for Handling Missing Data:

1. Remove Data:

Remove Rows: Delete rows with missing values. This is viable when the proportion of missing data is very small.

Remove Columns: Delete columns with a high percentage of missing values, especially if they are not critical features.

2. **Understand the Nature of Missingness**: Determine if the data is Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). The choice of imputation technique can depend on this.

3. **Evaluate the Impact**: After imputation, evaluate how the filled-in values affect your model's performance.

4. **Use Domain Knowledge**: Sometimes domain knowledge can suggest reasonable imputations that simple statistical methods may not capture.

5. **Avoid Defaulting to Mean/Mode Imputation**: While simple, mean/mode imputation can distort the data, especially with a high proportion of missing values.

6. **Consider Multiple Imputations**: When uncertainty in imputations is high, using multiple imputations can give a more robust result.

Recommendations of various Imputation Techniques:

1. Simple Imputation:

Mean/Median/Mode Imputation: Replace missing values with the mean (for numerical data), median (for numerical data with outliers), or mode (for categorical data).

2. Advanced Imputation Techniques: like

K-Nearest Neighbours (KNN) Imputation.

Multivariate Imputation by Chained Equations (MICE)

Multiple Imputation.

Regression Imputation.

3. Model-Based Imputation: Use Predictive Modeling: Train a model to predict missing values.

4. **Imputation Using Algorithms**: By using Random Forest Imputation.

Question no. 12

Answer:

A/B testing (also known as split testing): is a randomized controlled experiment used to compare two versions of a variable to determine which one performs better. It involves the following key steps:

Key Steps in A/B Testing:

1.Formulate Hypotheses:

Null Hypothesis (H_0): There is no difference between the control (A) and the variation (B).

Alternative Hypothesis (H_1): There is a significant difference between the control (A) and the variation (B).

2.Random Assignment:

Randomly assign subjects (e.g., website visitors) into two groups: one group is exposed to the control (A) and the other to the variation (B). This ensures that any differences observed between the groups can be attributed to the changes made rather than to other confounding factors.

3. Data Collection:

Collect data on the performance of both groups based on a predefined metric (e.g., conversion rate, click-through rate).

4. Statistical Analysis:

Use statistical tests (e.g., t-tests, chi-square tests) to compare the performance of the two groups. The goal is to determine whether any observed differences are statistically significant, meaning they are unlikely to have occurred by chance.

5.Interpret Results:

Determine whether to reject the null hypothesis. If the results show a statistically significant difference, the variation (B) is considered to have a different performance compared to the control (A).

In conclusion, A/B testing in statistics provides a rigorous framework for making data-driven decisions and optimizing processes, products or experiences based on empirical evidence.

Question no. 13

Answer**: Yes, mean imputation of missing data is acceptable practice. Mean imputation** is a simple and common method for handling missing data where the missing values in a dataset are replaced with the mean value of the observed data for that variable. However, while mean imputation is straightforward and easy to implement, it has several drawbacks and limitations.

Pros of Mean Imputation:

Simplicity: Easy to understand and implement.

Maintains Dataset Size: Keeps the dataset size the same, which can be important for some analyses.

Cons of Mean Imputation:

Distortion of Variability: Mean imputation reduces the natural variability in the data by creating more central tendency and can underestimate the variance.

Bias: It can introduce bias, especially if the data are not missing completely at random (MCAR). The relationships between variables might be distorted.

Impact on Correlations: It can artificially inflate correlations between variables because the same value (mean) is imputed across multiple instances.

Ignores Uncertainty: Does not account for the uncertainty introduced by the missing values.

When Mean Imputation Might Be Acceptable:

Small Proportion of Missing Data: If the proportion of missing data is very small and the dataset is large, mean imputation might not significantly affect the results.

Preliminary Analysis: For quick exploratory data analysis (EDA) where precision is not crucial, mean imputation can provide a rough idea of the dataset.

In Conclusion, while mean imputation is a quick and simple method for handling missing data, it is generally not recommended due to its potential to introduce bias and distort variability. More sophisticated imputation techniques like KNN, multiple imputation, or regression-based methods are preferred for better accuracy and reliability in data analysis.

Question no. 14

Answer:

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting straight line through the data points, which is represented by the equation:

$y= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$

where:

y is the dependent variable.

β_0 is the y-intercept.

β_1 β_2 ,…,βn are the coefficients of the independent variables x_1, x_2,…, x_n.

ϵ is the error term.

In simple linear regression, there is only one independent variable, and the equation simplifies to:

$y = \beta_0 + \beta_1 x + \epsilon$.

The goal of linear regression is to estimate the coefficients (β values) that minimize the sum of the squared differences between the observed values and the values predicted by the model.

This method is widely used for prediction, forecasting as well as for understanding the strength and type of relationships between variables.

Question no. 15

Answer: **Statistics** is a broad field with several branches, each focusing on different aspects of data collection, analysis, and interpretation.

Various branches of statistics are:

Descriptive Statistics: Summarizes and describes the main features of a data set. This includes measures of central tendency (mean, median, mode) and measures of variability (range, variance, standard deviation).

Inferential Statistics: Makes predictions or inferences about a population based on a sample of data. This includes hypothesis testing, confidence intervals, and regression analysis.

Probability Theory: Studies random events and the likelihood of their occurrence. It provides the mathematical foundation for inferential statistics.

Bayesian Statistics: Uses Bayes' theorem to update the probability of a hypothesis as more evidence becomes available. It contrasts with frequentist statistics, which only considers the probability of data given a hypothesis.

Biostatistics: Applies statistical methods to biological and health sciences. It is used in designing experiments, analyzing medical data, and interpreting public health data.

Econometrics: Applies statistical and mathematical methods to economic data to test hypotheses and forecast future trends.

Actuarial Science: Uses statistical and mathematical methods to assess risk in insurance, finance, and other industries.

Quality Control: Focuses on ensuring products and services meet certain quality standards, using statistical methods to monitor and control processes.

Environmental Statistics: Applies statistical techniques to environmental science, including the study of climate change, pollution, and natural resources.

Psychometrics: Focuses on the theory and technique of psychological measurement, including the measurement of knowledge, abilities, attitudes, and personality traits.

Social Statistics: Applies statistical methods to social science data, including sociology, political science, and demography.