Statistics Work book 1

Q.no.1 Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

Answer A) True

Q.no. 2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Answer a) Central Limit Theorem

Q.No.3 Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

Answer b) Modeling bounded count data

Q.No.4 Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

Answer c) The square of a standard normal random variable follows what is called chi-squared distribution

Statistics Work book 1

Q.No.5 _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer c) Poisson

Q.No.6 10 Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

Answer a) True

Q.No.7 Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Answer b) Hypothesis

Q.No.8 Normalized data are centered at_____and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

Answer a) 0

Q.No.9 Which of the following statement is incorrect with respect to outliers?

Statistics Work book 1

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

Answer c) Outliers cannot conform to the regression relationship

Q.no.10 What do you understand by the term Normal Distribution?

Answer The normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is symmetric around its mean. It is characterized by its bell-shaped curve, with the highest point at the mean and the curve tapering off symmetrically on both sides.

In a normal distribution:

- The mean (μ) represents the central tendency of the distribution.

- The standard deviation (σ) measures the spread or variability of the distribution.

- The shape of the curve is determined by the mean and standard deviation.

- Approximately 68% of the data falls within one standard deviation of the mean (μ ± σ), 95% within two standard deviations (μ ± 2σ), and 99.7% within three standard deviations (μ ± 3σ).

The normal distribution is widely used in statistics, probability theory, and various fields of science and engineering due to its mathematical tractability and its tendency to describe many natural phenomena and random processes. It serves as a foundation for many statistical methods and models, including hypothesis testing, confidence intervals, and regression analysis.

Q.No.11. How do you handle missing data? What imputation techniques do you recommend?

Answer Handling missing data is a crucial step in data preprocessing to ensure the quality and integrity of the dataset. Here are some common approaches to handling missing data and recommended imputation techniques:

1. Identify and Understand Missing Data : Begin by identifying missing values in the dataset and understanding the pattern and mechanism of missingness. Missing data can be categorized as Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR).

Statistics Work book 1

2. Deletion: Delete rows or columns with missing values. This approach is straightforward but may result in loss of valuable information, especially if missing values are not randomly distributed.

3. Mean/Median/Mode Imputation: Replace missing values with the mean, median, or mode of the corresponding feature. This method is simple and effective for numerical or categorical variables with missing values.

4. Forward Fill/Backward Fill : Use the value of the previous or next observation to fill missing values, particularly useful for time-series data where observations are ordered.

5. K-Nearest Neighbors (KNN) Imputation: Replace missing values with the average of k-nearest neighbors' values based on similarity metrics. KNN imputation can capture the local structure of the data and is effective for datasets with complex relationships.

6. Linear Regression Imputation: Predict missing values using linear regression models based on other variables in the dataset. This method is suitable for numerical variables with linear relationships.

7. Multiple Imputation: Generate multiple imputed datasets and combine results to obtain more robust estimates of missing values. Multiple imputation accounts for uncertainty associated with missing data and is suitable for datasets with complex dependencies.

8. Iterative Imputation: Iteratively impute missing values using statistical models such as linear regression, logistic regression, or decision trees. This method is useful for datasets with multiple variables and nonlinear relationships.

The choice of imputation technique depends on various factors such as the nature of the data, the extent of missingness, the distribution of missing values, and the assumptions about the missing data mechanism. It's essential to evaluate the performance of different imputation methods and consider the impact on downstream analyses or modeling tasks. Additionally, documentation of the imputation process is crucial for transparency and reproducibility of the results.

Q.No.12 What is A/B testing?

Answer A/B testing, also known as split testing or bucket testing, is a method used in marketing, product development, and user experience (UX) research to compare two versions of a web page, email, advertisement, or other digital assets to determine which one performs better.

Statistics Work book 1

A/B testing, two versions of a variable (e.g., a webpage layout, call-to-action button, email subject line) are compared by randomly assigning users or visitors to either version A or version B. Both versions are shown simultaneously to different groups of users, and their responses or behaviors are tracked and analyzed to determine which version performs better in achieving the desired outcome, such as click-through rate, conversion rate, or engagement rate.

The key steps involved in A/B testing include:

1. Identify the Objective: Clearly define the goal or metric you want to improve, such as increasing click-through rates, conversion rates, or revenue.

2. Create Variations: Develop two or more versions (A and B) of the element you want to test, with each version differing in one or more variables.

3. Random Assignment: Randomly assign users or visitors to either version A or version B, ensuring that each user has an equal chance of being exposed to either version.

4. Run the Experiment: Deploy both versions of the element simultaneously and collect data on user interactions, behaviors, or outcomes.

5. Measure Performance: Analyze the data collected from both versions to determine which version performs better in achieving the desired outcome. This is typically done using statistical hypothesis testing to determine if the differences in performance are statistically significant.

6. Draw Conclusions: Based on the results of the experiment, draw conclusions about which version is more effective and make data-driven decisions for future iterations or optimizations.

A/B testing allows businesses and organizations to make informed decisions based on empirical evidence rather than relying on intuition or assumptions. It helps optimize digital assets and improve user experience, ultimately leading to better engagement, conversions, and business outcomes.

Q.N.o 13 Is mean imputation of missing data acceptable practice?

Statistics Work book 1

Answer Mean imputation, where missing values are replaced with the mean of the observed values for that variable, is a simple and commonly used method for handling missing data. While it may be convenient and straightforward to implement, mean imputation has several limitations and potential drawbacks:

1. Alters Data Distribution: Mean imputation can introduce bias and alter the distribution of the variable, particularly if the missing values are not missing completely at random (MCAR). This can lead to inaccurate estimates of central tendency and variability.

2. Underestimates Variability: Mean imputation tends to underestimate the variability of the data because it replaces missing values with a single value (the mean). This can result in artificially narrowed confidence intervals and incorrect statistical inferences.

3. Distorts Relationships: Mean imputation can distort relationships between variables, especially if missing values are related to other variables in the dataset. Imputing missing values with the mean may obscure patterns and relationships present in the data.

4. Increases Correlation: Mean imputation can artificially inflate the correlation between variables, particularly if missing values are correlated with other variables in the dataset. This can lead to misleading interpretations of the data.

5. Doesn't Capture Uncertainty: Mean imputation assumes that the missing values are missing completely at random and ignores uncertainty associated with imputed values. It does not account for the variability or uncertainty in the imputed values.

Despite these limitations, mean imputation may still be acceptable under certain conditions, such as when the missing data are missing completely at random and the proportion of missing values is small. However, it's important to consider the potential impact of mean imputation on the validity and reliability of the analysis results and to explore alternative imputation methods when appropriate. Additionally, sensitivity analyses and robustness checks can help assess the robustness of the results to different imputation methods.

Q.No.14 What is linear regression in statistics?

Answer - Linear regression is a statistical method used to model the relationship between a dependent variable (often denoted as Y) and one or more independent variables (often denoted as X). It

Statistics Work book 1

assumes a linear relationship between the independent variables and the expected value of the dependent variable.

In simple linear regression, there is only one independent variable, X, and the relationship between X and Y is modeled using a straight line:

$$ Y = \beta_0 + \beta_1 X + \epsilon $$

Where:

- Y is the dependent variable (response variable).

- X is the independent variable (predictor variable).

- β_0 is the intercept (the value of Y when X is 0).

- β_1 is the slope (the change in Y for a one-unit change in X).

- ϵ is the error term (the difference between the observed value of Y and the predicted value of Y).

The goal of linear regression is to estimate the coefficients (β_0 and β_1) that best fit the observed data and minimize the sum of squared errors (SSE) between the observed and predicted values of Y. This is typically done using the method of least squares, which finds the line that minimizes the sum of the squared vertical distances between the observed data points and the line.

In multiple linear regression, there are multiple independent variables, and the relationship between Y and the independent variables is modeled using a linear equation:

$$ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon $$

Where:

- $X_1, X_2, \ldots, X_p$ are the independent variables.

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the coefficients.

- ϵ is the error term.

Statistics Work book 1

Linear regression is widely used for prediction, forecasting, and understanding the relationship between variables in various fields, including economics, finance, social sciences, and engineering.

Q.No.15 What are the various branches of statistics?

Answer - Statistics is a broad field that encompasses various branches and subfields, each focusing on different aspects of data analysis, inference, and interpretation. Some of the major branches of statistics include:

1. Descriptive Statistics: Descriptive statistics involves summarizing and describing the features of a dataset, such as measures of central tendency (e.g., mean, median, mode) and measures of variability (e.g., standard deviation, variance).

2. Inferential Statistics: Inferential statistics involves making inferences or predictions about a population based on sample data. It includes techniques such as hypothesis testing, confidence intervals, and regression analysis.

3. Probability Theory: Probability theory is the mathematical framework for quantifying uncertainty and randomness. It deals with the study of random variables, probability distributions, and stochastic processes.

4. Biostatistics: Biostatistics is the application of statistical methods to biological and health-related data. It involves designing experiments, analyzing clinical trials, and studying the effects of treatments or interventions on health outcomes.

5. Econometrics: Econometrics is the application of statistical methods to economic data. It involves analyzing economic relationships, forecasting economic variables, and testing economic theories using empirical data.

6. Social Statistics: Social statistics focuses on analyzing data related to social phenomena, such as demographics, education, crime, and public opinion. It involves studying patterns and trends in social data and making inferences about social processes.

7. Spatial Statistics: Spatial statistics deals with the analysis of spatially referenced data, such as geographical or environmental data. It includes techniques for spatial interpolation, spatial autocorrelation, and spatial regression.

Statistics Work book 1

8. Time Series Analysis: Time series analysis involves analyzing data collected over time to understand patterns, trends, and relationships. It includes techniques for forecasting, smoothing, and modeling time-varying phenomena.

9. Multivariate Statistics: Multivariate statistics deals with the analysis of datasets with multiple variables. It includes techniques for dimensionality reduction, cluster analysis, factor analysis, and discriminant analysis.

10. Bayesian Statistics: Bayesian statistics is a framework for statistical inference that uses probability distributions to represent uncertainty. It involves updating beliefs or probabilities based on new evidence and making decisions using Bayesian principles.

These are just a few of the major branches of statistics, and there are many other specialized areas within the field, such as environmental statistics, financial statistics, and industrial statistics, each focusing on specific applications and methodologies.