FLIP ROBO TECHNOLOGIES
STATISTICS

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

ANSWER -- TRUE

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 -- CENTRAL LIMIT THEOREM

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 -- MODELLING BOUNDED COUNT DATA

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 -- ALL OF THE MENTIONED

5._____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

ANSWER -- POISSON

6. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

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 -- HYPOTHESIS

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 -- 0

9. Which of the following statement is incorrect with respect to outliers?
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 -- OUTLIERS CANNOT CONFORM TO THE REGRESSION RELATIONSHIP

10. What do you understand by the term Normal Distribution?

ANSWER --
Normal distribution, also known as Gaussian distribution, is a continuous probability distribution characterized by its symmetric, bell-shaped curve. Here are some key features of normal distribution:

Symmetry: The curve is symmetric around its mean, meaning that the left and right sides of the curve are mirror images.

Mean, Median, and Mode: In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.

Standard Deviation: The spread of the distribution is determined by its standard deviation. A smaller standard deviation results in a steeper curve, while a larger standard deviation results in a flatter curve.

Empirical Rule: About 68% of the data falls within one standard deviation of the mean, approximately 95% falls within two standard deviations, and about 99.7% falls

within three standard deviations. This is often referred to as the 68-95-99.7 rule.

11. How do you handle missing data? What imputation techniques do you recommend?

ANSWER --
Handling missing data is crucial in data analysis, as it can significantly impact the results and conclusions. Here are several strategies and techniques for dealing with missing data, including various imputation methods:

1. Understanding the Missing Data Mechanism
Missing Completely at Random (MCAR): The missingness is unrelated to both observed and unobserved data.
Missing at Random (MAR): The missingness is related to observed data but not to the missing data itself.
Missing Not at Random (MNAR): The missingness is related to the unobserved data.
Understanding the type of missingness helps determine the appropriate handling method.

2. Imputation Techniques
Mean/Median/Mode Imputation:

Replace missing values with the mean (for continuous data), median (to reduce the effect of outliers), or mode (for categorical data).
Pros: Simple to implement.
Cons: Reduces variability; can distort relationships between variables.
Forward/Backward Fill:

In time series data, replace missing values with the next or previous observation.
Pros: Maintains the time order.
Cons: May not be appropriate if the data is not time-related.
K-Nearest Neighbors (KNN) Imputation:

Use the average of the nearest neighbors' values to impute missing data.
Pros: Takes into account the similarity between data points.
Cons: Computationally expensive; requires careful selection of
k
k.
Multiple Imputation:

Create multiple datasets by imputing missing values in several different ways, analyze each dataset separately, and then combine the results.
Pros: Provides a more robust estimate of uncertainty.
Cons: More complex and computationally intensive.
Regression Imputation:

Use regression models to predict missing values based on other available data.
Pros: Utilizes relationships between variables.
Cons: May lead to underestimation of variability; can introduce bias if the model is not well-specified.

Interpolation:

Estimate missing values within the range of existing data points, often used in time series.
Pros: Maintains trends in data.
Cons: May not accurately represent extreme values.
Last Observation Carried Forward (LOCF):

Commonly used in longitudinal studies, it fills in missing values with the last observed value.
Pros: Simple and maintains some data continuity.
Cons: Can introduce bias if the data is not stable.
Use of Machine Learning Models:

Employ algorithms like Random Forest or Neural Networks to predict missing values based on other features in the dataset.
Pros: Can capture complex relationships.
Cons: Requires more computational resources and careful tuning.

3. Other Considerations
Deletion: If the amount of missing data is small, one might consider deleting records with missing values (listwise or pairwise deletion).
Data Augmentation: In certain cases, adding new data through surveys or additional studies can help mitigate missingness.
Sensitivity Analysis: After imputation, it’s important to perform sensitivity analyses to see how different methods of handling missing data affect the results

12. What is A/B testing?

ANSWER--
A/B testing, also known as split testing or bucket testing, is a statistical method used to compare two versions of a variable (such as a webpage, app feature, or marketing campaign) to determine which one performs better. Here’s a detailed overview of A/B testing:

Key Components of A/B Testing
Control and Variation:

Control (A): The original version or current state of the variable.
Variation (B): The modified version that includes changes meant to improve performance.
Hypothesis Formation:

Before conducting an A/B test, a hypothesis is formed about how the change will affect user behavior. This hypothesis is often based on prior research or intuition.
Random Assignment:

Participants (users, customers, etc.) are randomly assigned to either the control group or the variation group. This randomization helps ensure that the results are statistically valid and not biased by external factors.

Data Collection:

During the test, relevant metrics are collected for both groups. These metrics can include conversion rates, click-through rates, sales, user engagement, or other key performance indicators (KPIs).
Statistical Analysis:

After a predetermined duration, the data is analyzed using statistical methods to determine if there are significant differences between the performance of the two versions. Common statistical tests include t-tests or chi-square tests.
Decision Making:

Based on the analysis, a decision is made whether to implement the change (variation) or retain the original version (control). If the variation shows statistically significant improvement, it may be adopted.
Benefits of A/B Testing
Data-Driven Decisions: A/B testing allows organizations to make informed decisions based on actual user behavior rather than assumptions.
Improved Performance: By systematically testing changes, businesses can optimize their products, services, and marketing strategies for better performance.
Reduced Risk: Implementing changes based on A/B test results reduces the risk associated with rolling out new features or campaigns without evidence of their effectiveness.
Common Use Cases
Web Design: Testing different layouts, colors, or content on a webpage to see which version results in higher engagement or conversion rates.
Email Marketing: Comparing subject lines, content, or calls to action in email campaigns to identify which approach yields better open or click rates.
Product Features: Testing new features in a software application to determine if they improve user satisfaction or increase usage.
Advertising: Evaluating different ad creatives or targeting strategies to see which generates more leads or sales.

13. Is mean imputation of missing data acceptable practice?

ANSWER--
Mean imputation is a commonly used method for handling missing data, but whether it is an acceptable practice depends on the context and the specific characteristics of the data. Here are some pros and cons of mean imputation, along with considerations for its use:

Pros of Mean Imputation
Simplicity: Mean imputation is easy to implement and understand. It involves replacing missing values with the mean of the observed data, making it a straightforward method for dealing with missingness.

Preservation of Data Size: By filling in missing values, mean imputation maintains the dataset's size, allowing for the inclusion of all available observations in analyses.

Speed: The process of calculating the mean and replacing missing values is computationally efficient, making it suitable for large datasets.

Cons of Mean Imputation
Loss of Variability: Mean imputation reduces the natural variability in the data. Since all missing values are replaced with the same mean, it can lead to an underestimation of the standard deviation and variance, affecting statistical analyses and interpretations.

Bias Introduction: If the data is not missing completely at random (MCAR), mean imputation can introduce bias. For example, if higher values are more likely to be missing, imputing with the mean can skew the dataset lower.

Distortion of Relationships: Mean imputation can distort relationships between variables, potentially affecting correlation and regression analyses. The method assumes that the missing values are similar to the observed values, which may not always be true.

Ignoring the Data Structure: Mean imputation does not consider the relationships among other variables. It treats the imputed values as if they are known data, which can mislead analyses.

When to Use Mean Imputation
When the Proportion of Missing Data is Small: If only a small percentage of the data is missing, mean imputation may not significantly distort the results.
When the Data is MCAR: If it can be reasonably assumed that the data is missing completely at random, mean imputation may be more acceptable.
As a First Step: In some cases, mean imputation can be used as a preliminary step before applying more sophisticated imputation techniques or analyses.
Alternatives to Mean Imputation
Given the limitations of mean imputation, several alternative methods are often recommended:

Median Imputation: This is less affected by outliers and can be a better option for skewed data.
K-Nearest Neighbors (KNN) Imputation: This considers the values of similar observations to impute missing values.
Multiple Imputation: This method creates several imputed datasets and combines results to reflect uncertainty about the missing data.
Regression Imputation: Using relationships with other variables to predict and fill in missing values

14. What is linear regression in statistics?

ANSWER--
Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the response or outcome variable) and one or more independent variables (also known as predictors or features). The goal is to find the best-fitting line (or hyperplane in the case of multiple regression) that describes how the independent variables influence the dependent variable.

Key Assumptions of Linear Regression
Linearity: The relationship between the independent and dependent variables is linear.

Independence: The residuals (errors) are independent. There should be no correlation between the errors.

Homoscedasticity: The residuals have constant variance at every level of the independent variable(s).

Normality: The residuals should be approximately normally distributed, particularly for small sample sizes.

Estimation of Coefficients
The coefficients (β β values) are typically estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of the squared differences between the observed values and the values predicted by the linear model.

Evaluation of the Model
Once a linear regression model is built, several metrics are used to evaluate its performance, including:

R-squared : Represents the proportion of variance in the dependent variable that can be explained by the independent variables. Values range from 0 to 1, with higher values indicating a better fit.

Adjusted R-squared: Adjusts

for the number of predictors in the model, providing a more accurate measure when comparing models with different numbers of predictors.

p-values: Indicate the statistical significance of each independent variable's coefficient. A low p-value (typically < 0.05) suggests that the variable is a significant predictor of the dependent variable.

Residual Analysis: Analyzing residuals helps assess the validity of the model's assumptions, including linearity and homoscedasticity.

15. What are the various branches of statistics?

ANSWER--
Statistics is a broad field with various branches that focus on different aspects of data collection, analysis, interpretation, and presentation. Here are the main branches of statistics:

1. Descriptive Statistics
Definition: This branch focuses on summarizing and organizing data to describe its main features.
Key Concepts: Measures of central tendency (mean, median, mode), measures of variability (range, variance, standard deviation), and graphical representations (histograms, pie charts, box plots).
2. Inferential Statistics
Definition: Inferential statistics involves making predictions or generalizations about a population based on a sample of data drawn from that population.
Key Concepts: Hypothesis testing, confidence intervals, p-values, and regression analysis.
3. Probability Theory
Definition: This branch studies the likelihood of different outcomes in uncertain situations.
Key Concepts: Probability distributions (normal, binomial, Poisson), random variables, and laws of probability (addition and multiplication rules).
4. Biostatistics
Definition: Biostatistics applies statistical methods to biological and health-related processes.
Key Concepts: Clinical trials, epidemiological studies, survival analysis, and genetic data analysis.
5. Econometrics
Definition: This branch combines statistical methods with economic theory to analyze economic data.
Key Concepts: Regression models, time series analysis, and economic forecasting.
6. Psychometrics
Definition: Psychometrics focuses on the theory and technique of psychological measurement, including the development and validation of measurement instruments.
Key Concepts: Reliability and validity, factor analysis, and item response theory.
7. Quality Control and Industrial Statistics
Definition: This branch uses statistical methods to monitor and improve processes in manufacturing and service industries.
Key Concepts: Control charts, process capability analysis, and Six Sigma methodologies.
8. Multivariate Statistics
Definition: Multivariate statistics deals with the analysis of data that involves multiple variables simultaneously.
Key Concepts: Principal component analysis (PCA), factor analysis, cluster analysis, and discriminant analysis.
9. Nonparametric Statistics
Definition: This branch includes statistical methods that do not assume a specific distribution for the data.
Key Concepts: Wilcoxon tests, Kruskal-Wallis tests, and chi-square tests.
10. Spatial Statistics
Definition: Spatial statistics involves analyzing spatial and geographical data to understand patterns and relationships in a spatial context.
Key Concepts: Geographic Information Systems (GIS), spatial autocorrelation, and kriging.
11. Time Series Analysis
Definition: This branch focuses on analyzing time-ordered data to identify trends,

cycles, and seasonal variations.
Key Concepts: Autoregressive integrated moving average (ARIMA) models, exponential smoothing, and forecasting.
12. Survival Analysis
Definition: Survival analysis deals with time-to-event data, often used in medical research to analyze the time until an event of interest occurs.
Key Concepts: Kaplan-Meier estimator, Cox proportional hazards model, and hazard functions.