STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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

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

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

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

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

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

7. 1. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10

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:-

1 - a	2 - a	3 - b	4 – a, c
5 - c	6 - b	7 – b	8 - a
9 - d			

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly

10. What do you understand by the term Normal Distribution?
Answer:- The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetrical and bell-shaped. It describes how the values of a variable are distributed. Here are some key characteristics and concepts related to the normal distribution:

1. Symmetry:- The distribution is symmetric around its mean. This means the left and right sides of the distribution are mirror images of each other.
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 graph of the normal distribution is bell-shaped, with most of the data points clustering around the mean. The frequencies of values decrease as you move away from the mean in either direction.

4. Standard Deviation:- The spread of the distribution is determined by the standard deviation. A smaller standard deviation results in a steeper, narrower bell curve, while a larger standard deviation produces a wider, flatter curve.

5. Empirical Rule (68-95-99.7 Rule):-
- Approximately 68% of the data within a normal distribution lies within one standard deviation of the mean.
- About 95% lies within two standard deviations.
- Around 99.7% lies within three standard deviations.

6. Probability Density Function (PDF):- The normal distribution is defined by its probability density function.
7. Standard Normal Distribution:- A special case of the normal distribution is the standard normal distribution, which has a mean of 0 and a standard deviation of 1. It is used as a reference to compare other normal distributions.

The normal distribution is widely used in statistics and various fields because many variables are naturally distributed this way, and it has several useful properties for inferential statistics.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer:- Handling missing data is crucial for maintaining the integrity of data analysis. Here are several strategies and imputation techniques commonly used to handle missing data:
Strategies for Handling Missing Data

1. Remove Missing Data:
- Listwise Deletion: Remove entire rows with any missing values. This is simple but can result in significant data loss.
- Pairwise Deletion: Use all available data for each analysis without removing entire rows. This preserves more data but can be complex to implement and interpret.

2. Keep Missing Data as a Separate Category:
- Useful for categorical variables where "missing" itself can be an informative category.
3. Fill with a Constant Value:
- Replace missing values with a specific constant, like zero or a large number. This is generally not recommended unless the constant has a meaningful interpretation in context.
Imputation Techniques
1. Mean/Median/Mode Imputation:
- Replace missing values with the mean, median, or mode of the non-missing values.
- Pros: Simple and quick.
- Cons: Can distort the variance and reduce variability.

2. K-Nearest Neighbors (KNN) Imputation:
   - Replace missing values with the values from the nearest neighbors.
   - Pros: Can handle nonlinear relationships and maintain variability.
   - Cons: Computationally intensive, especially with large datasets.

3. Multivariate Imputation by Chained Equations (MICE):
   - Use other variables in the dataset to predict and fill in missing values iteratively.
   - Pros: Creates multiple imputed datasets, captures the uncertainty of missing data.
   - Cons: Complex to implement, requires sufficient computational resources.

4. Regression Imputation:
   - Predict missing values using regression models based on other variables.
   - Pros: Utilizes relationships between variables.
   - Cons: Can underestimate variability if not handled properly.

5. Hot Deck Imputation:
   - Replace missing values with observed responses from similar units.
   - Pros: Maintains data distribution.
   - Cons: Can be challenging to define "similar" units.

6. Multiple Imputation:
   - Create several different plausible imputed datasets and combine results from each.
   - Pros: Reflects the uncertainty about the right value to impute, leading to more robust statistical inferences.

   - Cons: More complex and computationally intensive.

Advanced Techniques

1. Machine Learning Models:
   - Use models like Random Forests, Gradient Boosting Machines, or Neural Networks to predict missing values.
   - Pros: Can capture complex patterns and interactions in the data.
   - Cons: Requires more computational resources and can be more complex to interpret.

2. Time Series Specific Methods:
   - For time series data, use techniques like forward-fill, backward-fill, or interpolation.
   - Pros: Takes into account the temporal order of data.
   - Cons: Can introduce biases if trends or seasonality are not considered.

12. What is A/B testing?

Answer:- A/B testing, also known as split testing, is a method used to compare two versions of a webpage, app, or other user experience to determine which one performs better. This technique is widely used in marketing, user experience (UX) design, and product development to make data-driven decisions. Here's a detailed breakdown of A/B testing:

Steps Involved in A/B Testing

1. Identify the Objective: Define the goal you want to achieve with the test. This could be increasing click-through rates, conversions, sign-ups, or any other key performance indicator (KPI).

2. Develop Hypotheses: Formulate hypotheses about what changes might improve the metric you're focusing on. For example, you might hypothesize that changing the color of a call-to-action button will increase click rates.

3. Create Variants: Develop two versions of the content to be tested:
   - A (Control): The original version.
   - B (Variation): The modified version with the changes you hypothesized will improve performance.

4. Randomly Assign Users: Randomly divide your audience into two groups. One group is shown version A, and the other is shown version B.

5. Run the Test: Allow the test to run for a sufficient period to collect enough data. The duration should be long enough to achieve statistical significance and account for any external factors that could influence the results.

6. Measure Results: Compare the performance of the two versions based on the predefined metrics. Use statistical analysis to determine if any observed differences are statistically significant.

7. Analyze and Implement: If the variation (B) shows a significant improvement over the control (A), implement the changes. If not, you may need to revisit your hypotheses and test different variations.

Key Concepts in A/B Testing

- Statistical Significance: A measure of whether the results observed are likely due to the changes made rather than random chance. Common significance levels used are 0.05 or 0.01.
- Sample Size: The number of users or observations included in the test. Larger sample sizes increase the reliability of the results.

- Conversion Rate: The percentage of users who complete the desired action (e.g., clicking a button, making a purchase) out of the total number of users exposed to the test.
- Confidence Interval: A range of values that is likely to contain the true effect of the variation, providing an estimate of the uncertainty around the results.

Benefits of A/B Testing

- Data-Driven Decisions: Enables decisions based on actual user behavior and data rather than assumptions or intuition.
- Improved User Experience: Helps optimize various elements of a user interface to enhance overall user satisfaction and engagement.
- Increased Conversion Rates: Identifies the most effective changes to boost key performance metrics, leading to better business outcomes.
- Risk Mitigation: Allows for testing changes on a subset of users before rolling out to the entire audience, minimizing potential negative impacts.

13. Is mean imputation of missing data acceptable practice?
Answer:- Mean imputation, where missing values are replaced with the mean of the observed data, is a common and straightforward method for handling missing data. However, its acceptability depends on the context and the specific characteristics of the data. Here are some pros and cons to consider:

Pros of Mean Imputation

1. Simplicity: Easy to implement and understand.
2. Maintains Sample Size: Keeps the dataset size intact, which can be beneficial for analysis.
3. Fast: Computationally inexpensive, making it suitable for large datasets.
Cons of Mean Imputation

1. Underestimates Variability: Reduces the natural variability in the data by introducing a constant value, which can lead to biased statistical estimates.
2. Distorts Distribution: Can distort the distribution of the data, especially if the data are not normally distributed.
3. Ignores Relationships: Does not account for relationships between variables, which can result in misleading conclusions.
4. Bias: Can introduce bias, especially if the data are not missing completely at random (MCAR).
When Mean Imputation Might Be Acceptable

1. Small Proportion of Missing Data: If the proportion of missing data is very small, mean imputation might not significantly impact the results.

2. Non-Critical Analysis: For preliminary analysis or non-critical exploratory data analysis (EDA), mean imputation might be a practical choice.
3. Symmetric Distributions: If the data are approximately normally distributed, the impact of mean imputation might be less severe.

While mean imputation can be acceptable in some situations, it is generally not recommended for most analyses due to its limitations. More advanced imputation techniques, such as multiple imputation or KNN imputation, are usually preferred because they preserve the relationships in the data and provide more accurate and unbiased estimates.

14. What is linear regression in statistics?
Answer:- Linear regression is a fundamental statistical method used to model and analyze the relationships between a dependent variable and one or more independent variables. The primary objective of linear regression is to predict the value of the dependent variable based on the values of the independent variables.

Key Concepts of Linear Regression

1. Dependent and Independent Variables:
- Dependent Variable (Y): The variable you are trying to predict or explain.
- Independent Variable(s) (X): The variable(s) used to make predictions about the dependent variable.

2. Linear Relationship:
- Assumes that there is a linear relationship between the dependent and independent variables, which can be represented as a straight line in a scatter plot of the data.

3. Regression Line:
- The regression line (also called the line of best fit) is the line that best represents the data points in a scatter plot. It minimizes the sum of the squared differences (residuals) between observed values and the values predicted by the line.

Steps in Performing Linear Regression

1. Data Collection: Gather data on the dependent variable and independent variables.
2. Exploratory Data Analysis (EDA): Visualize the data using scatter plots to check for linear relationships and identify potential outliers.
3. Model Fitting: Use statistical software or methods like Ordinary Least Squares (OLS) to fit the linear regression model to the data.
4. Assumption Checking: Validate the assumptions of linear regression, which include:
- Linearity: The relationship between the dependent and independent variables is linear.

- Independence: Observations are independent of each other.
- Homoscedasticity: Constant variance of residuals across all levels of the independent variables.
- Normality: Residuals (errors) are normally distributed.

5. Model Evaluation: Assess the model's performance using metrics such as R-squared, Adjusted R-squared, Mean Squared Error (MSE), and p-values for significance testing of coefficients.
6. Prediction and Interpretation: Use the model to make predictions and interpret the coefficients to understand the relationship between variables.

Applications of Linear Regression

- Economics: Forecasting economic indicators like GDP, inflation, and unemployment rates.
- Finance: Modeling relationships between stock prices and various financial metrics.
- Medicine: Predicting patient outcomes based on clinical measurements and patient characteristics.
- Marketing: Analyzing the impact of advertising spend on sales revenue.
- Engineering: Estimating the effect of design parameters on product performance.

Linear regression is a powerful and widely used statistical tool for understanding and predicting the relationships between variables. Its simplicity and interpretability make it a fundamental method in many fields of research and application.

15. What are the various branches of statistics?
Answer:- Statistics is a broad field that encompasses various methodologies and techniques for collecting, analyzing, interpreting, and presenting data. It can be divided into several branches, each with its specific focus and applications. Here are the primary branches of statistics:

1. Descriptive Statistics

Descriptive statistics involve summarizing and organizing data to describe the main features of a dataset. This branch provides simple summaries and visualizations, such as graphs and tables, to present data in an informative way. Key concepts include:

- Measures of Central Tendency: Mean, median, mode.
- Measures of Dispersion: Range, variance, standard deviation, interquartile range.
- Graphical Representations: Histograms, bar charts, pie charts, box plots.

2. Inferential Statistics

Inferential statistics involve making inferences and predictions about a population based on a sample of data. This branch uses probability theory to draw conclusions and make decisions under uncertainty. Key concepts include:

- Estimation: Point estimates and confidence intervals.
- Hypothesis Testing: Null and alternative hypotheses, p-values, significance levels.
- Regression Analysis: Simple and multiple linear regression, logistic regression.
- ANOVA (Analysis of Variance): Testing differences between group means.

3. Probability Theory

Probability theory is the mathematical foundation of statistics, focusing on the analysis of random phenomena and the likelihood of events occurring. Key concepts include:

- Probability Distributions: Binomial, Poisson, normal, exponential distributions.
- Random Variables: Discrete and continuous random variables.
- Expected Value and Variance: Measures of the center and spread of a probability distribution.
- Law of Large Numbers and Central Limit Theorem: Theoretical underpinnings of inferential statistics.

4. Experimental Design

Experimental design involves planning experiments to ensure that the data collected can provide valid and objective conclusions. This branch focuses on controlling variability and bias to establish cause-and-effect relationships. Key concepts include:

- Randomization: Randomly assigning subjects to treatments.
- Replication: Repeating experiments to ensure reliability.
- Blocking: Controlling for variables that might affect the outcome.
- Factorial Designs: Studying the effects of multiple factors simultaneously.

5. Bayesian Statistics

Bayesian statistics involves updating the probability of a hypothesis as more evidence or information becomes available. This approach incorporates prior knowledge or beliefs in addition to the data. Key concepts include:

- Bayes' Theorem: A mathematical formula for updating probabilities.
- Prior, Posterior, and Likelihood: Components of Bayesian inference.
- Bayesian Networks: Graphical models representing probabilistic relationships.

6. Multivariate Statistics

Multivariate statistics involve analyzing data that includes multiple variables to understand their relationships and effects. This branch is used in many fields to analyze complex data structures. Key concepts include:

- Principal Component Analysis (PCA): Reducing the dimensionality of data.
- Factor Analysis: Identifying underlying factors that explain the data.
- Cluster Analysis: Grouping similar observations.
- Canonical Correlation Analysis: Exploring relationships between sets of variables.

7. Nonparametric Statistics

Nonparametric statistics involve methods that do not assume a specific probability distribution for the data. These techniques are useful when data do not meet the assumptions required for parametric tests. Key concepts include:

- Rank-Based Tests: Wilcoxon signed-rank test, Mann-Whitney U test.
- Resampling Methods: Bootstrap, permutation tests.
- Chi-Square Tests: Tests for independence and goodness-of-fit.

8. Time Series Analysis

Time series analysis involves analyzing data collected over time to identify trends, seasonal patterns, and other temporal dynamics. This branch is essential for forecasting and modeling time-dependent processes. Key concepts include:

- Autoregressive Models (AR): Modeling the relationship between an observation and a number of lagged observations.
- Moving Average Models (MA): Modeling the relationship between an observation and a residual error from a moving average model.
- ARIMA (AutoRegressive Integrated Moving Average): Combining AR and MA models with differencing to make data stationary.
- Seasonal Decomposition: Breaking down time series data into trend, seasonal, and residual components.

9. Survey Sampling

Survey sampling involves designing and analyzing surveys to collect data from a subset of a population, with the aim of making inferences about the entire population. Key concepts include:

- Sampling Methods: Simple random sampling, stratified sampling, cluster sampling.
- Sampling Bias: Understanding and minimizing bias in survey data.
- Survey Weighting: Adjusting sample data to represent the population accurately.

10. Survival Analysis
Survival analysis involves analyzing time-to-event data, often used in medical research to study the time until the occurrence of an event of interest (e.g., death, disease recurrence). Key concepts include:
- Survival Function: Probability of surviving beyond a certain time.
- Hazard Function: Instantaneous rate of event occurrence.
- Cox Proportional Hazards Model: A regression model for survival data.

These branches of statistics provide a comprehensive toolkit for understanding, analyzing, and interpreting data in various fields, from scientific research to business analytics.