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

Ans: a) 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

Ans: a) 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

Ans: b) Modeling 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

Ans: c) **Correct**

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

Ans: c) Poisson

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

Ans: 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

Ans: b) 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

Ans: a) 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

Ans: c) Outliers cannot conform to the regression relationship

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

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

Ans: Normal Distribution, often referred to as a Gaussian distribution, is a continuous probability distribution characterized by its bell-shaped curve. Here are the key features and properties of the normal distribution:

1. **Symmetry**: The normal distribution is symmetric about its mean. This means that the left and right halves of the curve are mirror images of each other.
2. **Mean, Median, and Mode**: In 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, indicating that data points closer to the mean are more frequent than those further away.
4. **Standard Deviation**: The spread of the distribution is determined by the standard deviation. A smaller standard deviation results in a steeper curve, while a larger standard deviation results in a flatter curve.
5. **68-95-99.7 Rule**: In a normal distribution:
    - Approximately 68% of the data falls within one standard deviation of the mean.
    - About 95% falls within two standard deviations.
    - Around 99.7% falls within three standard deviations.
6. **Central Limit Theorem**: The normal distribution plays a crucial role in statistics, particularly in the Central Limit Theorem, which states that the distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the original distribution of the population.

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

Ans: Handling missing data is an important aspect of data analysis, as missing values can lead to biased results and reduced statistical power. Here are some common strategies and imputation techniques for dealing with missing data:

1. Understanding the Missing Data Mechanism

- **Missing Completely at Random (MCAR)**: The missingness is unrelated to the data itself.
- **Missing at Random (MAR)**: The missingness is related to observed data but not the missing data.
- **Missing Not at Random (MNAR)**: The missingness is related to the unobserved data.

2. Imputation Techniques

Here are some common imputation techniques:

a) Mean/Median/Mode Imputation

- **Mean Imputation**: Replace missing values with the mean of the available values for that feature. This is useful for numerical data.
- **Median Imputation**: Replace missing values with the median, which is more robust to outliers.
- **Mode Imputation**: Replace missing values with the mode for categorical data.

b) Forward/Backward Fill

- **Forward Fill**: Replace missing values with the last observed value (useful in time series data).
- **Backward Fill**: Replace missing values with the next observed value.

c) K-Nearest Neighbors (KNN) Imputation

- This method uses the k-nearest neighbors to impute missing values based on the values of similar observations. It can capture local patterns in the data.

d) Regression Imputation

- Use regression models to predict and impute missing values based on other available variables. This involves building a regression model using complete cases and then predicting the missing values.

e) Multiple Imputation

- This approach involves creating multiple datasets by imputing missing values several times, then analyzing each dataset and combining the results. This method accounts for uncertainty about the missing data.

f) Hot Deck Imputation

- In this method, missing values are imputed using observed responses from similar units in the dataset. This is often done in surveys.

g) Interpolation

- This technique estimates missing values based on the values surrounding them in a time series or sequential dataset.

3. Deleting Missing Data

- **Listwise Deletion**: Remove any rows with missing values. This can lead to loss of data and bias if the missingness is not MCAR.
- **Pairwise Deletion**: Use all available data without excluding entire cases, which can be beneficial in some analyses.

4. Using Machine Learning Algorithms

- Some machine learning algorithms can handle missing values internally (e.g., decision trees, random forests), allowing you to bypass imputation altogether.

5. Evaluate the Impact of Imputation

- Always assess how your chosen imputation method impacts the analysis. This can be done by comparing the results of analyses using imputed datasets versus complete-case analyses.

12. What is A/B testing?

A/B testing, also known as split testing, is a method of comparing two versions of a webpage, app, or other product to determine which one performs better in terms of a specific metric. The goal is to identify changes that can improve user experience, engagement, or conversion rates. Here's a detailed overview of A/B testing:

Key Concepts of A/B Testing:

1. **Control and Variant**:
    - **Control (A)**: This is the original version of the product, webpage, or feature.
    - **Variant (B)**: This is the modified version that includes one or more changes designed to improve performance.
2. **Random Assignment**:
    - Participants (users, visitors, etc.) are randomly assigned to either the control group or the variant group. This randomization helps ensure that any differences in outcomes can be attributed to the changes made, rather than other factors.
3. **Performance Metrics**:
    - A/B testing requires clear metrics to evaluate success. Common metrics include:
        - Conversion rate (e.g., purchases, sign-ups)
        - Click-through rate (CTR)
        - Bounce rate
        - Time spent on the page
        - Any other relevant user engagement metric
4. **Statistical Analysis**:
    - After running the test for a predetermined period, the data collected is analyzed using statistical methods to determine whether any observed differences in performance are statistically significant. This often involves hypothesis testing, confidence intervals, and p-values.
5. **Sample Size**:
    - A/B tests require a sufficient sample size to ensure reliable results. Sample size calculations should be performed before the test to avoid inconclusive results.
6. **Duration**:
    - The duration of the test should be long enough to gather meaningful data but also consider seasonal variations in user behavior.

Steps in Conducting A/B Testing:

1. **Define the Goal**: Determine what you want to test and the desired outcome (e.g., increased sales, higher engagement).
2. **Create Hypotheses**: Formulate a hypothesis about how the changes will impact the performance metric.
3. **Develop Variants**: Create the variant version(s) with specific changes (e.g., changing a button color, modifying text).

4. **Split Traffic**: Randomly assign users to the control or variant group.
5. **Collect Data**: Monitor user interactions and collect data on the defined performance metrics.
6. **Analyze Results**: Use statistical analysis to compare the performance of the control and variant versions.
7. **Make Decisions**: Based on the analysis, decide whether to implement the changes from the variant version, revert to the original, or continue testing.
8. **Iterate**: A/B testing is an iterative process. Continue testing new ideas to optimize performance over time.

Benefits of A/B Testing:

- **Data-Driven Decisions**: A/B testing provides empirical evidence to guide decisions, reducing reliance on assumptions or intuition.
- **Optimization**: It allows for systematic optimization of user experiences, leading to improved conversion rates and user satisfaction.
- **Risk Reduction**: Testing changes on a small segment of users before a full rollout minimizes the risk of negative impacts on overall performance.

Use Cases for A/B Testing:

- Website design (e.g., layout, colors, call-to-action buttons)
- Email marketing (e.g., subject lines, content layout)
- Ad campaigns (e.g., visuals, messaging)
- Product features (e.g., user interface changes)

13. Is mean imputation of missing data acceptable practice?

Ans: Mean imputation is a common technique for handling missing data, but it has several important considerations that affect its acceptability and effectiveness. Here's a breakdown of the advantages and disadvantages of mean imputation:

Advantages of Mean Imputation:

1. **Simplicity**: Mean imputation is straightforward to implement and understand. It requires minimal computational effort and can be done quickly.
2. **Preservation of Sample Size**: By filling in missing values, mean imputation allows for the retention of all observations in the dataset, which can be particularly useful in small datasets.
3. **No Loss of Data**: Since it replaces missing values rather than removing entire rows, it preserves the overall structure of the dataset.

Disadvantages of Mean Imputation:

1. **Bias**: Mean imputation can introduce bias into the dataset, especially if the missing data is not missing completely at random (MCAR). If the missingness is related to the values themselves, this can skew results.
2. **Reduced Variability**: By replacing missing values with the mean, the overall variability of the dataset is reduced. This can lead to underestimating the standard deviation and other statistical metrics.
3. **Distortion of Relationships**: Mean imputation can distort the relationships between variables, particularly in regression analyses. It assumes that the missing values are randomly distributed around the mean, which may not be true.
4. **Ignores Other Information**: Mean imputation does not take into account other available information in the dataset that might provide a more accurate estimate of the missing values.

When is Mean Imputation Acceptable?

- **Small Amount of Missing Data**: If the proportion of missing data is very small (e.g., less than 5%), mean imputation may be acceptable as a quick fix, although more robust methods are generally recommended.
- **Data is MCAR**: If you can reasonably assume that the missing data is completely at random, mean imputation might be less problematic.
- **Non-critical Analysis**: In exploratory data analysis or scenarios where high accuracy is not critical, mean imputation can be a useful initial approach.

Alternatives to Mean Imputation:

Given the drawbacks of mean imputation, consider these alternative methods for handling missing data:

1. **Median Imputation**: More robust to outliers and skewed distributions than mean imputation.
2. **K-Nearest Neighbors (KNN) Imputation**: Uses the values of similar observations to impute missing values.
3. **Regression Imputation**: Predicts missing values based on the relationships in the dataset.
4. **Multiple Imputation**: Creates multiple datasets with different imputations, accounting for uncertainty

and variability.

5. **Machine Learning Algorithms**: Some models can handle missing values internally without requiring imputation.

14. What is linear regression in statistics?
Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is called "linear" because it assumes a linear relationship between the dependent and independent variables.

Key Concepts in Linear Regression:

1. **Dependent Variable (Y):** The outcome or the variable being predicted or explained.
2. **Independent Variable (X):** The predictor or explanatory variable used to predict the dependent variable.
3. **Linear Relationship:** The relationship between the dependent and independent variables is modeled as a straight line.

Types of Linear Regression:

1. **Simple Linear Regression:** Involves one dependent variable and one independent variable.
2. **Multiple Linear Regression:** Involves one dependent variable and two or more independent variables.

Objectives and Assumptions:

- **Objective:** The primary objective of linear regression is to find the line (or hyperplane in the case of multiple regression) that best fits the data, minimizing the sum of the squared differences between the observed and predicted values of the dependent variable.
- **Assumptions:**
  - **Linearity:** The relationship between the dependent and independent variables is linear.
  - **Independence:** Observations are independent of each other.
  - **Homoscedasticity:** The variance of error terms is constant across all levels of the independent variables.
  - **Normality:** The error terms are normally distributed.

Estimation and Evaluation:

- **Estimation:** The coefficients (\(\beta_0, \beta_1, \ldots, \beta_n\)) are typically estimated using the method of least squares, which minimizes the sum of the squared residuals (the differences between observed and predicted values).
- **Evaluation:**
  - **R-squared (\(R^2\)):** Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
  - **Adjusted R-squared:** Adjusted for the number of predictors in the model, providing a more accurate measure for models with multiple predictors.
  - **F-test:** Tests the overall significance of the model.
  - **t-tests:** Test the significance of individual coefficients.

Linear regression is a foundational technique in statistics and machine learning, widely used for predictive modeling, hypothesis testing, and understanding relationships between variables.

15. What are the various branches of statistics?

Ans: Statistics is a broad field with various branches that focus on different aspects of data collection, analysis, interpretation, and presentation. The primary branches of statistics are:

1. Descriptive Statistics

Descriptive statistics summarize and describe the features of a dataset. They provide simple summaries about the sample and the measures. Key concepts include:

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

2. Inferential Statistics

Inferential statistics use a random sample of data taken from a population to describe and make inferences about the population. This branch involves:

- **Hypothesis Testing:** Procedures to test assumptions (e.g., t-tests, chi-square tests, ANOVA).
- **Confidence Intervals:** Range of values used to estimate the true value of a population parameter.
- **Regression Analysis:** Modeling the relationship between dependent and independent variables.
- **Estimation:** Point and interval estimation of population parameters.

3. Probability Theory

Probability theory underpins statistical methods by providing the mathematical foundation for understanding randomness and uncertainty. Key concepts include:

- **Probability Distributions:** Normal, binomial, Poisson distributions, etc.
- **Random Variables:** Discrete and continuous random variables.
- **Expected Value and Variance:** Measures of central tendency and dispersion for probability distributions.

4. Experimental Design

Experimental design focuses on planning experiments to ensure that the data obtained can provide valid and objective conclusions. Key concepts include:

- **Randomization:** Randomly assigning subjects to treatment groups.
- **Replication:** Repeating experiments to assess variability.
- **Blocking:** Grouping subjects with similar characteristics to reduce variability.
- **Factorial Design:** Studying the effects of multiple factors simultaneously.

5. Multivariate Statistics

Multivariate statistics involve the observation and analysis of more than one statistical outcome variable at a time. Techniques include:

- **Principal Component Analysis (PCA):** Reducing the dimensionality of data.
- **Factor Analysis:** Identifying underlying relationships between variables.
- **Cluster Analysis:** Grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.
- **Multivariate Regression:** Extending regression analysis to multiple dependent variables.

6. Bayesian Statistics

Bayesian statistics incorporate prior knowledge or beliefs, along with current evidence, to update the probability of a hypothesis being true. Key concepts include:

- **Bayes' Theorem:** A mathematical formula to update probabilities based on new evidence.
- **Prior Distribution:** The initial belief about the parameter before seeing the data.
- **Posterior Distribution:** The updated belief after considering the data.

7. Nonparametric Statistics

Nonparametric statistics do not assume a specific distribution for the data. They are used when data do not meet the assumptions of parametric tests. Examples include:

- **Rank Tests:** Wilcoxon signed-rank test, Mann-Whitney U test.
- **Chi-Square Test:** Testing relationships between categorical variables.
- **Kruskal-Wallis Test:** Comparing more than two groups.

8. Time Series Analysis

Time series analysis involves analyzing data points collected or recorded at specific time intervals. Techniques include:

- **Trend Analysis:** Identifying long-term movement in time series data.
- **Seasonal Decomposition:** Breaking down data into seasonal, trend, and residual components.
- **Autoregressive Models:** Modeling the relationship between an observation and a number of lagged observations.

9. Spatial Statistics

Spatial statistics deal with the analysis of spatial and spatiotemporal data. Techniques include:

- **Spatial Autocorrelation:** Measuring the degree of dependency among observations in geographic space.
- **Kriging:** Interpolating the value of a random field at an unobserved location.
- **Spatial Regression:** Modeling the relationship between spatially located variables.

10. Biostatistics

Biostatistics applies statistical methods to biological, medical, and health-related fields. Applications include:

- **Clinical Trials:** Designing and analyzing data from clinical trials.
- **Epidemiology:** Studying the distribution and determinants of health-related states and events.