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.
   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?
   Answer: Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
   Answer: Modeling bounded count data

4. Point out the correct statement.
   Answer: All of the mentioned

5. _______random variables are used to model rates.
   Answer: Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   Answer: False

7. Which of the following testing is concerned with making decisions using data?

   Answer: Hypothesis

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.
   Answer: 0

9. Which of the following statement is incorrect with respect to outliers?
   Answer: None of the mentioned

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

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

The **Normal Distribution**, also known as the **Gaussian Distribution** or **Bell Curve**, is a continuous probability distribution that is widely used in statistics and probability theory.

A normal distribution is characterized by the following properties:

1. **Symmetry**: The distribution is symmetric around the mean, meaning that the probability of a value being above or below the mean is equal.
2. **Bell-shaped**: The distribution has a bell-shaped curve, with the majority of the data points clustered around the mean and tapering off gradually towards the extremes.
3. **Mean (μ)**: The mean is the average value of the distribution, which is also the midpoint of the symmetry.
4. **Standard Deviation (σ)**: The standard deviation is a measure of the spread or dispersion of the data, with about 68% of the data points falling within 1 standard deviation of the mean, about 95% within 2 standard deviations, and about 99.7% within 3 standard deviations.
5. **Probability Density Function (PDF)**: The PDF of a normal distribution is given by:

math

VerifyOpen In EditorEditCopy code

1f(x) = (1/σ√(2π)) * e^(-((x-μ)^2)/(2σ^2))

The normal distribution is commonly used to model real-valued random variables that are expected to be distributed symmetrically around the mean, such as:

- Human heights and weights
- IQ scores
- Errors in measurement
- Stock prices
- and many more...

The normal distribution has many important properties and is used extensively in statistical inference, hypothesis testing, and confidence intervals.

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

Handling missing data is a crucial step in data preprocessing. Here's a general framework for handling missing data:

1. **Identify the type of missing data**:
   - **Missing Completely at Random (MCAR)**: Missing data is independent of observed and unobserved data.
   - **Missing at Random (MAR)**: Missing data is dependent on observed data, but not on unobserved data.
   - **Not Missing at Random (NMAR)**: Missing data is dependent on both observed and unobserved data.
2. **Explore the data**:
   - **Summary statistics**: Calculate summary statistics (e.g., mean, median, mode) to understand the distribution of the data.
   - **Visualizations**: Use plots (e.g., histograms, scatter plots) to visualize the data and identify patterns.
3. **Imputation techniques**:
   - **Listwise deletion**: Remove rows with missing values (not recommended, as it can lead to biased results).
   - **Mean/Median imputation**: Replace missing values with the mean/median of the observed values.
   - **Regression imputation**: Use a regression model to predict missing values based on observed values.
   - **K-Nearest Neighbors (KNN) imputation**: Replace missing values with the values from the K most similar observations.
   - **Multiple Imputation**: Create multiple versions of the dataset with different imputed values, and then analyze each version separately.
   - **Last Observation Carried Forward (LOCF)**: Replace missing values with the last observed value (commonly used in time-series data).
4. **Evaluation metrics**:
   - **Mean Absolute Error (MAE)**: Calculate the average difference between imputed and actual values.
   - **Mean Squared Error (MSE)**: Calculate the average squared difference between imputed and actual values.

Recommended imputation techniques:

1. **Multiple Imputation**: This is a robust method that can handle complex missing data patterns.
2. **Regression imputation**: This method is suitable when there are strong relationships between variables.
3. **KNN imputation**: This method is suitable when there are no strong relationships between variables.

12. What is A/B testing?

A/B testing, also known as **split testing**, is a method of comparing two or more versions of a product, web page, or application to determine which one performs better in terms of achieving a specific goal or set of goals. The goal is to identify changes that can improve user engagement, conversion rates, or other desired outcomes.

Here's a step-by-step overview of the A/B testing process:

1. **Define the goal**: Identify the specific goal or metric you want to improve, such as increasing conversions, reducing bounce rates, or enhancing user experience.
2. **Hypothesize**: Formulate a hypothesis about which version of the product or page will perform better. This could be based on user feedback, analytics data, or industry best practices.
3. **Create variations**: Develop two or more versions of the product or page, with each version differing in one or more elements, such as:
   - **A (Control)**: The original version, which serves as a baseline for comparison.
   - **B (Treatment)**: The modified version, which includes the changes you want to test.
4. **Split traffic**: Divide your website traffic or user base into two or more groups, using a randomization process to ensure that each group is representative of the overall population.
5. **Run the test**: Serve each group the corresponding version of the product or page, and collect data on the desired metrics.
6. **Analyze results**: Compare the performance of each version, using statistical methods to determine whether the differences are significant.
7. **Draw conclusions**: If the results are statistically significant, determine which version performed better and why. Use this insight to inform future design decisions or iterate on the winning version.

Types of A/B testing:

1. **Simple A/B testing**: Compare two versions of a product or page.
2. **Multivariate testing**: Test multiple elements simultaneously to identify the most effective combination.
3. **Multi-armed bandit testing**: Dynamically allocate traffic to different versions based on their performance.

Tools and platforms:

1. **Google Optimize**: A free A/B testing and personalization tool.
2. **VWO**: A popular A/B testing and conversion optimization platform.
3. **Optimizely**: A comprehensive A/B testing and personalization platform.

Example: Suppose you want to increase the conversion rate on a landing page. You create two versions:

A (Control): The original page with a blue "Sign up" button. B (Treatment): The modified page with a green "Get started" button.

You split your website traffic, and after collecting data, you find that the green button version (B) has a 15% higher conversion rate than the blue button version (A). You can conclude that the green button is more effective and implement it as the new standard.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is a common technique for handling missing data, but it's not always an acceptable practice. Here's why:

Pros:

1. **Easy to implement**: Mean imputation is a simple and straightforward method to fill in missing values.
2. **Fast computation**: It's a quick way to impute missing data, especially when dealing with large datasets.

Cons:

1. **Loss of variability**: Mean imputation can reduce the variability in the data, as it replaces missing values with a single value (the mean). This can lead to underestimation of the true variability in the data.
2. **Introduces bias**: If the missing data is not Missing at Random (MAR), mean imputation can introduce bias into the data. For example, if the missing values are more likely to be extreme values, replacing them with the mean can skew the distribution.
3. **Ignores relationships**: Mean imputation doesn't take into account the relationships between variables, which can lead to inaccurate imputations.
4. **Not suitable for non-normal data**: Mean imputation assumes normality of the data, which may not always be the case. If the data is skewed or has outliers, mean imputation can be misleading.

When to use mean imputation:

1. **Exploratory data analysis**: Mean imputation can be a quick way to get a sense of the data distribution during exploratory data analysis.
2. **Simple datasets**: For small, simple datasets with minimal missing data, mean imputation might be sufficient.

Alternatives to mean imputation:

1. **Regression imputation**: Use a regression model to predict missing values based on the relationships between variables.
2. **K-Nearest Neighbors (KNN) imputation**: Replace missing values with the values from the K most similar observations.
3. **Multiple Imputation**: Create multiple versions of the dataset with different imputed values, and then analyze each version separately.

14. What is linear regression in statistics?

Linear Regression is a fundamental concept in statistics and machine learning that models the relationship between a dependent variable (target variable) and one or more independent variables (feature variables).

Definition: Linear regression is a statistical method that attempts to establish a linear relationship between a dependent variable (y) and one or more independent variables (x) by fitting a linear equation to the data. The goal is to create a model that can predict the value of the dependent variable based on the values of the independent variables.

Simple Linear Regression: In simple linear regression, there is only one independent variable (x) and the model takes the form:

$y = \beta 0 + \beta 1x + \varepsilon$

where:

- y is the dependent variable (target variable)
- x is the independent variable (feature variable)
- β0 is the intercept or constant term
- β1 is the slope coefficient
- ε is the error term (residual)

Multiple Linear Regression: In multiple linear regression, there are multiple independent variables (x1, x2, ..., xn) and the model takes the form:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn + \varepsilon$

where:

- y is the dependent variable (target variable)
- x1, x2, …, xn are the independent variables (feature variables)
- β0 is the intercept or constant term
- β1, β2, …, βn are the slope coefficients
- ε is the error term (residual)

Assumptions: For linear regression to be valid, certain assumptions must be met:

1. **Linearity**: The relationship between the dependent and independent variables should be linear.
2. **Independence**: Each observation should be independent of the others.
3. **Homoscedasticity**: The variance of the error term should be constant across all levels of the independent variable.
4. **Normality**: The error term should be normally distributed.
5. **No or little multicollinearity**: The independent variables should not be highly correlated with each other.

Interpretation: The coefficients (β) in the linear regression model represent the change in the dependent variable for a one-unit change in the independent variable, while holding all other independent variables constant.

Example: Suppose we want to predict the price of a house (y) based on its size (x1) and number of bedrooms (x2). A linear regression model might look like this:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + \varepsilon$

where:

- y is the price of the house
- x1 is the size of the house
- x2 is the number of bedrooms
- β0 is the intercept or constant term
- β1 and β2 are the slope coefficients
- ε is the error term (residual)

15. What are the various branches of statistics?

Statistics is a broad field that encompasses various branches, each focusing on a specific aspect of data analysis, interpretation, and decision-making. Here are some of the main branches of statistics:

1. **Descriptive Statistics**: Descriptive statistics involves summarizing and describing the basic features of a dataset, such as measures of central tendency (mean, median, mode), measures of variability (range, variance, standard deviation), and data visualization techniques.
2. **Inferential Statistics**: Inferential statistics focuses on making conclusions or inferences about a population based on a sample of data. This branch includes hypothesis testing, confidence intervals, and significance testing.
3. **Exploratory Data Analysis (EDA)**: EDA is an approach to analyzing data that emphasizes the use of visual and quantitative methods to understand the underlying structure of the data, identify patterns, and detect anomalies.
4. **Machine Learning**: Machine learning is a subfield of artificial intelligence that involves developing algorithms and models that enable machines to learn from data, make predictions, and improve their performance over time.
5. **Bayesian Statistics**: Bayesian statistics is a branch of statistics that uses Bayesian inference, which involves updating the probability of a hypothesis based on new data. This approach is particularly useful for modeling complex systems and making decisions under uncertainty.
6. **Time Series Analysis**: Time series analysis involves analyzing and modeling data that is collected over time, such as stock prices, weather patterns, or website traffic. This branch includes techniques like ARIMA, exponential smoothing, and spectral analysis.
7. **Spatial Statistics**: Spatial statistics deals with the analysis of data that is associated with geographic locations, such as disease outbreaks, climate patterns, or economic activity.
8. **Biostatistics**: Biostatistics is the application of statistical principles to medical and health-related data, including the design of experiments, analysis of clinical trials, and epidemiology.
9. **Econometrics**: Econometrics is the application of statistical methods to economic data, including the analysis of economic systems, forecasting, and policy evaluation.
10. **Computational Statistics**: Computational statistics involves the development and application of computational methods and algorithms for statistical analysis, such as Markov chain Monte Carlo (MCMC) and bootstrap resampling.
11. **Nonparametric Statistics**: Nonparametric statistics involves the use of statistical methods that do not require a specific distribution or parameter, such as rank-based tests and permutation tests.
12. **Survival Analysis**: Survival analysis is a branch of statistics that deals with the analysis of time-to-event data, such as the time until death, failure, or recovery.
13. **Quality Control**: Quality control involves the use of statistical methods to monitor and improve the quality of products, processes, and services.
14. **Operations Research**: Operations research is an interdisciplinary field that combines statistics, mathematics, and computer science to optimize decision-making in complex systems.