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 mention.

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

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

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

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
   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, also known as Gaussian distribution or bell curve, is a continuous probability distribution that is symmetric and bell-shaped. It is characterized by its mean (μ) and standard deviation (σ), which determine the shape and location of the distribution.

In a normal distribution, the values are symmetrically distributed around the mean, creating a bell-shaped curve. The mean, median, and mode of a normal distribution are all equal and located at the center of the distribution. The distribution is defined by the probability density function (PDF), which provides the likelihood of observing a particular value within the distribution.

The properties of the normal distribution are significant in various fields, especially in statistics and data analysis. Many natural phenomena and random variables, such as heights, weights, test scores, and errors, tend to follow a normal distribution. This distribution is important for statistical inference, hypothesis testing, and constructing confidence intervals. It serves as a reference for comparing and analyzing other data distributions.

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

Ans:- Handling missing data is an important aspect of data analysis and requires careful consideration. Here are some common approaches for handling missing data:

Complete Case Analysis (CCA): In this approach, any data point with missing values is excluded from the analysis. It is suitable when the missing data is minimal and randomly distributed. However, this method can lead to reduced sample size and potential bias if the missingness is not random.

Mean/Mode/Median Imputation: In this method, missing values are replaced with the mean (for continuous variables), mode (for categorical variables), or median (for skewed distributions) of the available data. It is a simple technique but can distort the distribution and underestimate the uncertainty in the data.

Last Observation Carried Forward (LOCF): This imputation technique replaces missing values with the last observed value in time series or longitudinal data. It assumes that the missing values are similar to the most recent observed values. However, it may not be suitable if the missingness is non-random or if the variable changes over time.

Multiple Imputation (MI): MI involves creating multiple plausible imputed datasets by estimating missing values based on the observed data. Statistical models are used to impute the missing values, and the analysis is performed separately on each imputed dataset. The results are then combined to obtain the final estimates, taking into account the uncertainty introduced by imputation.

Model-Based Imputation: This technique involves using predictive models to estimate missing values based on other variables in the dataset. It takes into account the relationships and patterns in the data to impute missing values more accurately. Examples include regression imputation, k-nearest neighbors imputation, and machine learning-based imputation methods.

The choice of imputation technique depends on the nature and pattern of missingness, the distribution of data, and the analysis goals. Multiple imputation is generally considered a robust approach when the missingness is non-random and allows for appropriate handling of uncertainty. However, it is always important to carefully assess the limitations and potential biases associated with any imputation method used.

12. What is A/B testing?

Ans:- A/B testing, also known as split testing, is a method used to compare two versions of a webpage, app, or other digital elements to determine which one performs better in terms of user engagement, conversions, or other key metrics. It is a controlled experiment where users are randomly divided into two groups: Group A and Group B.

In A/B testing, Group A is presented with the original or existing version of a webpage or feature (referred to as the "control" or "baseline"), while Group B is presented with a slightly modified version (referred to as the "variant" or "treatment"). The purpose is to evaluate whether the changes made to the variant produce statistically significant improvements compared to the control.

To conduct an A/B test, the following steps are typically followed:

1. **Hypothesis formulation**: Clearly define the goal and expected impact of the changes being tested. For example, "Changing the color of the call-to-action button will increase click-through rates."

2. **Random assignment**: Randomly assign users to either Group A or Group B to ensure a fair comparison. This helps minimize bias and ensures that the groups are similar in terms of user characteristics.

3. **Implementation**: Create and deploy the control and variant versions of the webpage or feature. It's crucial to ensure that only the specific element being tested differs between the two groups, while keeping everything else constant.

4. **Data collection**: Gather data on relevant metrics such as click-through rates, conversion rates, bounce rates, or any other key performance indicators (KPIs). Track and record the user behavior for both groups during the testing period.

5. **Statistical analysis**: Analyze the collected data to determine if the differences observed between the control and variant groups are statistically significant. Statistical tests, such as chi-square test or t-test, are often used to assess the significance of the results.

6. **Conclusion**: Based on the analysis, determine whether the variant outperforms the control. If the results are statistically significant and align with the initial hypothesis, it may be concluded that the changes made in the variant version have a positive impact.

A/B testing allows organizations to make data-driven decisions and optimize their digital elements by comparing different variations. It is widely used in marketing, user experience (UX) design, product development, and other areas where the effectiveness of different versions needs to be evaluated.

13. Is mean imputation of missing data an acceptable practice?

Ans:- Mean imputation of missing data is a commonly used technique due to its simplicity. However, it has certain limitations and potential drawbacks that need to be considered before deciding whether it is an acceptable practice. Here are some points to consider:

1. **Loss of variability**: Mean imputation replaces missing values with the mean of the available data. This can lead to an underestimation of the variability in the imputed variable since all imputed values will have the same value (i.e., the mean). This underestimation of variability can impact subsequent analyses and statistical inference.

2. **Distortion of distributions**: Mean imputation assumes that the missing values are missing completely at random (MCAR) and that the variable follows a normal distribution. However, if the missingness is not MCAR or if the variable has a non-normal distribution, mean imputation can introduce bias and distort

the true distribution of the variable.

3. **Artificially reducing uncertainty**: By imputing missing values with a single value (i.e., the mean), mean imputation tends to underestimate the uncertainty associated with the imputed values. This can lead to overly confident estimates and inappropriate standard errors in subsequent analyses.

4. **Potential impact on relationships**: Mean imputation can affect the relationships between variables. Imputing missing values based on the mean may introduce spurious correlations or alter existing correlations, which can lead to erroneous interpretations and conclusions.

5. **Not suitable for categorical variables**: Mean imputation is not appropriate for categorical variables since it operates on numeric values. For categorical variables, mode imputation or other techniques specifically designed for categorical data should be used.

Despite these limitations, mean imputation may still be considered acceptable in certain situations, such as:

- When missingness is minimal and randomly distributed, and the impact of imputation is expected to be minimal.

- As a quick exploratory analysis to get a sense of the data before more sophisticated imputation methods are applied.

- As a sensitivity analysis to compare results obtained with mean imputation against other imputation methods.

It's worth noting that more advanced imputation techniques, such as multiple imputation or model-based imputation, are generally recommended as they can provide more accurate and reliable results by capturing the uncertainty associated with missing values and preserving the characteristics of the original data distribution.

14. What is linear regression in statistics?

Ans:- Linear regression is a statistical technique used to model and analyze the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, aiming to find the best-fitting linear equation that describes the association between them.

In linear regression, the dependent variable (also called the response variable or outcome) is predicted or explained by one or more independent variables (also called predictor variables or features). The goal is to estimate the coefficients of the linear equation that minimizes the differences between the observed values of the dependent variable and the predicted values from the model.

The linear regression model can be represented as:

$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p + \varepsilon$

Where:

- Y is the dependent variable

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

- $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are the coefficients (also known as regression coefficients or slope parameters)

- ε represents the error term or residual (the difference between the observed and predicted values)

The coefficients ($\beta_0, \beta_1, \beta_2, ..., \beta_p$) in the equation represent the change in the dependent variable associated with a unit change in the corresponding independent variable, assuming all other independent variables are held constant. The error term (ε) captures the unexplained variation in the dependent variable.

The estimation of the coefficients is typically done using the method of least squares, which minimizes the sum of squared residuals. This estimation process calculates the best-fit line or hyperplane that represents the linear relationship between the variables.

Linear regression is widely used in various fields for prediction, forecasting, and understanding the relationship between variables. It provides valuable insights into the strength and direction of the relationship, as well as the contribution of each independent variable in explaining the variability in the dependent variable.

15. What are the various branches of statistics?

Ans:- Statistics is a broad field that encompasses various branches or sub-disciplines, each focusing on different aspects of data analysis and inference. Here are some of the main branches of statistics:

1. **Descriptive Statistics**: Descriptive statistics involves summarizing and describing data using measures such as central tendency (mean, median, mode), variability (standard deviation, range), and graphical representations (histograms, box plots) to provide a concise overview of the data.

2. **Inferential Statistics**: Inferential statistics aims to make inferences and draw conclusions about a population based on a sample. It involves techniques such as hypothesis testing, confidence intervals, and regression analysis to make predictions, test hypotheses, and generalize findings from the sample to the larger population.

3. **Biostatistics**: Biostatistics applies statistical methods to analyze and interpret data related to biological and health sciences. It includes the design and analysis of clinical trials, epidemiological studies, and the evaluation of healthcare interventions.

4. **Econometrics**: Econometrics is the branch of statistics that focuses on applying statistical methods to economic data. It involves the development of econometric models and the analysis of economic relationships, such as supply and demand, using techniques such as regression analysis and time series analysis.

5. **Psychometrics**: Psychometrics deals with the measurement and analysis of psychological data. It involves the development and validation of assessment tools, such as questionnaires and tests, and the statistical analysis of responses to measure constructs like personality traits, intelligence, and attitudes.

6. **Social Statistics**: Social statistics is concerned with the collection, analysis, and interpretation of data related to social phenomena. It involves studying social trends, demographics, public opinion, and conducting surveys to understand patterns and make informed decisions in social sciences.

7. **Statistical Computing**: Statistical computing focuses on developing computational algorithms, software tools, and techniques for efficiently analyzing and processing large datasets. It includes areas such as data visualization, data mining, machine learning, and big data analytics.

8. **Spatial Statistics**: Spatial statistics deals with the analysis and modeling of spatially referenced data, such as geographical or environmental data. It includes techniques for spatial interpolation, spatial autocorrelation, and the study of spatial patterns and processes.

These are just a few examples of the many branches of statistics. The field of statistics is interdisciplinary and finds applications in various domains, including business, finance, engineering, environmental sciences, and more.