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

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" refers to a symmetric, bell-shaped probability distribution that is defined by its mean and standard deviation. It is also referred to as the bell curve or the Gaussian distribution. The values in a normal distribution are symmetrically distributed about the mean, with the mean occurring most frequently. The distribution has a smooth, continuous curve that tapers out at both ends. The normal distribution has numerous applications in statistics, probability theory, and data analysis.

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

Handling missing data is an important task in data analysis to ensure accurate and reliable results. There are several common approaches to handle missing data:

1. Deleting Rows with missing values
2. Impute missing values for continuous variable
3. Impute missing values for categorical variable
4. Other Imputation Methods
5. Using Algorithms that support missing values
6. Prediction of missing values
7. Imputation using Deep Learning Library — Datawig

Choosing an appropriate imputation technique for handling missing data depends on various factors. Here are some commonly recommended imputation techniques:

Mean/Mode Imputation: This method involves replacing missing values with the mean (for numeric variables) or mode (for categorical variables) of the available data for that variable. Mean/Mode imputation is simple and easy to implement, but it does not account for any relationships or patterns in the data.

Regression Imputation: Regression imputation involves using regression models to predict missing values based on other variables. A regression model is built using the cases with complete data, and then the model is used to impute the missing values. This approach takes into account relationships between variables and can provide more accurate imputations.

Multiple Imputation: Multiple imputation generates multiple plausible imputations for missing values, taking into account the uncertainty associated with the imputed values. This approach incorporates the variability introduced by imputation into subsequent analyses, providing more reliable estimates and valid statistical inference.

Advanced Techniques: There are more advanced imputation techniques available, such as hot-deck imputation, k-nearest neighbors imputation, and predictive mean matching. These methods utilize additional information from the dataset or utilize similar cases to estimate missing values. They can be useful when there are complex patterns of missingness or when multiple variables need to be imputed simultaneously.
It is important to note that the choice of imputation technique should consider the nature of the data, the extent of missingness, the underlying assumptions, and the specific research goals.

12. What is A/B testing?

A/B testing, also known as split testing or bucket testing, is a method used to compare two or more versions of a webpage, app, or other elements to determine which version performs better. It is a common practice in marketing, UX design, and product development to make data-driven decisions and optimize results.
In A/B testing, different variants, typically referred to as the control (A) and the treatment (B), are created

and randomly presented to different segments of users or visitors. The performance of each variant is then measured and compared using specific metrics and statistical analysis. The objective is to identify the variant that performs better in terms of desired metrics such as conversion rate, click-through rate, engagement, or other key performance indicators.
A/B testing allows organizations to make informed decisions by testing different ideas, designs, or strategies and measuring their impact on user behavior and outcomes. It helps optimize conversion rates, user engagement, and overall performance by providing empirical evidence on what works best for the target audience.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is a simple and convenient approach to handle missing data, but it has limitations and may introduce biases if not appropriately applied. It is generally considered acceptable for missing data that is missing completely at random (MCAR) or missing at random (MAR) when certain assumptions are met. However, more advanced imputation methods, such as multiple imputation or regression imputation, are often recommended as they can provide more accurate estimates and account for the uncertainty associated with missing data.

14. What is linear regression in statistics?

Linear regression is a statistical method to establish a linear relationship between a dependent variable and one or more independent variables. It predicts or explains the value of the dependent variable based on the values of the independent variables.

In linear regression, the equation $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p + \varepsilon$ represents the relationship between the dependent variable (Y) and the independent variables ($X_1, X_2, ..., X_p$). The coefficients ($\beta_0, \beta_1, \beta_2, ..., \beta_p$) quantify the impact of each independent variable on the dependent variable. The error term (ε) captures the random variation or unexplained factors in the relationship.

15. What are the various branches of statistics?

Statistics has various branches:

a. Descriptive Statistics: Summarizing and presenting data.
b. Inferential Statistics: Making inferences about populations based on sample data.
c. Probability Theory: Study of uncertainty and likelihood of events.
d. Biostatistics: Application of statistics in biological and health sciences.
e. Econometrics: Applying statistics to economic data for analysis and prediction.
f. Bayesian Statistics: Updating probabilities based on new evidence.
g. Multivariate Statistics: Analyzing data with multiple variables.
h. Time Series Analysis: Analyzing time-dependent data.
i. Experimental Design: Planning controlled experiments.
j. Statistical Computing: Using computation for statistical analysis.