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 D) All of the mentioned

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 is a probability distribution which is symmetric and bell shaped. It is also known as Gaussian distribution. The normal distribution is defined by two parameters the first one is Mean and Standard deviation. Mean represents the central tendency of distribution and Standard deviation measures the variability of distribution. These determine the shape of normal distribution such that the majority of data get covered within one standard deviation of the mean. The probability of an observation being farther decreases drastically.

Normal distribution is an essential tool in statistical analysis. It is used to model and analyze. In addition, the central limit theorem states that the sum or average of a large number of independent and identically distributed random variables, regardless of their underlying distribution, tends to follow a normal distribution.

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

Ans Handling missing data is important task for data analysis, as missing values could lead to biased and inefficient results. Here are some common techniques for handling missing data:

1. Deleting missing value either row or column wise. While it is simple to implement, it can result in a loss of information and reduced sample size.

2. Imputation: This method involves filling in missing values with estimates based on the available data. Some of the imputation techniques available are:

- Mean imputation: Replacing missing values with the mean of the observed values for that variable.
- KNN imputation: It will refer to the given column will all the no NAN and take the K_neighbor and compare their data and take the mean of the data and fill it to missing value
- Iterative imputation: This method treats other columns (which does not have nulls as feature and train on them and treat Null column as label. Finally, it will predict the NaN data and impute. It's just like regression problem. Here null column is label

The choice of imputation technique depends on the nature of the missing data and the goals of the analysis.

12. What is A/B testing?

Ans A/B testing is also known as split testing. It is a statistical method used to compare two versions of a product or service to determine which one performs better. The goal is to identify which version leads to more desirable outcomes.

The process of A/B testing involves randomly dividing the audience into two groups and exposing each group to a different version of the product or service. The groups are then compared based on their response to the different versions, and statistical analysis is used to determine if there is a significant difference between the two groups. A/B testing can be used to test different features, layouts, content, pricing, or any other variable that can potentially impact user behavior.

13. Is mean imputation of missing data acceptable practice?

Ans Mean imputation of missing data is a simple and one of the widely used imputation technique, but it has several limitations and is not always an acceptable practice.

One limitation of mean imputation is that it assumes that the probability of missingness does not depend on the values of the observed or unobserved data. Mean imputation can lead to biased and inefficient

estimates, as it does not account for the underlying relationships between variables.

Another limitation of mean imputation is that it can lead to an underestimation of the variability , as it reduces the observed variability by filling in missing values with the same value (the mean). This can lead to an inflation of the statistical significance of the observed results and a decrease in the precision of the estimates.

Hence, mean imputation of missing data is not acceptable practice.

14. What is linear regression in statistics?

Ans Linear regression is a statistical method used to model the relationship between a dependent variable (Label) and one or more independent variables (Features). The goal of linear regression is to find the best-fit line that minimizes the distance between the predicted values and the actual values of the dependent variable.

The line is characterized by two parameters: the intercept, which is the point where the line or hyperplane intersects the y-axis, and the slope, which is the rate at which the dependent variable changes for a unit change in the independent variable. Linear regression is often used for prediction and forecasting.

The goodness of fit of the regression model can be assessed using metrics such as R-squared, adjusted R-squared, and root-mean-square error (RMSE).

15. What are the various branches of statistics?

Ans Types of Statistics Descriptive and Inferential

Descriptive: - When population is very small and we can describe it we use Descriptive statistics. This Branch deals with collection, analysis, interpretation and presentation of data by summarizing and visualizing data using central tendency, measures of dispersion, and graphical displays.

Inferential: - When population is very large and we take small random samples and infer these sample result as whole population result it is known as Inferential statistics. This branch deal with making inferences and predictions about population using sample data. It includes hypothesis testing, regression analysis.