STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False
Answer: - 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
Answer: - 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
Answer: - b) modeling bounded count data

4. Point out the correct statement.
a) The exponent of 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
Answer: - d) All of the mentioned

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

6. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False
Answer: - 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
Answer: - 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
Answer: - 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
Answer: - c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
Answer: - Normal Distribution is also called as Gaussian distribution or bell curve. It is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and kutosis of 3. The normal distribution has high use in statistics due to its adaptability.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer: - Data can be missing in the following ways:-

- Missing Completely At Random (MCAR): When missing data does not depend on observed data or missing data, then we consider the data to be missing completely at random.
- Missing At Random (MAR): When missing values depend on observed data.
- Missing Not At Random (MNAR): When the missing values depend on missing data, even when the observed data are given.

Imputation Techniques: -

1. Single Imputation
In simple mean imputation, you can replace the missing value with the mean

2. Multivariate Imputation by Chained Equations (MICE)
Generating imputed values by drawing from estimated conditional distributions.

3. Random Forest
Random forest imputation can accommodate nonlinearities and interactions, not requiring a particular regression model.

4. Dropping rows with null values

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value.

5. Dropping features with high nullity
In Pandas, there are two very useful methods: isnull () and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

12. What is A/B testing?
Answer: - A/B testing is statistical inference. An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

13. Is mean imputation of missing data acceptable practice?
Answer: - It is easy to implement, and thus popularly used, but should be considered as last resort. It is used to predict the missing data. It also can be used for both i.e. continuous as well as categorical data and so it makes advantageous over other imputations.

The disadvantages: -
1. Mean imputation does not preserve the relationship among variables. It preserves the mean of observed data. If data is

missing completely at random, the estimate of the mean remains unbiased.
2. Mean Imputation leads to an underestimate of standard errors. Because the imputations are estimates themselves, there is some error associated with them.

14. What is linear regression in statistics?
Answer:- Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

A linear regression line has an equation of the form $Y = mx + c$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is m, and c is the intercept (the value of y when $x = 0$).

Types of linear regression: -

1. Simple linear regression
2. Multiple linear regressions
3.Logistic regression
4. Ordinal regression
5. Multinomial regression

15. What are the various branches of statistics?
Answer: - Various branches of statistics are given below: -

1. Descriptive Methods:-

- Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

2. Inferential Methods: -

- If the data is too big, we use inferential statistics. A few samples of the data is taken, and the average is found. The average is then applicable to all the data from where the samples are selected. Regression analysis is a widely used technique of statistical inference used to determine the strength and nature of the relationship (i.e., the correlation) between a dependent variable and one or more explanatory (independent) variables.