STATISTICS WORKSHEET-1

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. Usually replacing the standard error by its estimated value does change the CLT.
a) True
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

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

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?

Normal distribution has a bell shaped curve, which is symmetrical. In this distribution, the mean median and the mode are same. This distribution has more data around the mean.

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

Missing data can affect the performance of the models.
It can be handled by deletion of any records with missing values. However, it can be used only when the dataset is large and the amount of deleted data is minute.

Various impution methods
Mean/Median/Mode Imputation
It replaces the missing values with the mean, median or mode. It is suitable only with numerical data.
Forward fill and backward fill:- It fills the missing values with last observed value or next observed value.
K-Nearest Neighbors:- Using the value of the k-nearest neighbours to impute the missing values.
Regression imputation:-
Use regression models to predict and fill in missing values based on other available variables.

12) What is A/B testing?

It is also know as split testing. It basically compares two versions of a variable which can be a web page or an app to find out which performs better.It is widely used in marketing, product development, and user experience design.

13. Is mean imputation of missing data acceptable practice?

If the missing data values are comparatively small, it is an ideal technique which can be used.
For exploratory data analysis, it can be used as a choice.Also for symmetric data, this can used, where there is no skewness.
However, mean imputation , reduces the variance. Also it does not consider the relationship between variables.

14. What is linear regression in statistics?

It is a statistical method used to model the relationship between a dependent variable and one or more independent variables.

Simple Linear Regression: Involves one dependent variable and one independent variable
Multiple Linear Regression: Involves one dependent variable and two or more independent variables.

Uses: It is used to predict the value of the dependent variable based on the values of the independent variables. And also to understand the relationship between the dependent and independent variables, including the strength and direction of this relationship.

15. What are the various branches of statistics?

Descriptive Statistics: involves summarizing and organizing data to describe the main features of a dataset.
Measures of Central Tendency: Mean, median, mode.

Measures of Dispersion: Range, variance, standard deviation, interquartile range.

Inferential Statistics: involves making inferences about a population based on a sample of data drawn from that population.

Hypothesis Testing: t-tests, chi-square tests, ANOVA.
Regression Analysis: Linear regression, logistic regression.