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?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

10.The Normal Distribution, also called the Gaussian distribution.
The normal distribution is the most commonly seen continuous distribution in nature. In the Normal Distribution the mean, median and mode all line up such that the center of the distribution is the mean. Because of this, exactly half of the results fall to either side of the mean.
The normal distribution is also identifiable by its bell shape and may sometimes be referred to as a bell curve.

11. We can handle the missing data by two methods:
I. Deleting the missing values
II. Imputing missing values.
Techniques recommended for imputation:
1. Mean or Median Imputation: we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low.
2. Multivariate Imputation by Chained Equations (MICE)
MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables

12.A/B testing is the process of comparing two variations of a page element, usually by testing users' response to variant A vs. variant B and concluding which of the two variants is more effective, different statistical analysis methods are used to determine which variant drives more conversions.
It is also known as split testing or bucket testing,
A/B testing lets you increase user engagement, reduce bounce rates, increase conversion rates, minimize risk, and effectively create content. Running an A/B test can have significant positive effects on your site or mobile app.

13.Mean imputation is typically considered terrible practice since it ignores feature correlation.
It has two big problems:
Firstly,Mean imputation does not preserve the relationships among variables:- Mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate.
Secondly,Mean Imputation Leads to An Underestimate of Standard Errors:-Applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low.

14.A statistical way of measuring the relationship between two variables It strives to show the relationship between the variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable.
For example, the weight of the person is linearly related to his height. Hence this shows a linear relationship between the height and weight of the person. As the height is increased, the weight of the person also gets increased.

15.Various branches of statistics are:
I. Mathematical Statistics: It helps in forming the experimental and statistical distribution.
II. Statistical methods: It helps in the collection, tabulation and interpretation of data. It helps in analyzing the data and returns insight from the data.
III. Descriptive Statistics: It helps in summarizing and organizing any data set characteristics. This is a branch of statistics which deals with methods of collection of data, its presentation and organization in various forms, such as distribution tables, graphs and diagrams.
IV. Inferential Statistics: This is a branch of statistics which deals with techniques used for analysis of data, making estimates Inferential statistics is used to make comparisons or predictions about a larger group, known as population, using information gathered about a small part of that population called a sample.