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?

Answer: Normal distribution, also known as the Gaussian distribution, 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 graphical form, the normal distribution appears as a "bell curve". Values are equally likely to plot either above or below the mean. It is a probability distribution that (roughly) describes many common datasets in the real world.

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

Answer: Techniques for Handling the Missing Data:

a. List wise or case deletion.
b. Pairwise deletion.
c. Mean substitution.
d. Regression imputation.
e. Last observation carried forward.
f. Maximum likelihood.
g. Expectation-Maximization.
h. Multiple imputation.

Imputation Techniques are:

a. Complete Case Analysis (CCA):- This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e. we consider only those rows where we have complete data i.e. data is not missing.
b. Arbitrary Value Imputation.
c. Frequent Category Imputation.

12. What is A/B testing?

Answer: A/B testing is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment.

13. Is mean imputation of missing data acceptable practice?

Answer: Yes, Mean imputation is typically considered terrible practice since it ignores feature correlation.

14. What is linear regression in statistics?

Answer: In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (or independent variables). It is used to predict the value of a variable based on the value of another variable.

15. What are the various branches of statistics?

Answer: There are three real branches of statistics, namely:

- **Data collection** – It is all about how the actual data is collected.
- **Descriptive statistics** – It is the part of statistics that deals with presenting the data we have.
- **Inferential statistics** – It is the aspect that deals with making conclusions about the data.