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 metioned

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

6. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Ans :- 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
   Ans :- 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
   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 :-

- The Normal Distribution is a probability function that describes how the values of a variable are distributed.
- It is a continuous probability distribution.
- Its shape is that of a Bell. And also known as Bell Curve or Gaussian distribution
- Mean = Median = Mode (in case of Normal distribution)
- Total area under the curve = 1.
- It is perfectly symmetrical in shape. (It can be divided into 2 equal halves, if divided from the middle)
- It shows that the data near the mean are more frequent in occurrence than data far from the mean.

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

Missing Data is Handled in the following ways :-
1) Deleting Rows with missing values
2) Listwise or case deletion/Pairwise Deletion
3) Mean Substitution
4) Last Observation Carried Forward
5) Maximum Likelihood
6) Expectation Maximization
7) Sensitivity Analysis

 IMPUTATION Techniques

1) Mean or Median Imputation
2) Multivariate Imputation by Chained Equations (MICE)
3) Hot Deck Imputation
4) Cold Deck Imputation
5) Regression Imputation
6) Stochastic Regression Imputation
7) Interpolation and Extrapolation
8) Single or Multiple Imputation

12. What is A/B testing?
Ans :-

 A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B.It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

 A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

13. Is mean imputation of missing data acceptable practice?

Ans:-

     Outliers data points will have significant impact on the mean and hence, in such cases, it is not recommended to use mean for replacing the missing values. Using mean value for replacing missing values may not create a great model and hence gets ruled out.

     Mean imputation does not preserve the relationships among variables.

True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased.

14. What is linear regression in statistics?

Ans :-

     Linear regression models the relationship between two variables by fitting a linear equation to observed data. ... 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 case of one explanatory variable is called simple linear regressio*n*; for more than one, the process is called multiple linear regression

15. What are the various branches of statistics?

Ans :-

     The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

1) **Descriptive Statistics**

   Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

   Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

2) **Inferential Statistics**

   Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

   Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.