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** Answer- 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** Answer – 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) M**odeling bounded count data** Answer – 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** Answer – d) All of the mentioned

5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson Answer – 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** Answer - b) False

7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) **Hypothesis** Answer - 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** Answer – 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** Answer – 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?

Answer –

10 . A normal distribution is **a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme**. The middle of the range is also known as the mean of the distribution.

11 . Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you.

Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

The following are some of the most prevalent methods:

Mean imputation

Substitution

Hot deck imputation

Cold deck imputation

Regression imputation

Stochastic regression imputation

Interpolation and extrapolation.

Single or Multiple Imputation

- Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single.
- The term "single" refers to the fact that you only use one of the seven methods to estimate the missing number outlined above.
- It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.
- When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.
- Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with listwise deletion, which is most software's default.
- The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set.

12 . A/B testing is **a way to compare multiple versions of a single variable**, for example by testing a subject's response to variant A against variant B, and determining which of the variants is more effective.

13 . Mean imputation is **typically considered terrible practice** since it ignores feature correlation.

14 . 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.

15 . There are three real branches of statistics: **data collection, descriptive statistics and inferential statistics**.

FLIP ROBO