Statistics– WORKSHEET 4

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 mentioned

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

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

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
   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?

- Normal distribution may be defined as a probability distribution where the mean is zero and standard deviation is 1.
- Graphical representation of normal distribution is a proper bell curve.
- I normal distribution the skew is zero and value of kurtosis is 3.
- In practical it is hard to achieve a normal distribution

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

- Missing data in a dataset are the result no response or non-availability of the required data. These are very common occurrence in dataset
- Higher amount of missing data adversely affects the model performance.
- However, there are several ways to treat the missing data.
- Either we can replace those missing data by a statistical value like the mean, median, mode or zero. Depending on the percentage of missing values w.r.t the dataset we can also choose to discard them.
- If the missing value percentage are lower than 10% of the dataset, we can opt to drop them.
- If the missing values is a categorical / ordinal value then we can impute those values by the mode.
- If the missing values is a continuous/discreet data then we can impute those missing values by their mean or median.

12. What is A/B testing
    - A/B testing is basically statistical hypothesis testing. It is an method which helps to take decision that estimate population parameter based on sample statistics.
    - For A/B testing we need to generate hypothesis, sample size and strategy.
    - Hypothesis may be defined as a statement which will represent the target, we want to achieve. This must be clear and precise.
    - We must have two hypothesis named as null hypothesis (H0) and alternate Hypothesis (H1).
    - The null hypothesis supports our base argument and alternate hypothesis rejects it.

13. Is mean imputation of missing data acceptable?

14. What is linear regression in statistics?
    Liner regression can be defined as a model used to establish relation between two variables using the liner approach. It establishes a relationship between the dependent and independent variable

15. What are the various branches of statistics?
    Statistics can broadly be divided into two categories named as;
    1) Descriptive statistics
    2) Inferential Statistics