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

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?

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) Modelling event/time data

Answer: b) Modelling bounded count data

c) Modelling 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

Answer: d) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

Answer: c) Poisson

d) All of the mentioned

6. Usually replacing the standard error by its estimated value does change the CLT.

a) True

Answer: b) False

7. Which of the following testing is concerned with making decisions using data?

a) Probability

Answer: b) Hypothesis

c) Causal

d) None of the mentioned

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

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

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?

- A normal distribution is also called a Gaussian Distribution or a bell curve.
- It is the probability distribution of a continuous variable.
- The mean and the median are equal for a normally distributed variable.
- The distribution shows a symmetry above and below the mean.
- From the distribution we can see that most of the datapoints of that variable occurs very close and around the mean.
- There very less data near the tails of the distribution.

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

- If the variable is a continuous numerical variable, we can use mean of the variable for imputation.
- If the variable is a categorical variable or a discrete numerical variable, we can use the most occurring value in the variable.
- If most of the values are missing, for example more than 50%, then it makes sense to just drop the column.
- There are imputers like SimpleImputer, KNN Imputer, MICE imputation etc can be used to fill the missing values.
- Linear Regression algorithm can also be used to predict the missing continuous values.
- Datawig is a Deep learning library that can be used for missing values imputation.

12. What is A/B testing?

AB testing is also called split testing. AB testing is mostly used to experiment the user experience for different variants A and B. Usually two or more variants of landing pages or CTAs or email or new features etc will be used. The Variant A is considered the 'Control' and the other variants are called 'Treatment'. The control and the treatment variants are presented to the users randomly and the user engagement is captured. Then using statistical methods, a hypothesis testing is conducted to check if the new variants are significantly increasing the user engagement or not.

13. Is mean imputation of missing data acceptable practice?

No. Imputing with mean can reduce the model's accuracy and bias the result. It does not take into account any correlation factor in the dataset between the features. Using mean for imputation will also reduce the variance in the dataset this concentrating more datapoints towards the mean.

14. What is linear regression in statistics?

Linear regression in statistics is a method of approximating a linear relationship between numeric or categorical variable(s) to a continuous variable. The linear equation y = mx+ c is used where the coefficients in the equation are the slope m and the intercept c of a line. The output is always a continuous data.

15. What are the various branches of statistics?

There are 2 branches of statistics: Descriptive statistics and Inferential statistics. The descriptive statistics are the quantitative summary of a dataset. The Inferential statistics is the analysis/experiments performed on sample data to infer information or conclusions about the population which the sample is taken from.