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.

- ✓ **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?

- ✓ **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.

- ✓ **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.

- ✓ **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

WORKSHEET

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 is the probability/frequency of distribution of the data, it ranges from -1 to +1,
- In Normal distribution, we expect the Mean, Median and Mode are equal and at expected to be at 0
- It's a bell-shaped curve, which shows that most of the data is distributed at the Mean of the total data.

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

- The Best way to Handle the missing data for a very large data is to just remove the Missing Value or do nothing
- Impute by Mean/Mode
- Simple Imputer- Simply uses Strategy-Mean/Median/Mode from multiple columns
- KNN Imputer-Can be imputer based on near K neighbors of the feature variable data to impute the value in Label variable
- Iterative Imputer-It uses the Nan column as label variable and uses all the columns are feature variables and it predicts the Nan based on real data.

I recommend Iterative Imputer for Imputation over KNN because it used all other columns to predict my NaN.

12. What is A/B testing?

The A/B testing is an example of Statistics Hypothesis testing between 2 Variables A and B, to see the relationship between each other, Establish a equation between the variables and whether this relationship is significant or not

And this this established equation between A and B can be used to predict the incoming new records for A or B

13. Is mean imputation of missing data acceptable practice?

No, According to me the mean imputation of missing data is not recommended practice.

14. What is linear regression in statistics?

It is Modelling techniques which using Linear approach for understanding relationship between 2 continuous Variables in a data, Where One variable is a independent variable and the other is a dependent variable

It also helps in predicting the data based on the established linear equation for the observed values.

15. What are the various branches of statistics

There are 2 main branches of Statistics

- Descriptive statistics- These are the parameter which would be used to describe the data
  Ex-Mean, Median, SD, Var
- Inferential statistics- These are the parameters which would help us in using the data for inferring/ concluding/making discission on the data available
  Ex-ANOVA, chi square test, T test, Z test etc

---Internal Use---