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

ANSWER:(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

ANSWER:(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

ANSWER:(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

ANSWER:(D)[All of the above mentioned]

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

ANSWER:(C)[Poisson]

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
   c) Causal
   d) None of the mentioned

ANSWER:(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

ANSWER:(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

ANSWER:(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: Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- In reality, most pricing distributions are not perfectly normal.

Let us say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and (x + dx).

- $f(x) \geq 0 \ \forall \ x \in (-\infty, +\infty)$
- And $\int_{-\infty}^{+\infty} f(x) = 1$

The probability density function of normal or gaussian distribution is given by ; Where,

- x is the variable

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- μ is the mean
- σ is the standard deviation.

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

Ans: The missing data of the given dataset can be handled by replacing the null or NaN values using some imputing techniques. Some of the techniques used are Simple Imputer, Knn-Imputer and Iterative Imputer.

- Simple - Imputer: It is the basic imputer technique, in which the missing values are imputed using mead or mode of the non null values in that particular column. Here, we import Simple Imputer from Sk-learn and impute ,by default mean of the not null values are taken and replaced.
- Knn - Imputer: It is one of the most common used imputer-technique, in which the missing values are imputed using knn model technique . Here, n-neighbors are taken and replace the nearest neighbor value to the null value(NAN).
- Iterative Imputer: It is one of the commonly used imputer-technique. Here, it uses Regression model. In which, it predicts the missing value by treating it as label and the rest of the non-null values as features using training the features and testing the label.

12. What is A/B testing?

Ans: 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. It is a '**hypothetical**

testing' methodology for making decisions that estimate population parameters based on sample statistics. Ttest and ANOVA ae the 2 mostly used hypothesis testing tools.

1) Make a Hypothesis such as,

Null hypothesis:(Ho)-Decisions always leas to status quo. Current status or assumption doesn't change.

Alternate hypothesis:(Ha)-decision leads to opposite of Ho. we have to collect enough evidence through our tests to reject the null hypothesis.

2)create control group and test group.3)Conduct the A/B test and collect the data

Statistical significance of the Test:

Type1 Error-If 'Ho' is True and you reject 'Ho'.

Type2 Error-If 'Ho' is False and you fail to reject 'Ho'

13. Is mean imputation of missing data acceptable practice?

Ans: 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. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

14. What is linear regression in statistics?

Ans: Linear Regression is one of the supervised technique and can only be used for regression. Process of predicting a label and features. It uses discrete/continuous variables. Best fit regression line. Also, uses continuous dependent variables. For example:

$$Y = a + bx + e$$

where a = intercept, b=slop of the line, e=error term,

x=Independent data(feature),y=dependent data(label).

15. What are the various branches of statistics?

Ans: Statistics is a study of presentation, analysis, collection, interpretation and organization of data

There are **two main branches** of statistics

- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:

Inferential statistics used to make inference and describe about the population. These stats are more useful when it's not easy or possible to examine each member of the population.

Descriptive Statistics:

Descriptive statistics are used to get a brief summary of data. You can have the summary of data in numerical or graphical form.