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
   ANS-a)

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

   A

   Ans-a)
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

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

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

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?

10- ANS- 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 mean of the distribution.
The normal distribution is also known as a *Gaussian distribution* or probability bell curve. It is symmetric about the mean and indicates that values near the mean occur more frequently than the values that are farther away from the mean.

11-ANS-**NORMAL IMPUTATION**
we have *a* feature that has missing values. We can replace the missing values with the below methods depending on the data type of feature

- Mean
- Median
- Mode

If the data is numerical, we can use mean and median values to replace else if the data is categorical, we can use mode which is a frequently occurring value.
IMPUTATION BASED ON CLASS LABEL
Here, instead of taking the mean, median, or mode of all the values in the feature, we take based on class.
Take the average of all the values in the feature that belongs to class 0 or 1 and replace the missing values. Same with median and mode.
MODEL-BASED IMPUTATION
This is an interesting way of handling missing data. We take feature *f1* as the class and all the remaining columns as features. Then we train our data with any model and predict the missing values.

12-ANS- /B testing, also called **split testing** or bucket testing, is an experiment that compares two versions of content (A and B) that have one or more deliberate changes. The goal is to see which version appeals more to visitors or customers based on key metrics or conversion goals. A/B testing helps businesses be more data-driven and remove the guesswork from their marketing decisions.

13- Yes

14- Linear regression is a **statistical algorithm used to predict or visualize the relationship between two variables**. In linear regression, there are two types of variables: the independent variable, which is not impacted by the other variable, and the dependent variable, which is being studied and predicted by the regression model. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. The equation of a line relating the independent variable to the dependent variable uses the slope-intercept form of a line. Simple linear regression is used to estimate the relationship between two quantitative variables

15- Statistics has two main branches: **descriptive statistics and inferential statistics**[1]. Descriptive statistics summarizes or presents data, while inferential statistics draws conclusions or makes predictions based on data. There are also other branches of statistics that focus on specific fields or applications, such as biostatistics, econometrics, environmental statistics, psychometrics, and social statistics.