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 1 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 2 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 3 b) Modeling bounded count data

4. Point out the correct statement.

a) The exponent of 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 4 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 5 c) Poisson

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

a) True
b) False

Ans 6 b) False

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

a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Ans 7 b) Hypothesis

8. 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 8 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 9 c) Outliers cannot conform to the regression relationship

Q10 to Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

Ans 10

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 graphical form, the normal distribution appears as a "bell curve".

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

Ans 11

Missing data is the data value that is not stored for a variable in the observation of interest.

There are 2 primary ways of handling missing data:

a) Deleting the Missing values

List wise Deletion

Dropping rows/columns

b) Imputing the Missing Values

The recommended imputation techniques are:

- **Simple Imputer: Imputation with a constant value or Imputation using the statistics (mean, median (for continuous variable) and mode (for categorical variable).**
- **K-Nearest Neighbor Imputer**
- **Iterative Imputer**

12. What is A/B testing?

Ans 12

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. A/B testing is often associated with websites and apps, and it is extremely common on large social media platforms.

13. Is mean imputation of missing data acceptable practice?

Ans 13

The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eighty-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does. Secondly, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Ans 14

Linear regression establishes the linear relationship between two variables based on a line of best fit. Linear regression is thus graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other. Linear regression has two primary purposes—understanding the relationships between variables and forecasting.

1) The coefficients represent the estimated magnitude and direction (positive/negative) of the relationship between each independent variable and the dependent variable.

2) A linear regression equation allows you to predict the mean value of the dependent variable given values of the independent variables that you specify.

15. What are the various branches of statistics?

Ans 15

There are two main branches of statistics

- **Descriptive Statistics**
- **Inferential Statistics**

Descriptive Statistics:

Descriptive statistics are used to get a brief summary of a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).The summary of data can be in numerical or graphical form.

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.