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) 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: 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: 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

Ans: 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: c) Poisson

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

Ans: 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: 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: 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: 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: In probability theory and statistics, the Normal Distribution, also called the Gaussian Distribution, is the most significant continuous probability distribution. Sometimes it is also called a bell curve. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics. Furthermore, it can be used to approximate other probability distributions, therefore supporting the usage of the word 'normal 'as in about the one, mostly used.

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

Ans: Missing data is a common headache in any field that deals with datasets. It can arise for various reasons, from human error during data collection to limitations of data gathering methods. Luckily, there are strategies to address missing data and minimize its impact on your analysis. Here are two main approaches:

Deletion: This involves removing rows or columns with missing values. This is a straightforward method, but it can be problematic if a significant portion of your data is missing. Discarding too much data can affect the reliability of your conclusions.

Imputation: This replaces missing values with estimates. There are various imputation techniques, each with its strengths and weaknesses. Here are some common ones:

Mean/Median/Mode Imputation: Replace missing entries with the average (mean), middle value (median), or most frequent value (mode) of the corresponding column. This is a quick and easy approach, but it can introduce bias if the missing data is not randomly distributed.

K-Nearest Neighbors (KNN Imputation): This method finds the closest data points (neighbors) based on available features and uses their values to estimate the missing value. KNN is useful when you have a lot of data and the missing values are scattered.

Model-based Imputation: This involves creating a statistical model to predict the missing values based on other features in the data. This can be a powerful technique, but it requires more expertise and can be computationally expensive.

12. What is A/B testing?

Ans: A/B testing is a type of experiment in which you split your web traffic or user base into two groups, and show two different versions of a web page, app, email, and so on, with the goal of comparing the results to find the more successful version. The data science behind A/B testing can get complex pretty quickly. But, we've highlighted a few need-to-know terms to start with the basics. The null hypothesis, or H0, posits that there is no difference between two variables. In A/B testing, the null hypothesis would assume that changing one variable on a web page (or marketing asset) would have no impact on user behavior.

13. Is mean imputation of missing data acceptable practice?

Ans: mean substitution is a poor method for handling missing data, whereas both multiple imputation and full information maximum likelihood are recommended alternatives to this approach.

14. What is linear regression in statistics?

Ans: Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation. For instance, suppose that you have data about your expenses and income for last year. Linear regression techniques analyze this data and determine that your expenses are half your income. They then calculate an unknown future expense by halving a future known income.

At its core, a simple linear regression technique attempts to plot a line graph between two data variables, x and y. As the independent variable, x is plotted along the horizontal axis. Independent variables are also called explanatory variables or predictor variables. The dependent variable, y, is plotted on the vertical axis. You can also refer to y values as response variables or predicted variables.

15. What are the various branches of statistics?

Ans: Statistics have majorly categorised into two types:

Descriptive statistics
Inferential statistics

Descriptive Statistics:-
In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

Measure of frequency
Measure of dispersion
Measure of central tendency
Measure of position
The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Inferential Statistics:-
This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.