Batch- DS2307 **Submitted By-Urvashi Gupta**

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: c) The square of a standard normal random variable follows what is called chi-squared distribution.

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: The most significant probability distribution in statistics for independent, random variables is the normal distribution, sometimes referred to as the Gaussian distribution. In statistics, its well-known bell-shaped curve is generally recognised. The majority of the observations are centred around the central peak of the normal distribution, which is a continuous probability distribution that is symmetrical around its mean. The probabilities for values that are farther from the mean taper off equally in both directions. Extreme values in the distribution's two tails are likewise rare. Not all symmetrical distributions are normal, even though the normal distribution is symmetrical.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans: The missing data handling is quite tricky sometimes. As data analyst or data scientist it is very important aspect for the modelling. This should not be ignored at any cost. Proper addressing of missing data is always advisable in the pre-processing steps of data handling. There are various imputation techniques which can be used to fill the missing values in the data. Few of popular techniques is to fill the missing data with mean, mode or median of the respective column as per the requirement. And the choice of imputation technique mostly based on the type of dataset being processed.

12. What is A/B testing?
Ans: A/B testing, also known as split testing, is a statistical method used to compare two versions of a variable (typically a web page, advertisement, or other digital content) to determine which one performs better. It's commonly used in marketing and product development to make data-driven decisions about changes to user experiences. In A/B testing, a randomly selected group of users is divided into two segments: Group A and Group B. One group is exposed to the original version (control), while the other group is exposed to a modified version (treatment). The goal is to measure any differences in user behavior, engagement, or other relevant metrics between the two groups.

13. Is mean imputation of missing data acceptable practice?
Ans: Yes it is acceptable practice.

14. What is linear regression in statistics?
Ans: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is used to find the line of best fit through a set of data points, which can then be used to make predictions about future observations.
The line of best fit is represented by the equation: $Y = mX + b$
where Y is the dependent variable, X is the independent variable, m is the slope of the line and b is the y-intercept. The goal of linear regression is to find the values of m and b that minimize the difference between the predicted values of Y and the actual values of Y. Linear regression can be used for both simple linear regression (one independent variable) and multiple linear regression (multiple independent variables).To find the slope (m) of the line of best fit, the following equation is used: $m = (n\sum(xy) - (\sum x)(\sum y)) / (n\sum(x^2) - (\sum x)^2)$
where n is the number of data points, x and y are the independent and dependent variable. The y-intercept (b) is given by the following equation: $b = (\sum y - m\sum x) / n$

15. What are the various branches of statistics?

Ans: The various branches of statistics are: Descriptive, Inferential, Predictive, Prescriptive