STATISTICS WORKSHEET 1

Q1 to Q9 have only one correct answer

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 :- A) Modeling event/time 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 :- D) None of the mentioned

Q10 and Q15 are subjective answer type questions.

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

Ans :- Normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme and probability distribution that (roughly) describes many common datasets in the real world. It is the most common type of distribution, and it arises naturally in statistics through random sampling techniques. A normal distribution is a bell curve because of its flared shape. It is defined by the mean and standard deviation of data set. The normal curve is a probability distribution with a total area under the curve of 1.

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

Ans :-

12) What is A/B testing?

Ans :- AB test is an example of statistical hypothesis testing**,** a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not. A/B tests consist of a randomized experiment that usually involves two variants (A and B), although the concept can be also extended to multiple variants of the same variable. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics.

13) Is mean imputation of missing data acceptable practice?

Ans :- 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 eight-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. Second, 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 :- Linear regression analysis is used to predict the value of a variable based on the value of another variable. Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variable, one variable is called an independent variable, and the other is a dependent variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. The case of one explanatory variable is called simple linear regression, for more than one, the process is called multiple linear regression. Linear regression is commonly used for predictive analysis and modeling.

15) What are the various branches of statistics?

Ans :- There are two main branches of statistics.

1) Descriptive statistics
2) Inferential statistics

- Descriptive statistics – If we are able to describe the things and deals with describing of data using measure of central tendency (Mean, Median, Mode) and measure of dispersion (spread) (Variance and Standard Deviation). Descriptive statistics are use to get a full information of data. We can have the information of data in numerical, graphical, charts and tables form.

  Mean, Median and Mode are measure of central tendency. These are use by examine the value distribution centre.

  ● MEAN = Mean means average value of Dataset/column/series.
  ● MEDIAN = Median means centre point/value of Dataset/column/series.
  ● MODE = Mode means the number is maximum time in series/column.

  Variance and Standard Deviation are measure of dispersion. These are used to analyse the distribution of particular data.

  ● Variance = Variance means how much the data is varying from each other in particular column.

● Standard Deviation = Standard deviation means the main population is within plus or minus the standard deviation from the average.

- Inferential statistics – If we are not able to describe the population. We took only some samples and then we will describe that and then infer to the whole population. Inferential statistics are use of compare, test and predicts the data.
  Two types of inferential statistics. 1) Estimation of parameters and 2) Testing of hypothesis. We can have the information of data probability score.