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

Answer – 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

Answer – a. Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

   a. Modelling event/time data

   b. Modelling bounded count data

   c. Modelling contingency tables

   d. All of the mentioned

Answer – b. Modelling 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

Answer – d. All of the mentioned

5. _____ random variables are used to model rates.

   a. Empirical

   b. Binomial

   c. Poisson

   d. All of the mentioned

Answer – c. Poisson

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

Answer – 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

Answer – b. Hypothesis

8. Normalized data are cantered at _____ and have units equal to standard deviations of the original data.
   a. 0
   b. 5
   c. 1
   d. 10

Answer – 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.

Answer – c. Outliers cannot conform to the regression relationship

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

It is bell shaped frequency distribution curve in which, most of data points lies. It is divided into three parts and those states that 68% of data lies in ±1σ, 95 % of data lies in ±2σ and 99% of data lies in ±3σ (where σ is standard deviation). Most of algorithms follow normal distribution curve, so statisticians try to convert their model into normal distribution so they can apply required algorithm on data.

It helps to identify skewness of data (plotting mean, median and mode)

To find and eliminate outliers.

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

Most of missing data received is collected by secondary resources and dealing with such poor-quality data is very important. Because without loosing data, working on data so as to get required trained mode it difficult.

In ML, before applying any model, it is important to clean data. So, to handle missing values below approaches can be used.

1. Training secondary resources before collecting data
2. Using Data validation techniques for accepting online forms/feedback
3. Studying on variables before collecting data
4. Identifying engaging population to collect samples
5. Using imputation techniques
6. If having less/very few missing values then dropping records with missing values.

Imputation Techniques

1. Hot/cold deck imputation – As we trying to handle missing value considering all other features of record [at random/ systematically].
2. Regression imputation – Using predicted value by keeping relation.
3. Interpolation
4. Multiple imputations

12. What is A/B testing?

A/B testing is an optimization technique used on live data for analysing effect of variables on user engagement.

13. Is mean imputation of missing data acceptable practice?

For handling missing data, mean imputation is not an acceptable practice. Because when we use mean imputation, we are not preserving relationship among variables, and leads our model to less accuracy.

14. What is linear regression in statistics?

Linear regression –

Linear regression is basic and widely used Machine learning algorithm used for numerical data to identifying current trends in data and forecasting future trends in data based on current algorithm.

Linear regression model consists of developing mathematical equation to find value of dependent variable (**Y**) when given independent variable (**x**) and model developed coefficient (**m**) with constant (**c**). The equation turns out to be

$$\mathbf{Y = mx + c}$$

There are two types, simple linear regression analysis where there is only one independent variable.

$$\mathbf{Y = mx + c}$$

Then multiple linear regression, where there are more than two independent variables ($\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$ **and** $\mathbf{x_4}$) are considered for developing regression equation along with constant.

$$\mathbf{Y = \beta x_1 + \beta x_2 + \beta x_3 + \beta x_4 + c}$$

15. What are the various branches of statistics?

There are two branches of statistics.

1. Descriptive Statistics – It is branch of statistics that deals with collection, summarizing, organising and representing data. This includes – central tendency, standard deviation, variance, co-relation and various charts/plots.
2. Inferential Statistics – It branch of statistics, that comes into picture after descriptive statistics. In inferential statistics, we conclude/infer results for population based on sample. Inferential statistics consists of sampling methods, hypothesis testing etc.