Statistics Worksheet – 1

Q1. a) True

Q2. a) Central Limit Theorem

Q3. b) Modeling bounded count data

Q4. d) All of the mentioned

Q5. c) Poisson

Q6. b) False

Q7. b) Hypothesis

Q8. a) 0

Q9. c) Outliers cannot conform to the regression relationship

Q10. There are various types of distribution in statistics. But most commonly used distribution is normal distribution and z distribution in our course.

Normal Distribution:-

A normal distribution is the proper term for a probability bell curve.
In a normal distribution the mean is zero and the standard deviation is one
skewed right, symmetric distribution, skewed left.

Q11. Firstly if the data is not important or it if it don't affect the label data then surely we can ignore it. If the data having null values is of object type then it is not good to impute them. On the overall observation it is upto us that which nulls w should fill or replace or encode or just ignore them.

Imputation techniques :-

There are several imputation techniques but I use to recommend following techniques :-

1. K Nearest Neighbors Imputer
2. Iterative Imputer

Q12. A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric.

A/B testing is a form of :-

1. Statistical hypothesis testing – It is a method in which a sample dataset is compaired against the population data.

Two-sample hypothesis testing – Determines whether the difference between the two samples are statistically significant or not.

Q13. As mean imputation ignores feature correlation so it is not a good solution on filling null or missing values in given dataset.

Q14. Linear Regression ((y=mx+c))

It is one of the most fundamental and widely known ML algorithms

Bulding block of Linear Regression are:-

. Descrete or continous independent variables

. A best fit regression line

. Continuous dependent variable ie a linear regression model predicts the dependent variable using a regression line based on the independent variables. The equation for the Linear Regression is-

$Y = a+b*X + e$

Where

a = intercept

b = slope of the line

e = the error term

The equation above is used to predict the value of the target variable based on the given predictor variable(s)

Q15. There are three main branches of stats that are :-

1 . Data Collection – It is all about hoe the actual data is collected, there are significant issues to consider when actually collecting data.

2. Discriptive statistics : It presents the given data either visually or numerically.
3. Inferential statistics : Finally in this the conclusions made on given dataset.