STATISTICS -WORKSHEET 1

Ques.1– Ques.9 (choose one correct answer)

Ques1: Bernoulli random variable take (only) the values 1 and 0.

Ans1: a) True

Ques2: Which of the following theorem states that the distribution of average of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

Ans2: a) Central Limit Theorem

Ques3: Which of the following is incorrect with respect to use of Poisson distribution?

Ans3: b) Modeling bounded count data

Ques4: Point out the correct statement.

Ans4: d) All the mentioned

Ques5: _____ random variables are used to model rates.

Ans5: c) Poisson

Ques6: Usually replacing the standard error by its estimated value does change the CLT.

Ans6: b) False

Ques7: Which of the following testing is concerned with making decision using data?

Ans7: b) Hypothesis

Ques8: Normalized data are centered _____ and have units equal to standard deviations of the original data.

Ans8: a) 0

Ques9: Which of the following statement is incorrect with respect to outliers?

Ans9: c) Outliers cannot conform to the regression relationship

Ques.10-Ques.15

Ques10: What do you understand by the term Normal Distribution?

Ans10: Normal distribution is graphically represented as a probability belled shape curve so it is known as bell curve and also known as Gaussian distribution. It is symmetric about its mean, median and mode which is

$$\text{Mean} = \text{Median} = \text{Mode}$$

The parameters of normal distribution is mean μ and variance σ^2. Normal distribution works on numerical data not on a categorical data. If normal distribution transform into standard normal distribution so mean is equal to 0 and variance is equal to 1.

Ques11: How do you handle missing data? What imputation techniques do you recommend?

Ans11: We can handle missing data by imputing some new data in place of NAN. The imputation is possible if data is of integer or float type, but if it is of string type so we can't apply. It is not an easy task to give a new value because the value we insert that must have a relation with all values corresponding to that particular column values.

So, we have an imputation techniques to impute a missing data:

- Mean (): If our data is continuous so we calculate its mean and then replace NAN by the new value.
- Mode (): If our data is discrete (1, 2 …) so we cannot use mean because it gives any value (i.e., integer or float) or may be possible that it doesn't belongs to that range. So we use mode in such type of data which gives the highest frequency value.

Ques12: What is A/B testing?

Ans12: A/B testing is the comparison of two variables A and B & may be possible for more than two variables. In statistics, it termed as a statistical significance where we test the variables and decide which one is the better than the other one. The testing works on a percentage and whose percentage value is higher than the other value that variable is consider as the best option to choose and analyse.

Ques13: Is mean imputation of missing data acceptable practice?

Ans13: No, mean imputation is not always applicable in handling the missing values. It depend upon how the independent variable or label data is distributed.

If our data is continuously distributed so we can use mean imputation to get the value for a missing term. But if the data is of discrete type so we can't use it because it take range value(1,2,3 …) and by using mean it give any value which may or may not belong to that range or may give unexpected value. So, in discrete case we apply mode imputation to get better result. That's why mean is not always acceptable for the handle the missing value.

Ques14: What is linear Regression in statistics?

Ans14: Linear regression is a model in which we estimate the relationship between independent variable and dependent variable where dependent variable gives the prediction value on the bases of another independent variable. Linear regression is straight line and the equation of a linear regression is

$$Y = a + bX$$

where, Y = dependent variable

X = independent variable

a = intercept of Y

b = slope of the line

There are two types of linear regression

i. Simple Linear Regression (one dependent and one independent variable)
ii. Multiple Linear Regression (one dependent and more than one independent variable)

Ques15: What are the various branches of statistics?

Ans15: Statistics is mathematical tool in which we collect the data, analyse the data, interpret the data and in the last or final we have to make a better decision from it. Statistics is further classified into two branches:

1. Descriptive Statistics
2. Inferential Statistics

Descriptive Statistics:

If data is small so we can described the data without any statistical tools, then it is called descriptive statistics.

Inferential Statistics:

If our data is too large then we take few sample from the population and find averages of each sample on that basis we classified our data set. This is termed as inferential statistics.