STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
Ans:- 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?
Ans:- Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
Ans:-Modeling bounded count data

4. Point out the correct statement.
Ans:- All of the mentioned

5. ……random variables are used to model rates.
Ans:- Poisson

6**. Usually replacing the standard error by its estimated value does change the CLT.**
Ans:-False

7. Which of the following testing is concerned with making decisions using data?
Ans:-Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.

Ans:-0

9. Which of the following statement is incorrect with respect to outliers?

Ans:-Outliers cannot conform to the regression relationship

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

Ans: -Normal distribution is also known as gaussian distribution, it is important probability distribution in statistics for independent variable. Most people recognized it for its familiar bell-shaped curve in the statistical report.
Normal distribution is continuous probability distribution, normal distribution is symmetrical but not all symmetrical distribution is normal. The parameters define the shape and probabilities of normal distribution. There are two parameters mean and standard deviation.
Mean defines the location of peak while standard deviation defines measures of variability. The normal distribution mean median and mode are equal, half of population is less than the mean while half of population is greater than the mean.
There is empirical rule of distribution of data, where the data is distributed normally and standard deviation become particularly valuable where the data is fall within the specific no. of standard deviation from the mean.
It is important distribution because it accurately describes the value for many natural phenomena

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

Ans:-missing data is inevitable part of the process. No matter now accurate the data is pouring no matter how many times we clean the it, and prepare it we still get the data missing in data. There are ways to handle it to work and make the contingency plan

Missing data :-there are three types of missing data
1.MACR: -missing at completely random
2.MAR: - missing at random
3.NMAR: - not missing at random

Some techniques are :-

Using dealation method :-
There are several deletion methods two common one includes listwise deletion and pairwise deletion. It means deleting any large no. of data which deleting the data do not affect the volume of data
Using imputation techniques:-
The imputation techniques replacing the missing values with substituted values
There are some imputation techniques are follows -
Mean imputation – take the mean for missing value and impute it

Substitution imputation – impute the value for which is not selected for sample

Hot deck imputation – as an example if we have to choose between the 0 to 9 then every time, we see the value is between the 0 to 9. It gives the range between the number
Cold deck imputation – it is similar to hot deck imputation, it just removes the random variation
Regression imputation – instead of taking the mean we take predicted value, of another variable
Stochastic regression imputation – predicted value of regression as well as value from random residual value. It has the advantage of regression and the residual value
Two types of imputation single and multiple
Multiple: - So multiple imputation comes up with multiple estimates. Two of the methods listed above work as the imputation method in multiple imputation–hot deck and stochastic regression. Because these two methods have a random component, the multiple estimates are slightly different.

For my view in general the values in dataset which is missing should be depend on domain of data. The behaviour in which data is acting is any data set, we need to understand that, we just cannot use one techniques of imputation for all data set, so it is difficult to argue which is best imputation technique, by the way few of them are used mostly like knn-imputation and random forest , it seen they often perform best

12. What is A/B testing?

Ans:- A/B testing is an example of statistical hypothesis testing , it also know as split testing , it is the hypothesis made about the relationship between two data set, and those data set compared against each other to determine if they have any statistical relationship or not , on the other hand it is the marketing techniques that involves comparing the two version of product which perform better. With an example suppose the company wants to know product is better, it will scientifically or statistically use the method or do the research to know what is the behaviour of consumer.
So, it takes the hypothesis for that null hypothesis and alternative hypothesis
If we are accepting the null hypothesis, it automatically rejecting the alternative hypothesis and if we accept the alternative, it rejects the null hypothesis.
Although the this is old technique but now a days it is eminent for online environment and for big data. It is easier to use to conduct the test and use the result for better experience.
We must know the reason behind the data, there are different tool for A/B testing

13. Is mean imputation of missing data acceptable practice?

Ans:- In world of data science , it is terreable practice, the mean or average in data in dataset reduced the variance, any data in data set should have high and low variance of data, which might be true as per the domain of data, while we use the average in data set which brings all the data into average and narrow the variance between the data, which is different from the actual data. Yes, it decreased the variance of data but increased the bias. So, the mean is considered the terrible practice

14. What is linear regression in statistics?

Ans:-linear regression is the starting point of econometric analysis. The linear regression model has a dependent variable that is a continuous variable, while the independent variables can take any form (continuous, discrete, or indicator variables). A simple linear regression model has only one independent variable, while a multiple linear regression model has two or more independent variables. The linear regression is typically estimated using OLS (ordinary least squares). Examples include studying the effect of education on income; or the effect of recession on stock returns.

15. What are the various branches of statistics?

Ans:- two major ways of classifying statistics: (i) on the basis of function and (ii) on the basis of distribution.

1.on the basis of function: - three types of statistics have been described

Descriptive statistics: The branch which deals with descriptions of obtained data is known as descriptive statistics. On the basis of these descriptions a particular group of population is defined for corresponding characteristics. The descriptive statistics include classification, tabulation measures of central tendency and variability. These measures enable the researchers to know about the tendency of data or the scores, which further enhance the ease in description of the phenomena

Correlational statistics: The obtained data are disclosed for their inter correlations in this type of statistics. It includes various types of techniques to compute the correlations among data. Correlational statistics also provide description about sample or population for their further analyses to explore the significance of their differences.

Inferential statistics: Inferential statistics deals with the drawing of conclusions about large group of individuals (population) on the basis of observations of few participants from them or about the events which are yet to occur on the basis of past events. It provide tools to compute the probabilities of future behaviour of the subjects

2.on the Basis of Distribution of Data

Nonparametric statistics are those statistics which are not based on the assumption of normal distribution of population. Therefore, these are also known as distribution free statistics. They are not bound to be used with interval scale data or normally distributed data.

Parametric statistics is defined to have an assumption of normal distribution for its population under study. Parametric statistics refers to those statistical techniques that have been developed on the assumption that the data are of a certain type. In particular the measure should be an interval scale and the scores should be drawn from a normal distribution.