STATISTICS WORKSHEET

1. Bernoulli random variables take (only) the values 1 and 0.
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
3. Which of the following is incorrect with respect to use of Poisson distribution?
A. Modeling bounded count data
4. Point out the correct statement
A. All of the mentioned
5. _____ random variables are used to model rates.
A. Poisson
6 .Usually replacing the standard error by its estimated value does change the CLT.
A. False
7. Which of the following testing is concerned with making decisions using data?
A. Hypothesis
8. Normalized data are centered at _____ and have units equal to standard deviations of the original data.
A. 0
9. Which of the following statement is incorrect with respect to outliers?
A. Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
A. Normal distribution also known as Guassian distribution. It is a continuous distribution. We can denote Normal distribution by 'N' followed by mean 'mu' and variance 'sigma^{2}'. When dealing with actual data we know the numerical values of 'mu' and 'sigma'. There are so many examples for normal distribution in real life. For example, the weight of the new born baby, amount of rainfall…. The graph of the normal distribution is Bell shaped. Therefore , majority of data is centered around the mean. The graph is symmetric regards to the mean, so we can say thatequally far away from opposing direction. Another pecularly of normal distribution is Threesigma rule/Emperical rule. This law suggests that for any normally distributed event 68% of all outcomes falls within one standard deviation, 95% of all outcomes falls within two standard deviation and 99.7% of all outcomes falls within three standard deviation. The remaining 0.3% of chance of outliers. It means the chance of outliers is very rare.

11. How do you handle missing data? What imputation techniques do you recommend?
A. There are so many imputation techniques for handling the missing data. If there are low number of null values or missing data, we can handle these situation by deleting the rows which have null values. But in such cases there are high number of null values or missing data, i9n such cases we can use 'Mean imputation' technique for handle the missing data. Mean imputation technique means it will find the mean of the observed data and add to the missing values.
I recommend Random forest technique and Arbitrary value imputation, mean or median imputation. Random forest is non-parametric imputation method applicable to various variable types that works well both data missing at random and not missing at random. In arbitrary imputation, it consists of replacing all occurances of missing values within a variable. It is suitable for both numerical and categorical variables. When the dataset is large also the missing value is larger the simplest technique is the mean imputation technique. It give simplest possible approach and does not introduce any bias to the dataset.

12. What is A/B testing?
A. A/B testing is basically statistical hypothesis testing. It is an analytical method for making decisions that estimate population parameters based on sample statistics. In other words, it is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

13. Is mean imputation of missing data acceptable practice?
A. Yes, mean imputation is acceptable testing. When dataset is large and the missing values or null values are larger number, the best imputation technique is mean imputation technique. It gives the simplest approach for dataset. But in other words, basically everything have advantages and disadvantages. When we look for the disadvantages, in this technique it does not show the relation among variables, it only shows the result of the mean of observed data. So we can conclude that, mean imputation technique is acceptable.

14. What is linear regression in statistics?
A. Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variable. There are two types of linear regression:
1).Simple linear regression
2).Multiple linear regression
Simple linear regression contains one explanatory variable and Multiple linear regression contains one or more explanatory variable. In other words, it attempts to the model the relationship between two variable by fitting a linear equation to observed data. A linear regression line has an equation of the form, y=a + bx , where 'x' is the explanatory variable and 'y' is the dependent variable.

15. What are the various branches of statistics?
A. The major branches of statistics are:
1). Descriptive statistics
2). Inferential statistics

Descriptive Statistics

It is the one of the branch of statistics. Descriptive coefficients that summarize a given dataset, which can be either a representation of the entire population or a sample of population. There are two types of descriptive statistics: a).Measure of central tendency b).Spread of data
Measure of central tendency includes Mean , Median and Mode. While spread of data contains Variance, standard deviation, range, percentile, skew.

Inferential Statistics

It is another branch of statistics. It is the process of using data analysis to infer properties of an underlying distribution of probability. It analysis infers properties of a population.