STATISTICS WORKSHEET-1

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?

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

5._____ random variables are used to model rates

c) Poisson

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

b) False

7.Which of the following testing is concerned with making decisions using data?

b) 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?

c) Outliers cannot conform to the regression relationship

10.What do you understand by the term 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 1.It has zero skew and a kurtosis of 3.

Normal distributions are symmetrical, but not all symmetrical distributions are normal.

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

It can be done buy imputation. Imputation is the process of replacing missing values with substituted data. It is done as a pre-processing step.

Mean, Median and Mode. This is one of the most common methods of imputing values when dealing with missing data.

Model-Based Imputation. We take feature f1 as the class and all the remaining columns as features. Then we train our data with any model and predict the missing values.

12.What is A/B testing?

A/B testing is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behaviour. Also known as

split tests, allow you to compare 2 versions of something to learn which is more effective. A/B tests are widely considered the simplest form of controlled experiment.

13.Is mean imputation of missing data acceptable practice?

Bad practice in general

If just estimating means: mean imputation preserves the mean of the observed data

Leads to an underestimate of the standard deviation

Distorts relationships between variables by “pulling” estimates of the correlation toward zero

14.What is linear regression in statistics?

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is an explanatory variable, and the other is a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

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. The slope of the line is b, and a is the intercept (the value of y when $x = 0$).

15.What are the various branches of statistics?

Descriptive Statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation. Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures.

Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.