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) Modeling bounded count data

4. Point out the correct statement

d) All of the mentioned

5. ______ random variables are used to model rates.

c) Poisson

6.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?

Normal Distribution is a bell-shaped frequency distribution curve which helps describe all the possible values a random variable can take within a given range with most of the distribution area is in the middle and few are in the tails, at the extremes.

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?

Datasets may have missing values, and this can cause problems for many machine learning algorithms.As such, it is good practice to identify and replace missing values for each column in your input data prior to modeling your prediction task. This is called missing data imputation, or imputing for short.

Mostly used imputation techniques are

- Iterative Imputation
- knn imputation
- mean and median and mode

- Iterative Imputation

Iterative imputation refers to a process where each feature is modeled as a function of the other features, e.g. a regression problem where missing values are predicted. Each feature is imputed sequentially, one after the other, allowing prior imputed values to be used as part of a model in predicting subsequent features.

This approach may be generally referred to as fully conditional specification (FCS) or multivariate imputation by chained equations (MICE).

- knn imputation

One popular technique for imputation is a K-nearest neighbor model. A new sample is imputed by finding the samples in the training set "closest" to it and averages these nearby points to fill in the value and it is also known as "Nearest Neighbor Imputation" .

knn imputer will try to find the relationship between other features and impute the data using Euclidean distance measuring method to calculate the distance between members of the training dataset. KNNImputer can work with continuous, discrete and categorical data types

- Mean and median and Mode

This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.and for categorical data we can use mode of the colum of data .This is one of the basis method of handing the missing data

12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment

For instance, let's say you own a company and want to increase the sales of your product.It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

A/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Here you may compare pre and post-modification results to decide whether the changes are working as desired or not

- Before conducting an A/B testing, you want to state your null hypothesis and alternative hypothesis

- the next step is to create your control and test (variant) group. Random sampling helps to eliminate bias because you want the results of your A/B test to be representative of the entire population rather than the sample itself.

- Once you conduct your experiment and collect your data, you want to determine if the difference between your control group and variant group is statistically significant

14.Is mean imputation of missing data acceptable practice?

Mean imputation is typically considered terrible practice , mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Mean imputation does not preserve the relationships among variables. Any statistic that uses the imputed data will have a standard error that's too low.It also distorts relationships between variables by "pulling" estimates of the correlation toward zero

15. What is linear regression in statistics?

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables ). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression

15. What are the various branches of statistics
Statistics can be divided into 2 branches, they are

• Descriptive Statistics. Descriptive statistics is the first part of statistics that deals with the collection of data, summarize and organizing characteristics of a data set.

Example- political polling

• Inferential Statistics. The inferential statistics is a branch of statistics that makes the use of various analytical tools to draw inferences about the population data from sample data and help you decide whether your data confirms or refutes your hypothesis and whether it is generalizable to a larger population