STATISTICS WORKSHEET

Q1-> Bernoulli random variables take (only) the values 1 and 0?

Ans->true

Q2-> 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

Q3-> Which of the following is incorrect with respect to use of Poisson distribution?

Ans-> Modeling bounded count data

Q4-> Point out the correct statement

Ans-> All of the mentioned

Q5-> _____ random variables are used to model rates

Ans-> Poisson

Q6-> 10. Usually replacing the standard error by its estimated value does change the CLT

Ans->false

Q7-> 1. Which of the following testing is concerned with making decisions using data?

Ans-> Hypothesis

Q8->. Normalized data are centered at_____and have units equal to standard deviations of the original data

Ans->0

Q9-> Which of the following statement is incorrect with respect to outliers?

Ans-> Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly

Q10-> What do you understand by the term Normal Distribution?

Ans->A normal distribution is the proper term for a probability bell curve .In a normal distribution

The mean is zero. normal distribution is the most common type of distribution assumed in statistical analyses. However real life data rarely follows a perfect normal distribution . The standard normal distribution has two parameter the mean and standard deviation .For a normal distribution ,68% of the obserbation are with in +/- one standard deviation of the mean , 95% are with in +/- two standard deviation and 99.7% are within +/- are thrid standard deviation

The skewness measure how different the distribution is form a normal distribution and skewness measure symmetry of distribution .The normal distribution is symmetic and has a skewnessof zero.IF the distribution of a data set has a skewness less than zero , or negative skewness ,then the left tail of the distribution is longer than the right tail .positive skewness impiles that the right tail of the distribtuion is longer than the left.if kurthosis statistic measure a thickness of the tail ends of a distribtuion in relation to the tails of the normal distribution and noraml distribution model is motivated by the Cental limit theoram .

Q11-> How do you handle missing data? What imputation techniques do you recommend?

Ans->first see different types of missing data
i) Missing at Random (MAR):
Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.

ii) Missing Completely at Random (MCAR):
The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

iii) Missing not at Random (MNAR):
Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

In the first two cases, it is safe to remove the data with missing

values depending upon their occurrences, while in the third case removing observations with missing values can produce a bias in the model. Very basic type of imputation technique is Computing the overall mean, it is the only tested function that takes no advantage of the time series characteristics or relationship between the variables. It is very fast, but has clear disadvantages. One disadvantage is that mean imputation reduces variance in the dataset. There are other machine learning techniques like XGBoosting and Random Forest for data imputation but personally my favourite imputation technique is KNN(K Nearest Neighbor) and also it is widely used. In this method, k neighbour are chosen based on some distance measure and their average is used as an imputation estimate. The method requires the selection of the number of nearest neighbour, and a distance metric. KNN can predict both discrete attributes (the most frequent value among the k nearest neighbour) and continuous attributes (the mean among the k nearest neighbour)

Q12-> What is A/B testing?

Ans-> 1. A/B testing is basically used to compare two different products those comparison is made from the user input, whether the user is clicking on product 1 or user is clicking on product 2.

2. Technically we can say that 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.

3. For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

4. In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

5. 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.

6. 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.

7. A/B testing doesn't work well when testing major changes, like new products, new branding, or completely new user experiences. In these cases, there may be effects that drive higher than normal engagement or emotional responses that may cause users to behave in a different manner.

Q13-> Is mean imputation of missing data acceptable practice?

Ans-> *1. Mean imputation is a very basic type of imputation technique.*

2. It is the only tested function that takes no advantage of the time series characteristics or relationship between the variables.

3. Mean imputation reduces the variance of the imputed variables.

4. Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.

5. Mean imputation does not preserve relationships between variables such as correlations.

6. Thus for a professional Data Scientist mean imputation of missing data is not at all acceptable practice.

Q14-> What is linear regression in statistics?

Ans->Linear regression is one of the most fundamental algorithm in the machine learning world.It is the door of the magical world .Linear regression used to study linear relationship between a dependent variable y(bloodpressure) and one or more independent variable(age,sex,weight).More generally ,a linear makes a prediction by simple computing sum of the input features ,plus a constant called bias term ,also called intercept term.There are two types of linear regression one is simple linear regression and other one is multiple linear regression

In the simple linear regression ,we predict score on one variable from the score on a second variable the variable we are predicting is called criterion variable and is referred as Y .The variable we are basing our prediction is called the predictor variable is

referred as x .When there is only one predictor variable ,the predicting method is called simple regression in simple linear regression the prediction of y when plotted as the function of x from a straight line .when there is multiple predictor variables ,the prediction mrethod is called multiple linear regression .The statistical formula for multiple linear regression is y = 0 + 1×1 + 2×2 + pxp.

Q15-> What are the various branches of statistics?

Ans-> There are two branches of statistics they are:-
i) Descriptive statistics
ii) Inferential statistics

Descriptive statistics:-

It organizes raw data into meaningful information .An house hold articles manufacturing company would like to know what people feel about their products. For that purpose, the company forms a team of people and tries to collect information from the public. The team of people formed by the company is trying to collect data from the public directly. The data which is being collected directly from the public will always not be meaning full. Hence, the data which is being collected directly from the public has to be converted in to meaningful information. This is the work being done in this particular branch " descriptive-statistics". That is,it focuses on collecting, summarizing and presenting set of data.

For example, Industrial statistics, population statistics, trade statistics etc.

Inferential statistics:-

It analyses sample data to draw conclusion about population. Marketing research team of a company wants to know how far the people need a particular product manufactured by the company. There is one hundred thousand population in a particular city. It is bit difficult to go and ask all one hundred thousand people, due to time consumption and other factors. Hence, it takes a sample of 1000 people to draw conclusion for the whole population. That is making general statement from the study of particular cases or any treatment of data, which leads to prediction or inference concerning a larger group of data.

For example, we want to have an idea about percentage of illiterates

in a country. We take a sample from a population and the proportion of illiterates in the sample. That sample with the help of probability enables us to find the proportion to the original population.