**********Statistics Worksheet1**********

Answer 1. **A) True** [Bernouli, is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability (q = 1-p)]

Answer 2. **A) Central Limit Theorem** (Mean, Median, Mode)

Answer 3. **B) Modeling Bounded count data** (Poisson distribution is used for modeling unbounded count data)

Answer 4. **D) All of the mentioned** (Many of random variables, property normalized, limit to a normal distribution)

Answer 5. **C) Poisson** (Poisson distribution is used to model counts)

Answer 6. **B) False** (Because, usually replacing the standard error by its estimated value does not change the CLT)

Answer 7. **B) Hypothesis** (The null hypothesis consider as true and evidence of statistics, required to reject it in favor of an alternate hypothesis)

Answer 8. **A) 0** (In statistics, normalization can have a range of meanings.)

Answer 9. **C) Outliers cannot conform to the regression relationship**

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

************Normal Distribution************

- Normal distribution also known as Gaussian distribution and bell curve.

- It is a symmetrical, bell shaped distribution in which the mean, median and mode are equal. It always has a mean of Zero and standard deviation of one.

- The normal distribution is fully characterized by its mean & standard derivation. This makes the distribution symmetric and it is depicted as a bell-shaped curve when plotted. A normal distribution is defined by a mean (averages) of zero and a standard deviation of 1, 0.

- The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.

- In a normal distribution , approximately 68% of the data collected will fall within +/- one standard deviation of the mean: approximately 95% within +/- two standard deviations and 99.7% within three standard deviation

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

************Handling missing data************

We can handle missing data by performing following operations

- Ignore missing values row / delete row
- Fill missing value manually
- Global constant
- Measure of central tendency (mean, median & mode)
- Measure of central tendency for each class

- Most probable value (ML Algorithm) follow the rule (GIGO : Garbage in Garbage Out)

Ignore missing values row / delete row

1. In this operation, we remove the entire row when any NaN value present and if we find maximum number of cells NaN in a particular column then we delete that column.

2. Generally, we perform this drop operation on large datasets because, if we remove few rows or columns then large dataset will not be in high loss percentage of data.

Fill missing value manually

- Generally, we don't perform this manually filling of NaN data, because maximum time we work on large dataset which contain 500, 1000, 2500 rows, so, this operation can be applied only to the small datasets, not to the large datasets.

Global constant

- In this method we defined a value and then replace that value with all NaN cells in one operation.

Measure of central tendency (mean, median & mode)

- In this operation, we find mean, median and mode for each column of dataset and then we replace it with their NaN values of columns.
- We choose to find mean, median or mode as per their sns.distplot.
- For character variable we use mode value to replace with NaN

Measure of central tendency for each class

- In this method, we don't find mean or median for whole columns, we find mean or median for different groups of a particular column and then we replace it with their respective NaN values of different groups.

Most probable value (ML Algorithm) follow the rule (GIGO : Garbage in Garbage Out)

- In this operation, we train the Model with existing dataset rows which are Not NaN and then we predict the value for NaN data by using trained model of ML.

- This method is time consuming and costly to implement.

Conclusion: As I observed, drop row and column method is good for large data. But if we have to deal with small dataset then we should use Measure of central tendency for each class, in which no loss of data occurs and imputation occurs based on group.

Eg. Imputation based on class:

A	10
A	9
A	
A	6
C	5
C	6
C	3
C	

For this dataset we will find separate mean for A Group and B group then we impute that mean into this Nan value as per their respective group. We should use this technique when dataset is small.

Directly apply mean imputation will not be good, because it will ignore variable correlation.

If we find any dataset which is too large, mean so much variables exist in that dataset then we apply PCA (Principle Component Analysis) to prevent from the overfitting the model.

Answer 12.

*******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.

It is an experiment for determining which of different variations of an online experience performs better by presenting each version to users at random and

analyzing the results. So, A/B testing demonstrates the efficacy of potential changes, enabling data-driven decisions and ensuring positive impacts.

Lets we understand thid A/B testing by an example:

Suppose an e-commerce company XYZ. It wants to make some changes in its newsletter format to increase the traffic on its website. It takes the original newsletter and marks it A and makes some changes in the language of A and calls it B. Both newsletters are otherwise the same in color, headlines, and format.

Our objective here is to check which newsletter brings higher traffic on the website i.e the conversion rate. We will use A/B testing and collect data to analyze which newsletter performs better.

The user experience is the main aspect to achieve high end targets. There are times when a user does not like a change on the newspaper. When changes leads to less traffic then we don't implement changes in older version.

Answer 13: Is mean imputation of missing data acceptable practice?

Mean imputation of missing data is **not acceptable practice**

In general, quick and easy method is to substitute a mean for numerical data and use a mode for categorical ones. But this is not a good practice because mean and mode ignore variable correlations and when we look for variance, it also reduce a variance of the data while increasing bias. As result of the reduce variance, the model is less accurate and the confidence interval is narrow.

Answer 14: What is linear regression in statistics?

************Linear Regression in Statistics ************

It is a Machine learning model which is used to measure the relationship between variable or predictive analysis, in this model we estimates the relationship between independent variable and dependent variable using a straight line.

When there is a single input variable, the regression is referred to as Simple Linear Regression. We use the single variable (independent) to model a linear relationship with the target variable (dependent). We do this by fitting a model to describe the relationship. If there is more than predicting variable, the regression is referred to as Multiple Linear Regression.

Regression Coefficients

When performing simple linear regression, the four main components are:

Dependent Variable or Target variable: will be estimated and predicted

Independent Variable or Predictor variable: used to estimate and predict

Slope or Angle of the line: denoted as m or β

Intercept: Where function crosses the y-axis / denoted as c

Answer 15: What are the various branches of statistics?

***********Various branches of statistics************

The two main branches of Statistics are **descriptive statistics**, which describe the properties of sample and population data, and **inferential statistics**, which uses those properties to test hypothesis and draw conclusion.

Descriptive statistics are brief descriptive coefficient that summarize a given data set, which can be either a representation of the entire population or a sample of a population. It broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median and mode, while measures of variability include standard deviation, variance, minimum and maximum variables and skewness.

In short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics.

Inferential statistics

We usually cannot measure the entire population that we are interested in, so we measure a subset which infer population, and then use statistics to make predication about the population as a whole.

These prediction are based on probabilities given the information that we collected.

Inferential statistics is mainly used to derive estimates about a large group (or population) and draw conclusions on the data based on hypotheses testing methods.

Inferential statistics uses sample data because it is more cost-effective and less tedious than collecting data from an entire population. It allows one to come to reasonable assumptions about the larger population based on a sample's characteristics. Sampling methods need to be unbiased and random for statistical conclusions and inferences to be validated.