STATISTICS WORKSHEET-1

S.NO	Questions	Answers
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	d) All of the mentioned
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 is the correct answer
7	1. Which of the following testing is concerned with making decisions using data?	b) Hypothesis is the correct answer
8	4. Normalized data are centered at_____and have units equal to standard deviations of the original data.	a) 0 is the right answer
9	Which of the following statement is incorrect with respect to outliers?	c) Outliers cannot conform to the regression relationship

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

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

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve. 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. In reality, most pricing distributions are not perfectly normal.

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

- First we have to find the null values in the dataset by using IsNull(),isnan().
- Then we will check the percentage of missing data because if more than 50 percent of the data in a particular row is missing then we will need to accumulate proper information for that particular column else deleting the entire column would be logical instead of treating it manually and giving incorrect data for the model to be trained and tested upon.
- If there is very little data missing then depending on the data can use mean, median and mode options to fill the correct data.
- The imputation techniques that I will be using are mean imputation, simple imputer, iterative imputer and knn imputer.

12. What is A/B testing?

A/B testing is also known as split/bucket testing mainly used when trying out a new feature on an existing product. It is similar to the concept of main and branch used in GitHub where we create branch for new feature changes and then merge them into the main section if things go well or keep the main untouched. In A/B testing we create sample of an entire population and then use the Hypothesis testing mechanism to check if our Null Hypothesis is correct or our Alternative Hypothesis is right. We need to check where we are able to reject the Null Hypothesis or whether we fail to reject the Null Hypothesis keeping in the mind the Type 1 and Type 2 errors that can be checked via the confusion matrix. In Null Hypothesis the important values considered are the alpha (allowable percentage of error), p value and the confidence level of the test model.

13. Is mean imputation of missing data acceptable practice?

According to the research studies even though Mean Imputation is a commonly used technique but is not a good practice to fill the missing data with the mean of all the other observations of a column which does mean correct always and moreover it increases the number of same data in a single column making the model biased towards the usage of the same mean value over other legitimate observations secured from proper data collection channels. There are other better imputation methods that provide accurate data for filling the missing value. However, if every other technique fails then mean imputation can be used as a last resort instead of deleting null values, especially for a smaller data set.

14. What is linear regression in statistics?

Linear regression strives to show the relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable. For example, the weight of the person is linearly related to his height. Hence this shows a linear relationship between the height and weight of the person. As the height is increased, the weight of the person also gets increased. The equation of Linear Regression is $y=mx+c$. where m is the slope, c is the intercept and x is the independent variable and y is the dependent variable.

15. What are the various branches of statistics?

The branches of statistics are

- Descriptive
- Inferential

Descriptive:

Here this data is summarised through the given observation. Summarization means statistical describing is given for the data by providing mean, mean, mode values, standard Deviation, Range, and others.

Inferential:

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions, or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.