Statistics Worksheet 1

Ans1: a) True

Ans2: a) Central Limit Theorem

Ans3: b) Modeling bounded count data

Ans4:

Ans5: c) Poisson

Ans6: b) False

Ans7: b) Hypothesis

Ans8: a) 0

Ans9: c) Outliers cannot conform to the regression relationship

Ans10: Normal Distribution: It 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. It is the continuous probability distribution with a probability density function that gives you a symmetrical bell curve. it is a plot of the probability function of a variable that has maximum data concentrated around one point and a few points taper off symmetrically towards two opposite ends.

Ans11: Missing data can be handled using various methods. That are

Mode substitution

Median substitution

By dropping values

Depending upon the type and quantity of data we need to take decision

i.e If out of 100 rows 2-5 rows has missing data, one can choose to drop these values

But if in same case 25-30 rows has missing data, one cannot drop these rows, giving away the crucial data.

So in such cases where we have data in categorical type one can go for “Mode Substitution”

Where the missing data will be filled with maximum occurring value.

And when data is continuous or in numerical format one can use median substitution for missing values.

Ans12: A/B Testing: A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric. Typically, two consumer groups are exposed to two different versions of the same thing to see if there is a significant difference in metrics like sessions, click-through rate, and/or conversions.

A/B testing is a form of statistical and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

Ans13: The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, 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.

Ans14: Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable. Linear regression are of two types simple and Multiple.

Simple linear regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable. Simple Linear Regression Equation is $y= b_o + b_1x$ where y is dependent and x is independent variable.

Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship. Multiple Linear Regression equation is

$y= b_o + b_1x_1 + b_2x_2 + b_3x_3 \dots . +b_nx_n$ where y is dependent and x1,x2,…..xn are independent variables

Ans15: There are two branches of statistics: descriptive statistics and inferential statistics.

Descriptive Statistics:

It deals with the presentation and collection of data. This is usually the first part of a statistical analysis.

Inferential Statistics:

As the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics.