STATISTICS WORKSHEET-1 ANSWERS

Q1) Option a

Q2) Option a

Q3) Option b

Q4) Option d

Q5) Option c

Q6) Option b

Q7) Option b

Q8) Option a

Q9) Option c

Q10) **Normal Distribution: -**

It is a probability distribution that is symmetric about the mean, which means the occurrence of the data near the mean is more frequent than it is far from the mean. It is also called Gaussian Distribution. And it is graphically represented by a 'bell curve'.

We say the distribution of data is normal when it is symmetric around the mean, when +-68% of the data lies within one standard deviation of the mean, 95% within two standard deviation of the mean and 99.8% of it within three standard deviations of the mean.

Q11) **Handling the missing data: -**

While dealing with the missing data, there are two methods which we should consider using to solve the error. But before that we need to analyze each column with missing values carefully to understand the reasons behind the missing values as it is crucial to find out the strategy for handling the missing values.

1) Deleting the missing values, rows and columns.

Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values.

And there is a risk that we might end up deleting the useful data from the dataset.

Similarly, we cannot delete an entire row or column if we see any missing values in them.

2) **Listwise deletion: -**

It removes all observations from your data, which have a missing value in one or more variables.

2) Imputing

It is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset.

Simple Imputing: -

Simple Imputer is a scikit-learn class which is helpful in handling the missing data in the predictive model dataset. It replaces the Nan values with a specified placeholder.

Mean Imputation: -

We can replace the integer data or continuous data with by finding the mean or median of that an attribute which contains missing values and replace them with either its mean or median

Regression Imputation: -

Regression imputation fits a statistical model on a variable with missing values. Predictions of this regression model are used to substitute the missing values in this variable.

If the data is categorical, we can replace the missing values with mod, i.e., the categorical data which has highest frequency.

In this way we can replace the missing values without affecting or deleting the useful data. In this way we can use the most out of data and predict accurate results.

Q11) A**/B Testing**

It is a statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not. A/B Test tries to calculate implied impact from the experiment testing two randomized different variants. Randomization helps to ensure bias is minimized but this isn't always easy to accomplish.

It allows decision makers to choose the best design by looking at the results obtained with two possible alternatives A and B.

To make sure that you wouldn't evaluate an experiment based on random results, statisticians implemented a concept called **statistical significance** — which is calculated by using something called **p-value.**

P-value is created to show you the exact probability that the outcome of your A/B test is a result of chance.

Q12) Mean imputation is a popular and easy method to deal with the missing values. Although it is an easy methos, it has some drawbacks as well.

It does not preserve the relationship among variable: -

Although by imputing the mean, you are able to keep your sample size up to the full sample size. if the data are missing completely at random, mean imputation will not bias your parameter estimate. It might still bias the standard error. Which eventually reduces the relationship among the variables. This can be a great error and lead to many problems with the data.

And it underestimates the standard errors as well. Another reason is applying to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low.

Q14) **Linear Regression: -**

It is a statistical method that relates a dependent variable to one or more independent or explanatory variables.

It is mainly used to examine two things

1. Does a set of predictor variables doing a good job in predicting the dependent variable.
2. What are the values or predictors which are significantly contributing to the dependent variable or outcome and in what way do they impact the outcome variable.

Linear Regression formula: - $y = a + bx$

y is the outcome or dependent variable, a is the intercept, b is the coefficient and x is the predictor or independent variables. There is also e which denotes the errors.

Q15) There are mainly 3 real branches of statistics: -

1. Data Collection: -

It is all about how the actual data is collected. This isn't very important for us in the math's perspective.

2. Descriptive Statistics: -

It is a part of statistics that deals with presenting the data we have either visually (via graphs, charts, etc.) or numerically (via averages and so on).

It is divided into two types: -

Central Tendency: - mean, median, mode

Dispersion of Data: - range, variance, standard deviation, percentiles/quartiles, skewness and Kurtosis

1. **Inferential Statistics: -**

It is a branch of statistics that makes the use of various analytical tools to draw conclusions about the population data from sample data.

It basically helps to develop a good understanding of the population data by analyzing the samples obtained from it. It helps in making generalizations about the population by using various analytical tests and methods.

This is divided into two types of testing and Analysis: -

Hypothesis Testing: -

Z-test, F-test, T-test, ANOVA test, etc.

Regression Analysis: -

Linear, Nominal, Logistic and Ordinal Regression