Statistics Worksheet

1. A

2. A

3. B

4. D

5. C

6. B

7. B

8. A

9. C

10.

Answer - Normal Distribution is also called as bell curve. 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.
In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew. The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

11.

Answer - Data can be missing in the following ways -

1. **Missing Completely At Random (MCAR):** When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random.
2. **Missing At Random (MAR):** The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data.
3. **Not Missing At Random (NMAR):** When the missing data has a structure to it, we cannot treat it as missing at random.

Imputation Techniques I recommends -

1. Mean or Median Imputation
2. Multivariate Imputation by Chained Equations (MICE)
3. Random Forest

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value.
In Pandas, there are two very useful methods: isnull () and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value you could use the fillna() method.

12.

Answer - A/B testing also known as split testing. An AB test is an example of statistical hypothesis testing, a process in which 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.

13.

Answer - It is a non-standard, it uses Random Forest. It is use to predict the missing data. It also can be used for both i.e. continuous as well as categorical data and so it makes advantageous over other imputations.

14.

Answer - Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Types of linear regression are as follows -

1. Simple linear regression.
2. Multiple linear regressions.
3. Logistic regression.
4. Ordinal regression.
5. Multinomial regression.

15.

Answer - Various branches of statistics are as follows -

1. Descriptive Methods -

- This type of method consists of all the preliminary steps to final analysis and interpretation. As such this method includes the method of collection, methods of tabulation, measures of central tendency, measures of dispersion, measures of skewness, and analysis of time series. These methods bring out the various characteristics of data and help in summarizing and interpreting the salient features of the data. This method is also called descriptive statistics.

2. Analytical Methods -
   - This type of method consists of all those methods which help in the matter of analysis and comparison between any two or more variables. This includes the methods of correlation, regression analysis, association of attributes. This method is also called analytical statistics.

3. Inductive Methods -
   - This type of method consists of all those procedures that help in the generalization or estimation over a phenomenon on the basis of random observation or partial data. This includes the procedure of interpolation, extrapolation and theory of probability. This method is also called inductive statistics.

4. Inferential Methods -
   - This type of method consists of those procedures which help which in drawing inferences about the characteristics of the population on the basis of samples. This method includes the theory of sampling,

different tests of significance, statistical control etc. This method is also called inferential statistics.

5. Applied Methods -
   - This type of method consists of those procedures which are applied to the problems of real life. This includes the method of statistical quality control, sample survey, linear programming and inventory control.