STATISTICS WORKSHEET-1:

1. Option (a)
2. Option (a)
3. Option (b)
4. Option (d)
5. Option (c)
6. Option (b)
7. Option (b)
8. Option (a)
9. Option (c)

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

The Normal distribution is also known as **Gaussian** or **Gauss** distribution. Many groups follow this type of pattern. Here are several reasons why the normal distribution is crucial in statistics. Some of those are as follows:

1. The **statistical hypothesis test** assumes that the data follows a normal distribution.
2. Both linear and non-linear regression assumes that the **residual** follows the normal distribution.
3. Moreover, the central limit theorem states that as the **sample size** increases the distribution of the mean follows normal distribution irrespective of the distribution of the original variable.
4. Apart from this most of the **statistical software programs** support some of the probability functions for normal distribution as well.

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

The methods to handle missing data are:

1. Mean or Median Imputation

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

- There may not be enough observations with non-missing data to produce a reliable analysis
- In predictive analytics, missing data can prevent the predictions for those observations which have missing data.
- External factors may require specific observations to be part of the analysis

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations.

2. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

Imputation techniques are:

- Complete Case Analysis(CCA)
- Arbitrary Value Imputation
- Frequent Category Imputation

12.

13. Is mean imputation of missing data acceptable practice?

No, Mean imputation reduces the variance of the imputed variables. It shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval. It does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

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. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept. The most common method for fitting a regression line is the method of least-squares.This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.After a regression line has been computed for a group of data, a point which lies far from the line (and thus has a large residual value) is known as an outlier. Such points may represent erroneous data, or may indicate a poorly fitting regression line. If a point lies far from the other data in the horizontal direction, it is known as an influential observation. The reason for this distinction is that these points have may have a significant impact on the slope of the regression line. Once a regression model has been fit to a group of data, examination of the residuals allows the modeler to investigate the validity of his or her assumption that a linear relationship exists. Plotting the residuals on the y-axis against the explanatory variable on the x-axis reveals any possible non-linear relationship among the variables, or might alert the modeler to investigate lurking variables.If non-linear trends are visible in the relationship between an explanatory and dependent variable, there may be other influential variables to consider. A ***lurking variable*** exists when the relationship between two variables is significantly affected by the presence of a third variable which has not been included in the modeling effort. Since such a variable might be a factor of time a ***time series plot*** of the data is often a useful tool in identifying the presence of lurking variables. Whenever a linear regression model is fit to a group of data, the range of the data should be carefully observed. Attempting to use a regression equation to predict values outside of this range is often inappropriate, and may yield incredible answers. This practice is known as ***extrapolation***.

15. What are the various branches of statistics?

There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

- Data Collection:

Data collection is all about how the actual data is collected.For the most part, this needn't concern us too much in terms of the mathematics (we just work with what we are given), but there are significant issues to consider when actually collecting data.

- Descriptive statistics

Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on).

- Inferential statistics.

Inferential statistics is the aspect that deals with making conclusions about the data. This is quite a wide area; essentially you are asking 'What is this data telling us, and what should we do?'