STATISTICS WORKSHEET 1

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. C
10. The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.
11. Missing data can skew anything for data scientists, from economic analysis to clinical trials. After all, any analysis is only as good as the data. A data scientist doesn't want to produce biased estimates that lead to invalid results. The concept of missing data is implied in the name: it's data that is not captured for a variable for the observation in question. Missing data reduces the statistical power of the analysis, which can distort the validity of the results.

    When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data.

    The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

    The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

12. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

13. Mean Imputation is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean of the existing observations. However, when there are many missing variables, mean imputation can results can result in a loss of variation in the data. This method does not use time-series characteristics or depend on the relationship between the variables.
14. **Linear regression** is the simplest and most extensively used statistical technique for predictive modelling analysis. It is a way to explain the relationship between a dependent variable (target) and one or more explanatory variables(predictors) using a straight line. There are two types of linear regression - **Simple** and **Multiple**.
15. The study of statistics have two major branches. They are descriptive and inferential statistics.

    a) Descriptive Statistics: It helps in summarizing and organizing any data set characteristics. It also helps in representation of data in both classification and diagrammatic way.

    b) Inferential Statistics: It helps in finding the conclusion regarding the population after analysis on the sample drawn from it.