STATISTICS-WORKSHEET 1

Q1-Q9:

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. C

Q10-Q15:

10. The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. 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.
11. The best techniques to handle missing data are:
    a. Use deletion methods to eliminate missing data:

       The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data

entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings. Alternatively, data scientists can fill out the missing values by contacting the participants in question. The problem with this method is that it may not be practical for large datasets.

b. Use regression analysis to systematically eliminate data:

Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is. Of course, the one drawback with regression analysis is that it requires significant computing power, which could be a problem if data scientists are dealing with a large dataset.

c. Use data imputation techniques:

Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. Common-point imputation, on the other hand, is when the data scientists utilize the middle point or the most commonly chosen value.

The imputation techniques recommended are:

a. Complete Case Analysis (CCA)

b. Arbitrary Value Imputation

c. Frequent Category Imputation

12. A/B testing is basically statistical hypothesis testing, or, in other words, statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics. Or in other words, A/B Testing is designed to test the new feature of a website. The traffic on the website that is users are randomly split into two groups: control (A) and experiment (B). Here, the users are either shown the original website or the modified website but not both and based on the statistical analysis the better version of the website is chosen to launch.

    The A/B testing process can be simplified as follows:

    - You start the A/B testing process by making a claim (hypothesis).
    - You launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors.
    - The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

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

    Mean imputation is typically not considered a good 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.

14. Linear regression is a basic and commonly used type of predictive analysis.

    Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

    This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

15. Statistics is concerned with developing and studying different methods for collecting, analyzing and presenting the empirical data.

    The field of statistics is composed of two broad categories- Descriptive and inferential statistics. Both of them give us different insights about the data.

    Descriptive Statistics:

    It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc.

    Data can be summarized and represented in an accurate way using charts, tables and graphs.

Inferential Statistics:

It is about using data from sample and then making inferences about the larger population from which the sample is drawn. The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, Analysis of variance etc.