Statistics

1. A
2. B
3. B
4.D
5. C
6. B
7. B
8. A
9. C
10.While Binomial and Poisson distributions enable us to deal with the occurrences of distinct events , such as the number of defective items in a sample of a given size or the number of accidents occurring in factory, the normal distribution is used for dealing with the quantities whose magnitude vary continuously.

Properties of Normal Distribution of Curve

i. The curve of the normal distribution has single peak, thus it is bell shaped with the highest point over the mean.
ii. It is symmetrical about the vertical line through mean.
iii. As it is symmetrical about mean, the mean, median and node of the distribution also have the same value.
iv. The two tails of the curve extend indefinitely and never touch the horizontal axis.
v. The curve height decreases on both the sides of the peak which occurs at the mean.
vi. The area on either side of the peak is equal.

11. Missing data is a huge problem for data analysis because it distorts findings. It's difficult to be fully confident in the insights when you know that some entries are missing values.

Methods of handling Missing values

Deletion Method:-

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. Furthermore, some corporations obtain their information from third-party sources, which only makes it unlikely that organisations can fill out the gaps manually. Pairwise deletion is the process of eliminating information when a particular data point, vital for testing, is missing. Pairwise deletion saves more data compared to likewise deletion because the former only deletes entries where variables were necessary for testing, while the latter deletes entire entries if any data is missing, regardless of its importance.

Regression Method:-

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.

12. A/B testing is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, its complexity grows.

A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

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 considered bad practice since it ignores feature correlation. Consider the following scenario: we have a table with ID, designation and Salary, and an employee with the ID "8" has a missing salary. If we average the salary of all employees , and follow mean imputation technique then the employee with the ID "8" will appear to have a significantly greater salary than he actually does as his designation is Teaboy .
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. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = m*x+c, where y = Dependent variable, c = constant, m = slope, and x = independent variable.

15. The two main branches of statistics are **descriptive statistics** and **inferential statistics**.

i. **Descriptive statistics** mostly focus on the central tendency, variability, and distribution of sample data. Central tendency means the estimate of the characteristics, a typical element of a sample or population, and includes descriptive statistics such as mean, median, and mode. Variability refers to a set of statistics that show how much difference there is among the elements of a sample or population along the characteristics measured, and includes metrics such as range, variance, and standard deviation

ii. Inferential statistics is used to draw conclusions about the characteristics of a population, drawn from the characteristics of a sample, and to decide how certain they can be of the reliability of those conclusions. Based on the sample size and distribution statisticians can calculate the probability that statistics, which measure the central tendency, variability, distribution, and relationships between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which the sample is drawn.