Statistics Answers

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. Normal Distribution

Normal distribution, also known as the Gaussian distribution, 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 graph form, normal distribution will appear as a bell curve.

11. How To handle the missing data

We can handle the missing data with the help of the Imputers. Imputers treat NaN or none or nulls with some numbers using advance techniques.

- Knn imputer : Knn imputer will try to find the relation with other columns and impute the data according the relation with other columns.
- Iterative Imputer : This method treats other columns (which does not have nulls as feature) and train on them and treat null column as label. Finally it will predict the NaN data and impute. It's just like regression problem here null column is label.

The imputation techniques which I would recommend are Knn imputer and Iterative imputer.

12. A/B Testing

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 acceptable or not ?

Mean imputation is typically considered terrible 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

Linear Regression is one of the most fundamental algorithms in the machine learning world. It is one of the supervised ML algorithms that we use to predict the future data i.e. sales, profit, marks etc. Here the label is continuous.

Regression in statistics is the process of predicting a label (Dependent variable) based on the features (independent variable) at hand.

Regression analysis is an important tool for analyzing and modeling data. Here we fit a curve/line to the data points, in such a manner that the difference between the distances of the actual data points from the plotted curve/line is minimum. Regression analyses the relationship between two or more features. When the label is continuous then we use regression.

Building blocks of Linear regression are:

- Discrete/continuous independent variables.
- A best fit Regression line.
- Continuous dependent variable.

A Linear Regression model predicts the dependent variable using a regression line based on independent variables. The equation of Linear Regression is

$$Y=a+b*x+e$$

Where ‘a’ is the intercept, ‘b’ is the slope of the line and ‘e’ is the error term & ‘x’ is the data.

15. Types of Statistics

There are two types of statistics

- Descriptive Statistics -:

The data which we can describe is called Descriptive Statistics. Descriptive statistics give information that describes the data in some manner. For example, suppose a pet shop sells cats, dogs, birds and fish. If 100 pets are sold, and 40 out of the 100 were dogs, then one description of the data on the pets sold would be that 40% were dogs.

This same pet shop may conduct a study on the number of fish sold each day for one month and determine that an average of 10 fish were sold each day. The average is an example of descriptive statistics.

- Inferential Statistics -:

Now, suppose you need to collect data on a very large population. For example, suppose you want to know the average height of all the men in a city with a population of so many million residents. It isn't very practical to try and get the height of each man.

This is where inferential statistics comes into play. Inferential statistics makes inferences about populations using data drawn from the population. Instead of using the entire population to gather the data, the statistician will collect a sample or samples from the millions of residents and make inferences about the entire population using the sample.

The sample is a set of data taken from the population to represent the population. Probability distributions, hypothesis testing, correlation testing and regression analysis all fall under the category of inferential statistics.