**Statistics 4**

1. What is central limit theorem and why is it important?

Ans: Central limit theorem of statistics states that from a population given population mean X and population standard deviation SD, if we take more and more random samples of same size with replacement then than the distribution of sample means will be equal to the population mean. If the number of samples are increasing the average sample mean will be equal to the population mean and the distribution will approximately normal distributed.

Central limit theorem helps in predicting the characteristics of the population. If we can say that the data is normally distributed a lot of questions can be answered about the data.

2. What is sampling? How many sampling methods do you know?

Ans: – Sampling means picking up random data points from a whole set of data. Sampling can be done Randomly where the probability of picking up a data point is equiprobable. Sampling with replacement, Random Sampling, Proportionate sampling in this type of sample the probability of picking up a data point from the population depends upon the value of the data point (higher the value means high probability of getting picked up).

3. What is the difference between type1 and typeII error?

Ans: – Type 1 error arises when we reject the null hypothesis which is actually True, and Type 2 error is when we fail to reject the null hypothesis when it is not true

4. What do you understand by the term Normal distribution?

Ans: Normal distribution is a probability distribution which is like a Gaussian distribution. The shape of a normal distribution is like a Bell curve and symmetric from both the side. The mean , median and mode are equal in a normal distribution

5. What is correlation and covariance in statistics?

Ans: Covariance is a measure of how much two random variables vary together where as correlation is a measure that tells how the variables are correlated. The value of covariance lies between - infi to + infi, the value of correlation lies between -1 to +1 where -1 means variable is hihly negative correlated +1 means variables are highly positive correlated.

6. Differentiate between univariate ,Biavariate,and multivariate analysis

Ans: Univariate analysis means analysis of single variable, Bi-variate analysis means analysis of 2 variable and multi-variate analysis means analysis of more than 2 variables.

7. What do you understand by sensitivity and how would you calculate it?

Ans: – The recall score or true positive score in a confusion matrix is called sensitivity score. It is calculated as the number of correctly predicted positive points divided by the total actual positive points (False negative + True positive).

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Ans : – Hypothesis testing is a statistical technique to measure / determine if there is enough evidence from the sample data set to prove a condition/assumption about the population data set is true or not. Hypothesis testing includes 2 things. Null Hypothesis (H0) and Alternate hypothesis H1).

In two tailed test Our null hypothesis is that the sample mean is equal to the population mean. A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x. (x = sample mean)

9. What is quantitative data and qualitative data?

Ans: Quantitative data refers to the data with some numerical value, it shows some weights of a specific data point for example height, Salary, Sales, Profit etc are the types of quantitative data, where as qualitative data refers to that type of data where the information is in the form of some description, quality of the data for example Gender will have Male, Female or Transgender as the values which are not in numerical format.

Most of the time categorical type variable as classified as qualitative data.

10. How to calculate range and interquartile range?

Ans: To calculate the range 1st we need to sort the data in ascending order, after that we will subtract the lowest value with the highest value, for example we have a sorted data as 1,2,3,4,5,6,7,8,9 the range will be 9-1 = 8.

Interquartile range is the difference between the 75th percentile and 25th percentile

Take the median of the data, here median will be 5, now take the median of data before 5, that will be 2.5 so the 25th quantile is 25 and 7.5 is 75th quantile . IQR = 7.5-2.5 = 5

11. What do you understand by bell curve distribution?

Ans: Bell curve refers to a probability distribution of data, usually a normal distribution is referred as a bell curve, as the distribution (PDF) of a normal distributed data is symmetric. Mean, Median and Mode in a bell-shaped curve are equal

12. Mention one method to find outliers.

Ans: – IQR method of detecting the Outliers.

To find the outliers using interquartile range steps are:

1st find the 1st and 3rd quantile of the data then subtract 1st quantile which is 25 percentiles from 3rd quantile which is 75 percentile you will get the interquartile range. Then find the threshold value which shows that this is the limit of the data points beyond this point the point will be an outlier.

Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers. If a point > 1.5 IQR + 3rd quantile it will be an outlier or if a point < 1.5 IQR – 1st quantile it will be an outlier.

13. What is p-value in hypothesis testing?

Ans: P-value shows the probability of the test statistics to be true given that our null hypothesis is true.

14. What is the Binomial Probability Formula?

Ans: – The formula can be written as nCr · pr (1 − p)n−r where n refers to the number of the trials, r refers to the total number of successful events, p refers to the probability of successful event on a single trial and (1-p) refers to the probability of failure. This can be written in the form of permutation combination: [n!/r!(n−r)!]· pr (1 − p)n−r

15. Explain ANOVA and it’s applications.

Ans: ANOVA stands for analysis of variance; it is a statistical technique used to compare the means of two or more groups of observation. ANOVA can be used in medical areas to prove if a medical treatment were equally effective or not