Q1to Q15 are descriptive types. Answer in brief.

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

Ans Given the sample size is large, the sampling distribution of the mean will be always normally distributed. We know that it is highly inefficient to measure the whole population, rather we can optimize this process with the help of CTL thus helping us to analyse data with statistical techniques even without having the full data of the population hence saving us the time.

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

Sampling is a technique where a selected numbers of data is collected from a population. There are two methods of sampling:

1. Probability Sampling
2. Non Probability sampling

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

Type I error is a case of False Positive and occurs when a null hypothesis is rejected though it is actually true.

Type II error is a case of False Negative and occurs when the null hypothesis is not rejected though it is actually false.

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

Ans A dataset is normally distributed when most of the value’s cluster around the mean and the values fall symmetrically around it to obtain a bell-shaped curve. A normal distribution with have their mean median and mode to be equal.

5. What is correlation and covariance in statistics?

Covariance and correlation are terms used for showing the relation between two variables.

Covariance describes the variance between the variables. It gives us an overall idea of the direction with respect to the mean.

Correlation gives us a concrete magnitude of how much the two variables are related to each other. Its range is between -1 and 1.

6. Differentiate between univariate, Biavariate, and multivariate analysis.

In univariate analysis we describe only one variable such as it’s count, central tendencies quartiles etc.

In bivariate we see the relationship between two variables where we can use scatter plots etc

In multivariate analysis we analyse two or more variables at a time. We can use pairplots, heatmaps etc

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

Ans Sensitivity is one of the metric used in evaluating the performance of a classification model. It is measured as the proportion of True positives to total positives

Sensitivity =

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

Hypothesis testing is a method to assess if an assumption about a parameter of the population is supported by data or not. Ho is defined as the Null Hypothesis, which proposes that there is no credibility in the assumption, where as H1 is known as Alternative hypothesis which is the assumption that is being made.

In general, for a two tailed test:

H0 = 0 and

H1 ≠ 0

9. What is quantitative data and qualitative data?

Ans Quantitative data is measurable data expressed in numbers where as qualitative data expresses the nature and are not numerically measured.

10. How to calculate range and interquartile range?

Range is calculated by finding the difference between the lowest value and the maximum value in the data.

Inter-quartile rage is the difference between the 75th percentile and 25th percentile of the data.

Therefore,

IQR = Q3 – Q1

Where, Q3= 75th percentile

Q1 = 25th percentile

11. What do you understand by bell curve distribution?

Ans If we observe a bell curve distribution, it means that the datapoints are normally distributed. Most of the values will be centred around the mean and they are symmetrical as the values fall away from the mean. The mean, median and mode of the dataset will be equal. Based on the empirical rule data points beyond 3 standard deviations, i.e which is beyond 99.7% of the data is considered unlikely and taken as outliers.

12. Mention one method to find outliers.

Ans Using Inter Quartile Range (IQR). Where anything beyond 1.5 times the IQR of upper or lower quartile is considered to be an outlier.

13. What is p-value in hypothesis testing?

Ans The p-value is a statistical measure that helps us decide whether to reject or accept a null hypothesis as it describes the probability of the result occurring if the null hypothesis was true. For example if the p- value is very small, let say 0.004, then it means that there is only 0.4% chance for the null hypothesis to be true signifying more evidence towards alternate hypothesis .

14. What is the Binomial Probability Formula?

P(X) =

Where x is the number of successes

n is the number of trials

p is the probability of the success of one trial

q is the probability of getting a failure

15. Explain ANOVA and it’s applications

Analysis of Variance (ANOVA) is a statistical test that is used for evaluating the variations found in more than two groups to see how dependent variable is impacted by the independent. variable/variables. There are two main aspects we compute during an ANOVA test which is i) Variance within a group and ii) variance between the groups. The ratio, which is the F value indicates if the means of the groups are related or not. If F-value is high, then the groups may have different means.

**F= Variance between the groups/ Variance within a group**

Depending on the result, we decide whether or not to reject the null hypothesis. ANOVA can be classified into one-way ANOVA and two-way ANOVA. One way ANOVA has only one factor effect the dependent variable where as two-way ANOVA has more than one factor.

ANOVA has a lot of applications.

Healthcare: To test and compare different medications on patients and find medicines that shows statically significant improvements against a disease for further analysis.

Manufacturing: when comparison is required for the most reliable procedures.