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

   The CLT is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.
   The CLT has several applications. Look at the places where you can use it.
   - Political/election polling is a great example of how you can use CLT. These polls are used to estimate the number of people who support a specific candidate. You may have seen these results with confidence intervals on news channels. The CLT aids in this calculation.
   - You use the CLT in various census fields to calculate various population details, such as family income, electricity consumption, individual salaries, and so on.

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

   In statistics, sampling is a method when researchers determine a representative segment of a larger population that is then used to conduct a study. Sampling generally comes in two forms — probability sampling and non-probability sampling.

- Probability Sampling Method –
  - Simple random sampling
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
- Non-probability sampling –
  - Convenience Sampling
  - Voluntary response sampling
  - Purposive Sampling
  - Snowball Sampling

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

   - Type -1 Error (Error of the first kind)

   - It is also known as a false-positive.
   - It occurs if the researcher rejects a correct null hypothesis in the population.
     - i.e., incorrect rejection of the null hypothesis.
   - Measured by alpha (significance level).
   - If the significance level is fixed at 5%,
     - It means there are about five chances of type – 1 error out of 100.
   - Cause of Type – 1 Error
     - The significance level is decided before testing the hypothesis
     - Sample size is not considered
     - This may occur due to chance
   - It can be reduced by decreasing the level of significance.

   - Type -2 Error (Error of the second kind)

   - It is also known as a false negative.
   - It occurs if a researcher fails to reject a null hypothesis that is actually a false hypothesis.
   - Measured by beta (the power of test).
   - The probability of committing a type -2 error is calculated by 1 – beta (the power of test).
   - Cause of Type – 2 Error:

    - A statistical test is not powerful enough.
    - It is caused by a smaller sample size.
        - It may hide the significance level of the items being tested.
- It can be reduced by increasing the level of significance.

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

   A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution.
   Graphically, a normal distribution is a bell curve because of its flared shape. The precise shape can vary according to the distribution of the values within the population. The population is the entire set of data points that are part of the distribution.
   Regardless of its exact shape, a normal distribution bell curve is always symmetrical about the mean. A symmetrical distribution means that a vertical dividing line drawn through the maximum/mean value will produce two mirror images on either side of the line, in which half the population is less than the mean and half is greater. However, the reverse is not always true; that is, not all symmetrical distributions are normal. In the bell curve, the peak is always in the middle, and the mean, mode and median are all the same.

5. What is correlation and covariance in statistics?

   Correlation:
   - Correlation is a statistical measure that indicates how strongly two variables are related.
   - Correlation is limited to values between the range -1 and +1
   - It is a unit free measurement.

   Covariance:
   - Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.
   - The value of covariance lies in the range of -∞ and +∞.
   - It is NOT a unit free measurement.

6. Differentiate between univariate, Bivariate and multivariate analysis.

   - Univariate analysis –

     This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

   - Bivariate analysis –

     This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables. Bivariate data analysis involves comparisons, relationships, causes and explanations. These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

   - Multivariate analysis –

     When the data involves three or more variables, it is categorized under multivariate. Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).

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

   The technique used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions is defined as sensitive analysis.

   Below are mentioned the steps used to conduct sensitivity analysis:

1. Firstly, the base case output is defined; say the NPV at a particular base case input value (V1) for which the sensitivity is to be measured. All the other inputs of the model are kept constant.
2. Then the value of the output at a new value of the input (V2) while keeping other inputs constant is calculated.
3. Find the percentage change in the output and the percentage change in the input.
4. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

   This process of testing sensitivity for another input (say cash flows growth rate) while keeping the rest of inputs constant is repeated until the sensitivity figure for each of the inputs is obtained. The conclusion would be that the higher the sensitivity figure, the more sensitive the output is to any change in that input and vice versa.

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

   The process of hypothesis testing is to draw inferences or some conclusion about the overall population or data by conducting some statistical tests on a sample.
   In hypothesis testing there are two mutually exclusive hypotheses; the Null Hypothesis (H0) and the Alternative Hypothesis (H1).
   The hypothesis actually to be tested is usually given the symbol H0, and is commonly referred to as the null hypothesis. The null hypothesis is assumed to be true unless there is strong evidence to the contrary – similar to how a person is assumed to be innocent until proven guilty.
   The other hypothesis, which is assumed to be true when the null hypothesis is false, is referred to as the alternative hypothesis
   Two-tailed hypothesis tests are also known as nondirectional and two-sided tests because you can test for effects in both directions. When you perform a two-tailed test, you split the significance level percentage between both tails of the distribution.

9. What is quantitative data and qualitative data?

   Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

10. How to calculate range and interquartile range?

    The range is calculated by subtracting the lowest value from the highest value.
    The formula for finding the interquartile range takes the third quartile value and subtracts the first quartile value. Equivalently, the interquartile range is the region between the 75th and 25th percentile (75 – 25 = 50% of the data).

11. What do you understand by bell curve distribution ?

    A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.

12. Mention one method to find outliers.

    Sorting Method –
    You can **sort** quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.
    This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

13. What is p-value in hypothesis testing?

    The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis

14. What is the Binomial Probability Formula?

    The formula for binomial distribution is:
    P(x: n,p) = nCx px (q)n-x
    Where p is the probability of success, q is the probability of failure, n= number of trials

15. Explain ANOVA and it's applications.

    Analysis of Variance (ANOVA) is a statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios use it to determine if there is any difference between the means of different groups.
    Suppose in the Manufacturing Process, we want to compare and check which are the most reliable procedures, materials, etc. We can use the ANOVA test to compare different suppliers and select the best available.