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

The central limit theorem is that the all the input data average are in the secific size or on cental part . The central limit theorem is a theorem that helps to explain the behavior of a population’s average when samples are taken from it. The central limit theorem states that the distribution of the sample averages will be approximately normal, no matter what the population distribution is. he central limit theorem is important in statistics for two reasons:

**The normality assumption**

The information that the sample distributions could approximate a normal distribution has some important applications. The normality assumption is essential for the parametric hypothesis test of the mean. However, it may look like these tests are invalid because the data is not distributed normally. If the sample size is large enough, then the central limit theorem will help make the sampling distributions that will approximate a normal distribution.

**Precision of estimates**

When we prepare graphs, we notice that when the sample size increases, the sampling distributions of the mean are grouped more strongly around the population mean. When we use a sample to estimate the mean of the whole population, this property becomes applicable. When we have a large sample size, then it is possible that the mean would be close to the real population, or we can say that we have a more precise estimate.

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

Sampling is a technique to reflect the results of the entire population by studying the results of each sample taken from the population.

In Statistics, there are different sampling techniques available to get relevant results from the population. These are categorized into two different types of sampling methods. They are:  
Probability Sampling Methods  
Non-probability Sampling methods

methods of probability sampling?

The probability sampling methods are:  
Simple Random Sampling  
Systematic Sampling  
Stratified Sampling  
Clustered Sampling

non-probability sampling methods?

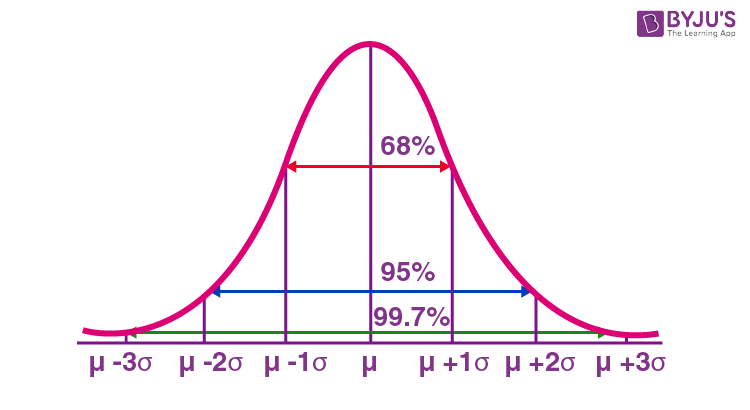
The non-probability sampling methods are:  
Convenience Sampling  
Consecutive Sampling  
Quota Sampling  
Purposive or Judgmental Sampling  
Snowball Sampling.

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

1. Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true. Type II error occurs when the sample results in the acceptance of null hypothesis, which is actually false.
2. Type I error or otherwise known as false positives, in essence, the positive result is equivalent to the refusal of the null hypothesis. In contrast, Type II error is also known as false negatives, i.e. negative result, leads to the acceptance of the null hypothesis.
3. When the null hypothesis is true but mistakenly rejected, it is type I error. As against this, when the null hypothesis is false but erroneously accepted, it is type II error.
4. Type I error tends to assert something that is not really present, i.e. it is a false hit. On the contrary, type II error fails in identifying something, that is present, i.e. it is a miss.
5. The probability of committing type I error is the sample as the level of significance. Conversely, the likelihood of committing type II error is same as the power of the test.
6. Greek letter ‘α’ indicates type I error. Unlike, type II error which is denoted by Greek letter ‘β’

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

distribution In probability theory and statistics, the **Normal Distribution**, also called the **Gaussian Distribution**, is the most significant continuous probability distribution. Sometimes it is also called a bell curve. The [random variables](https://byjus.com/maths/random-variable/) following the normal distribution are those whose values can find any unknown value in a given range.



In this image we see that the there are 3 standard deviation are available in positive as well as negative side the first deviation collect 68% of input data and 2nd deviation is collect 95% data and 3rd deviation collect the 99.7 % data and rest 0.3 % data are the outliers.

**5. What is correlation and covariance in statistics?**

Covariance:- Covariance measures how the two variables move concerning each other and is an extension of the concept of variance (which tells about how a single variable varies). It can take any value from -∞ to +∞ .The higher this value, the more dependent the relationship is. A positive number signifies positive covariance and denotes a direct connection. Effectively this means that an increase in one variable would also lead to a corresponding increase in the other variable, provided other conditions remain constant.

On the other hand, a negative number signifies negative covariance, which denotes an inverse relationship between the two variables. Though covariance is perfect for defining the type of relationship, it is not good for interpreting its magnitude.

Correlation:- Correlation is a step ahead of covariance as it quantifies the relationship between two random variables. In simple terms, it is a unit measure of how these variables change concerning each other (normalized covariance value).

The correlation has an upper and lower cap on a range, unlike covariance. It can only take values between +1 and -1. A correlation of +1 indicates that random variables have a direct and strong relationship.

On the other hand, the correlation of -1 indicates a strong inverse relationship, and an increase in one variable will lead to an equal and opposite decrease in the other variable. 0 means that the two numbers are independent.

1. **Differentiate between univariate ,Biavariate,and multivariate analysis.**

# Univariate Analysis :- Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable. There are many different ways people use univariate analysis. The most common univariate analysis is checking the central tendency (mean, median and mode), the range, the maximum and minimum values, and standard deviation of a variable.

# Bivariate Analysis :- Bivariate analysis is where you are comparing two variables to study their relationships. These variables could be dependent or independent to each other. In Bivariate analysis is that there is always a Y-value for each X-value.

# Multivariate Analysis :- Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables. For three variables, you can create a 3-D model to study the relationship (also known as Trivariate Analysis). However, since we cannot visualize anything above the third dimension, we often rely on other softwares and techniques for us to be able to grasp the relationship in the data.

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

it is also called the true positive rate, the recall, or probability of detection.It has been defined as the ability of a test to identify correctly all those who have the disease, which is “true-positive”.A 90 percent sensitivity means that 90 percent of the diseased people screened by the test will give a “true-positive” result and the remaining 10 percent a “false-negative” result.Thus, a highly sensitive test rarely overlooks an actual positive (for example, showing “nothing bad” despite something bad existing).

* The sensitivity of a diagnostic test is expressed as the probability (as a percentage) that a sample tests positive given that the patient has the disease.
* The following equation is used to calculate a test’s sensitivity:

1. **What is quantitative data and qualitative data?**

Quantitative data is, quite simply, information that can be quantified. It can be counted or measured, and given a numerical value—such as length in centimeters or revenue in dollars. Quantitative data tends to be structured in nature and is suitable for statistical analysis. If you have questions such as “How many?”, “How often?” or “How much?”, you’ll find the answers in quantitative data.

1. **How to calculate range and interquartile range?**

The **range** measures the difference between the minimum value and the maximum value in a dataset.

he **interquartile range** measures the difference between the first quartile (25th percentile) and third quartile (75th percentile) in a dataset. This represents the spread of the middle 50% of values.

## **Example: How to Calculate Range & Interquartile Range**

Suppose we have the following dataset:

**Dataset:** 1, 4, 8, 11, 13, 17, 19, 19, 20, 23, 24, 24, 25, 28, 29, 31, 32

We can use the following steps to calculate the **range**:

* Range = Maximum value – Minimum value
* Range = 32 – 1
* Range = **31**

We can use the [Interquartile Range Calculator](https://www.statology.org/interquartile-range-calculator/) to help us calculate the **interquartile range**:

* Interquartile Range = 3rd Quartile – 1st Quartile
* Interquartile Range = 26.5 – 12
* Interquartile Range = **14.5**

1. **What do you understand by bell curve distribution ?**

The term **bell curve** is used to describe the mathematical concept called normal distribution, sometimes referred to as Gaussian distribution. "Bell curve" refers to the bell shape that is created when a line is plotted using the data points for an item that meets the criteria of normal distribution. In a bell curve, the center contains the greatest number of a value and, therefore, it is the highest point on the arc of the line. This point is referred to the [mean,](https://www.thoughtco.com/the-mean-median-and-mode-2312604) but in simple terms, it is the highest number of occurrences of an element

1. **Mention one method to find outliers.**

Outliers are the values that look different from the other values in the data. Z-score method is method for detecting outliers. This method is generally used when a variable’ distribution looks close to Gaussian. Z-score is the number of standard deviations a value of a variable is away from the variable’ mean.

Z-Score = (X-mean) / Standard deviation

when the values of a variable are converted to Z-scores, then the distribution of the variable is called standard normal distribution with mean=0 and standard deviation=1. The Z-score method requires a cut-off specified by the user, to identify outliers. The widely used lower-end cut-off is -3 and the upper-end cut-off is +3. The reason behind using these cut-offs is, 99.7% of the values lie between -3 and +3 in a standard normal distribution. Let’s look at the implementation of the Z-Score method in Python.

1. **What is the Binomial Probability Formula?**

The binomial probability formula for any random variable x is given by**P (x : n, p)** = n C x p x q n-x n = the number of trials

**15. Explain ANOVA and it’s applications.**

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. We can use ANOVA to prove/disprove if all the medication treatments were equally effective or not. There are two main types of ANOVA viz. one–way ANOVA and two–way ANOVA.

* One Way ANOVA – It is also known as one factor ANOVA. Here, we are using one criterion variable (or called as a factor) and analyze the difference between more than two sample groups. Suppose in glass industry, we want to compare the variation of three batches (glass) for their average weight (factor).
* Two Way ANOVA – Here, we are using two independent variables (factors) and analyze the difference between more than two sample groups. Similarly, we want to compare the variation of three batches of glass w.r.t weight and hardness (two factors).