1. What is the Central Limit Theorem and why is it important? - 5 Marks

Ans: central limit theorem is a statistical theory which states that when the large sample size is having a finite variance, the samples will be normally distributed and the mean of samples will be approximately equal to the mean of the whole population.
In other words, the central limit theorem states that for any population with mean and standard deviation, the distribution of the sample mean for sample size N has mean μ and standard deviation σ / √n .

As the sample size gets bigger and bigger, the mean of the sample will get closer to the actual population mean. If the sample size is small, the actual distribution of the data may or may not be normal, but as the sample size gets bigger, it can be approximated by a normal distribution. This statistical theory is useful in simplifying analysis while dealing with stock index and many more.

The CLT can be applied to almost all types of probability distributions. But there are some exceptions. For example, if the population has a finite variance. Also this theorem applies to independent, identically distributed variables. It can also be used to answer the question of how big a sample you want. Remember that as the sample size grows, the standard deviation of the sample average falls because it is the population standard deviation divided by the square root of the sample size. This theorem is an important topic in statistics. In many real time applications, a certain random variable of interest is a sum of a large number of independent random variables. In these situations, we can use the CLT to justify using the normal distribution.

2. What is sampling? How many sampling methods do you know? - 5 Marks

Ans: Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights.
It is also a time-convenient and a cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in a research survey software for optimum derivation.

For example, if a drug manufacturer would like to research the adverse side effects of a drug on the country’s population, it is almost impossible to conduct a research study that involves everyone. In this case, the researcher decides a sample of people from each demographic and then researches them, giving him/her indicative feedback on the drug’s behavior.

Types of sampling

Sampling in market research is of two types – probability sampling and non-probability sampling.

1. Probability sampling: Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.
2. Non-probability sampling: In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

There are four types of probability sampling techniques:

- Simple random sampling: One of the best probability sampling techniques that helps in saving time and resources, is the Simple Random Sampling method. It is a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.

  For example, in an organization of 500 employees, if the HR team decides on conducting team building activities, it is highly likely that they would prefer picking chits out of a bowl. In this case, each of the 500 employees has an equal opportunity of being selected.
- Cluster sampling: Cluster sampling is a method where the researchers divide the entire population into sections or clusters that represent a population. Clusters are identified and included in a sample based on demographic parameters like age, sex, location, etc. This makes it very simple for a survey creator to derive effective inference from the feedback.

  For example, if the United States government wishes to evaluate the number of immigrants living in the Mainland US, they can divide it into clusters based on states such as California, Texas, Florida, Massachusetts, Colorado, Hawaii, etc. This way of conducting a survey will be more effective as the results will be organized into states and provide insightful immigration data.
- Systematic sampling: Researchers use the systematic sampling method to choose the sample members of a population at regular intervals. It requires the selection of a starting point for the sample and sample size that can be repeated at regular intervals. This type of sampling method has a predefined range, and hence this sampling technique is the least time-consuming.

For example, a researcher intends to collect a systematic sample of 500 people in a population of 5000. He/she numbers each element of the population from 1-5000 and will choose every 10th individual to be a part of the sample (Total population/ Sample Size = 5000/500 = 10).

- Stratified random sampling: Stratified random sampling is a method in which the researcher divides the population into smaller groups that don't overlap but represent the entire population. While sampling, these groups can be organized and then draw a sample from each group separately.

  For example, a researcher looking to analyze the characteristics of people belonging to different annual income divisions will create strata (groups) according to the annual family income. Eg – less than $20,000, $21,000 – $30,000, $31,000 to $40,000, $41,000 to $50,000, etc. By doing this, the researcher concludes the characteristics of people belonging to different income groups. Marketers can analyze which income groups to target and which ones to eliminate to create a roadmap that would bear fruitful results.

Four types of non-probability sampling

- Convenience sampling: This method is dependent on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street. It is usually termed as convenience sampling, because of the researcher's ease of carrying it out and getting in touch with the subjects. Researchers have nearly no authority to select the sample elements, and it's purely done based on proximity and not representativeness. This non-probability sampling method is used when there are time and cost limitations in collecting feedback. In situations where there are resource limitations such as the initial stages of research, convenience sampling is used.

  For example, startups and NGOs usually conduct convenience sampling at a mall to distribute leaflets of upcoming events or promotion of a cause – they do that by standing at the mall entrance and giving out pamphlets randomly.
- Judgmental or purposive sampling: Judgemental or purposive samples are formed by the discretion of the researcher. Researchers purely consider the purpose of the study, along with the understanding of the target audience. For instance, when researchers want to understand the thought process of people interested in studying for their master's degree. The selection criteria will be: "Are you interested in doing your masters in …?" and those who respond with a "No" are excluded from the sample.

- Snowball sampling: Snowball sampling is a sampling method that researchers apply when the subjects are difficult to trace. For example, it will be extremely challenging to survey shelterless people or illegal immigrants. In such cases, using the snowball theory, researchers can track a few categories to interview and derive results. Researchers also implement this sampling method in situations where the topic is highly sensitive and not openly discussed—for example, surveys to gather information about HIV Aids. Not many victims will readily respond to the questions. Still, researchers can contact people they might know or volunteers associated with the cause to get in touch with the victims and collect information.
- Quota sampling: In Quota sampling, the selection of members in this sampling technique happens based on a pre-set standard. In this case, as a sample is formed based on specific attributes, the created sample will have the same qualities found in the total population. It is a rapid method of collecting samples.

3. What is the difference between type I vs type II error? - 5 Marks

In statistics, type I error is defined as an error that occurs when the sample results cause the rejection of the null hypothesis, in spite of the fact that it is true. In simple terms, the error of agreeing to the alternative hypothesis, when the results can be ascribed to chance.
When on the basis of data, the null hypothesis is accepted, when it is actually false, then this kind of error is known as Type II Error. It arises when the researcher fails to deny the false null hypothesis. It is denoted by Greek letter 'beta (β)' and often known as beta error.

| BASIS FOR COMPARISON | TYPE I ERROR | TYPE II ERROR |
|---|---|---|
| Meaning | Type I error refers to non-acceptance of hypothesis which ought to be accepted. | Type II error is the acceptance of hypothesis which ought to be rejected. |
| Equivalent to | False positive | False negative |
| What is it? | It is incorrect rejection of true null hypothesis. | It is incorrect acceptance of false null hypothesis. |

| | | |
|---|---|---|
| Represents | A false hit | A miss |
| Probability of committing error | Equals the level of significance. | Equals the power of test. |
| Indicated by | Greek letter 'α' | Greek letter 'β' |

4. What is linear regression? What do the terms p-value, coefficient, and r-squared value mean? What is the significance of each of these components? - 5 Marks

Ans: Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

P-values and coefficients in regression analysis work together to tell you which relationships in your model are statistically significant and the nature of those relationships. The coefficients describe the mathematical relationship between each independent variable and the dependent variable. The p-values for the coefficients indicate whether these relationships are statistically significant.

R-squared ($R^2$) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the $R^2$ of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

5. What are the assumptions required for linear regression? -2 Marks

Ans: Regression is a parametric approach. 'Parametric' means it makes assumptions about data for the purpose of analysis. Due to its parametric side, regression is restrictive in nature. It fails to deliver good results with data sets which doesn't fulfill its assumptions. Therefore, for a successful regression analysis, it's essential to validate these assumptions.
So, how would you check (validate) if a data set follows all regression assumptions? You check it using the regression plots (explained below) along with some statistical test.

Let's look at the important assumptions in regression analysis:

A. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in $X^1$ is constant, regardless of the value of

$X^1$. An additive relationship suggests that the effect of $X^1$ on Y is independent of other variables.

B. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
C. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
D. The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.
E. The error terms must be normally distributed.

6. What do you understand by the term Normal Distribution? - 3 Marks

Ans: The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. It is also known as the Gaussian distribution and the bell curve.

7. What is correlation and covariance in statistics? - 3 Marks

Ans: Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.
The sample correlation coefficient, *r*, quantifies the strength of the relationship. Correlations are also tested for statistical significance.

covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables.

8. What is the goal of A/B Testing? - 2 Marks

Ans: A/B testing allows individuals, teams and companies to make careful changes to their user experiences while collecting data on the results. This allows them to construct hypotheses and to learn why certain elements of their experiences impact user behavior. In another way, they can be proven wrong—their opinion about the best experience for a given goal can be proven wrong through an A/B test.
More than just answering a one-off question or settling a disagreement, A/B testing can be used to continually improve a given experience or improve a single goal like conversion rate over time.

An A/B test, also known as a split test, is an experiment for determining which of different variations of an online experience performs better by presenting each version to users at random and analyzing the results. So, what is A/B testing? A/B testing demonstrates the efficacy of potential changes, enabling data-driven decisions and ensuring positive impacts.

9. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head? - 5 Marks

Ans: There are two ways of choosing the coin. One is to pick a fair coin and the other is to pick the one with two heads.
Probability of selecting fair coin = 999/1000 = 0.999
Probability of selecting unfair coin = 1/1000 = 0.001
Selecting 10 heads in a row = Selecting fair coin * Getting 10 heads + Selecting an unfair coin
P (A) = 0.999 * (1/2)^5 = 0.999 * (1/1024) = 0.000976
P (B) = 0.001 * 1 = 0.001
P( A / A + B ) = 0.000976 / (0.000976 + 0.001) = 0.4939
P( B / A + B ) = 0.001 / 0.001976 = 0.5061
Probability of selecting another head = P(A/A+B) * 0.5 + P(B/A+B) * 1 = 0.4939 * 0.5 + 0.5061 = 0.7531

10. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour? - 5 Marks

Ans: 20% probability = 20/100 = 1/5
Probability of Seeing a Star in 15 minutes = 1/5

Probability of no seeing a Star in 15 minutes = 1 - 1/5 = 4/5

probability that you see at least one shooting star in the period of an hour

= 1 - Probability of not seeing any Star in 60 minutes

= 1 - Probability of not seeing any Star in 15 * 4 minutes

= 1 - $(4/5)^4$

= 1 - 0.4096

= 0.5904

the probability that you see at least one shooting star in the period of an hour = 0.5904

11. How would you perform clustering on a million unique keywords, assuming you have 10 million data points—each one consisting of two keywords, and a metric measuring how similar these two keywords are? How would you create this 10 million data points table in the first place? - 5 Marks

Ans: First we will check whether the entire dataset has any missing value or not. Because Missing data is a huge problem for data analysis .. It's difficult to be fully confident in the insights when you know that some entries are missing values. Hence, why they must be addressed. According to data scientists, there are three types of missing data. These are Missing Completely at Random (MCAR) – when data is completely missing at random across the dataset with no discernable pattern. There is also Missing At Random (MAR) – when data is not missing randomly, but only within sub-samples of data. Finally, there is Not Missing at Random (NMAR), when there is a noticeable trend in the way data is missing.

So we will use Best techniques to handle missing data like *Use deletion methods to eliminate missing data,Use regression analysis to systematically eliminate data.*

Now that we have clean data, it's time to manipulate it in order to get the most value out of it. We have to start the data enrichment phase of the project by joining all your different sources and group logs to narrow your data down to the essential features.
we now have a nice dataset (or maybe several), so this is a good time to start exploring it by building graphs. When we're dealing with large volumes of data, visualization is the best way to explore and communicate our findings and is the next phase of our data analytics project.

By working with clustering algorithms (aka unsupervised), we can build models to uncover trends in the data that were not distinguishable in graphs and stats. These create groups of similar events (or clusters) and more or less explicitly express what feature is decisive in these results.

12. What are the differences between over-fitting and under-fitting? How to combat Overfitting and Underfitting? - 5 Marks

Ans:Underfitting A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data. *(It's just like trying to fit undersized pants!)* Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data. In such cases, the rules of the machine learning model are too easy and flexible to be applied on such minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data and also reducing the features by feature selection.
Overfitting:
A statistical model is said to be overfitted when we train it with a lot of data *(just like fitting ourselves in oversized pants!)*. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.