

Question1: What is Statistics?

Statistics is the science concerned with developing and studying methods for collecting and analyzing, interpreting, and presenting empirical data (information that comes from research).

Question2: What are the types of data?

Categorical – Describe category or groups

Example – Car Brands (Audi, BMW, TATA)

Numerical – Represent numbers

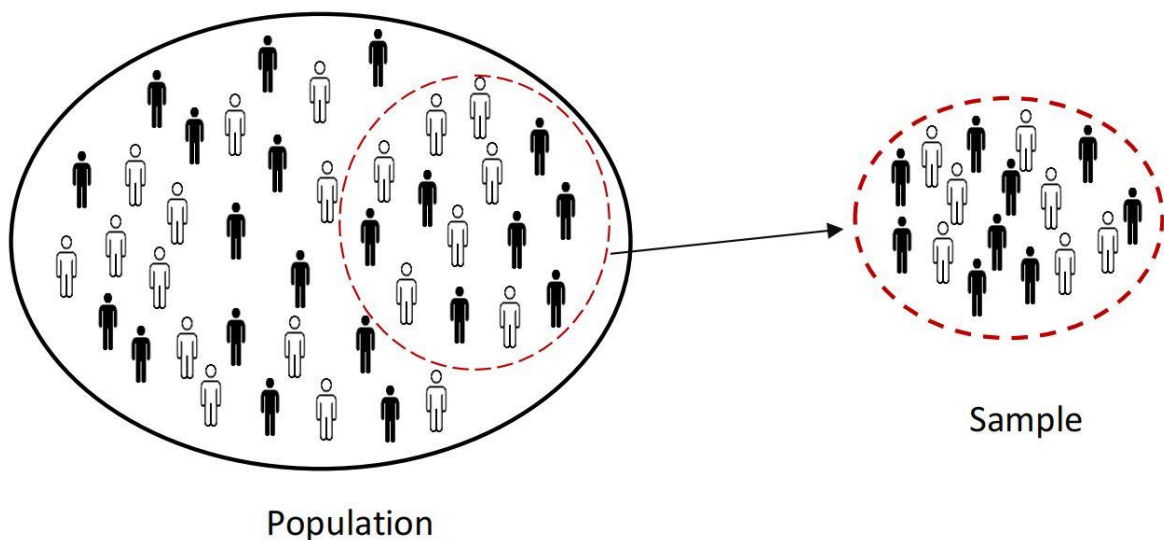
These are of two types:

- Discrete
Example – Grade, Number of Objects
- Continuous
Example – Weight, Height, Area

Question3: What is the difference between a population and a sample?

A population represents the entirety of all items that are being studied.

A sample is a finite subset of the population that is selected to represent the entire group. A sample is usually selected because the population is too large or costly to study in its entirety.

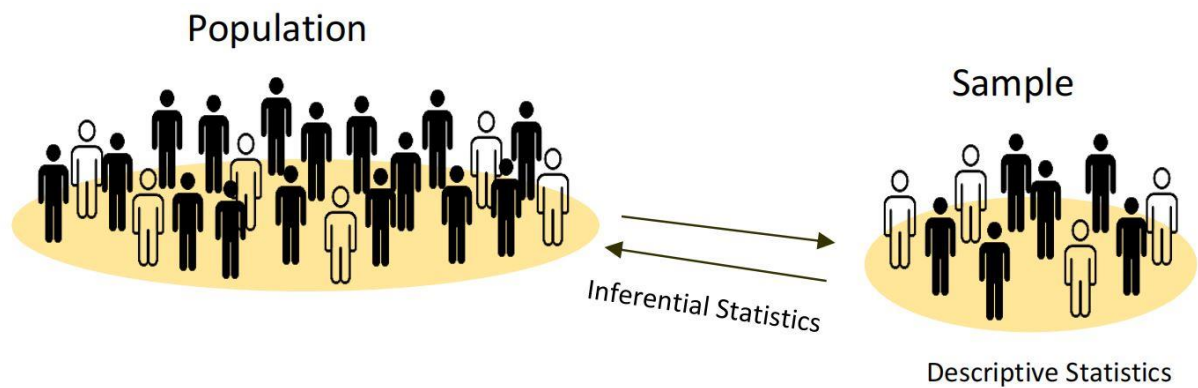
**Question4: Difference between Population and Sample?**

The Population is a collection of all items of interest while the Sample is the subset of the population. The numbers obtained from the population are called Parameters while the numbers obtained from the sample are called Statistics. Sample data are used to make conclusions on Population data.

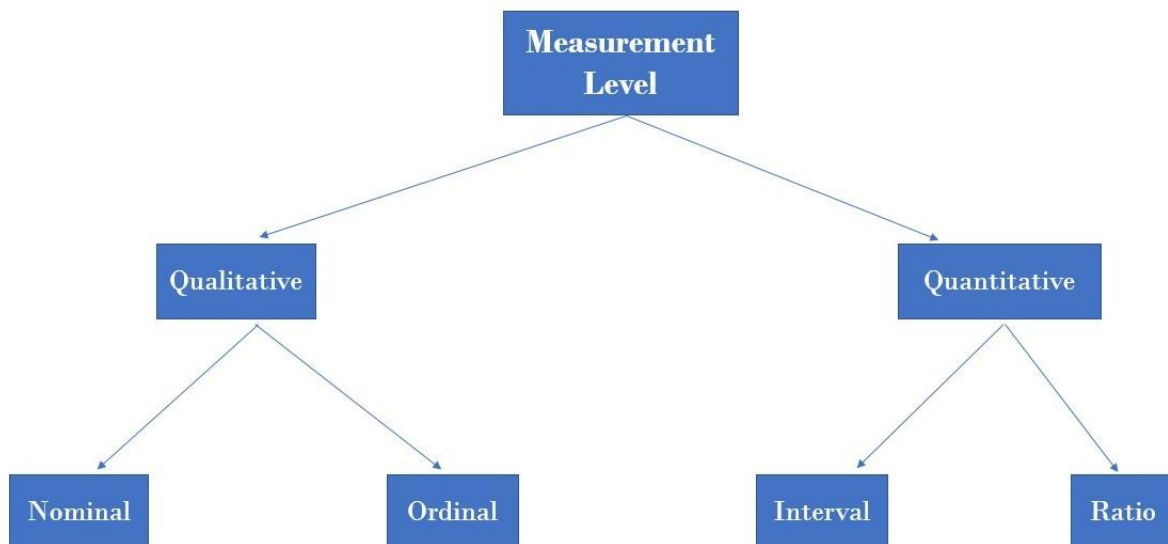
Question5: What is the difference between inferential and descriptive statistics?

Descriptive statistics describes some sample or population.

Inferential statistics attempts to infer from some sample to the larger population.



Question6: What are the different types of variables or measurement levels?



Question7: What are quantitative and qualitative data?

Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data refers to numerical data (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. Qualitative data is also known as categorical data.

Question8: What is the meaning of standard deviation?

Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. It is the average amount of variability in your dataset. It tells you, on average, how far each value lies from the mean.

A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.

The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

Question9: What are the Measures of Central Tendency?

The measure of central tendency is a single value that describes(represents) the central position within the dataset. Three most common measures of central tendency are Mean, Median, and Mode.

Mean:

Mean(Arithmetic Mean) is defined as the sum of all values divided by the number of values. If there are n values given ($x_1, x_2, x_3, \dots, x_n$) then,

$$Mean(\bar{x}) = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Median:

Median is the exact middle value when the data is ordered(i.e. arranged either in ascending or descending order). If there are n values given ($x_1, x_2, x_3, \dots, x_n$) then,

Case – I: if n is odd:

$$Median = \left(\frac{n+1}{2}\right)^{th} \text{ term}$$

Case – II: if n is even:

$$Median = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2} + 1\right)^{th}}{2}$$

i.e. mean of two middle values

Mode:

Mode is the most frequent value in the dataset. It may or may not be unique. i.e. in the dataset, more than one value can be the mode.

Question10: Which is the best measure of central tendency – Mean, Median, Mode?

- If the data is symmetrically distributed then
Mean = Median = Mode
- If the distribution is Skewed, then Median is the best measure of central tendency
- Mean is most sensitive for skewed data.
- Mode is the best measure for all levels of measurement, but more meaningful for Qualitative Data
- Variables and the corresponding best measures:

Types of Variable	Best Measurement
Nominal	• Mode
Ordinal	• Mode • Median
Interval/Ratio(Skew)	• Median
Interval/Ration(Not Skew)	• Mean

Question11: What are the Measures of Dispersion?

Dispersion or variability describes how items are distributed from each other and the centre of a distribution.

The measure of dispersion is a statistical method that helps to know how the data points are spread in the dataset.

There are 4 methods to measure the dispersion of the data:

- Range
- Interquartile Range
- Variance
- Standard Deviation

Question12: Which measure of dispersion is best?

Standard Deviation is considered the best measure of dispersion as

- Help to make a comparison between the distribution of two or more different datasets
- Based on all values
- Capable of further algebraic treatment

Question13: What is the Central Limit Theorem?

Central limit theorem states that, if you have a population mean (μ) and standard deviation (σ) and take large random samples from the population with replacement.

Then the distribution of the sample means will be approximately normally distributed regardless of whether the population is normal or skewed.

Provided that the sample size is sufficiently large ($n > 30$).

Question14: What is the difference between Covariance and Correlation?

Covariance

- Signifies the direction of the linear relationship between two variables
- In simple terms, It is a measure of variance between two variables
- It can take any value from positive infinity to negative infinity

Correlation

- It measures the relationship between two variables, as well as the strength between these two variables.
- It can take any value from -1 to 1
-

Question15: What is Normal Distribution?

Normal Distribution is a probability distribution that is symmetric about the mean. It is also known as Gaussian Distribution. The distribution appears as a Bell-shaped curve which means the mean is the most frequent data in the given data set.

In Normal Distribution:

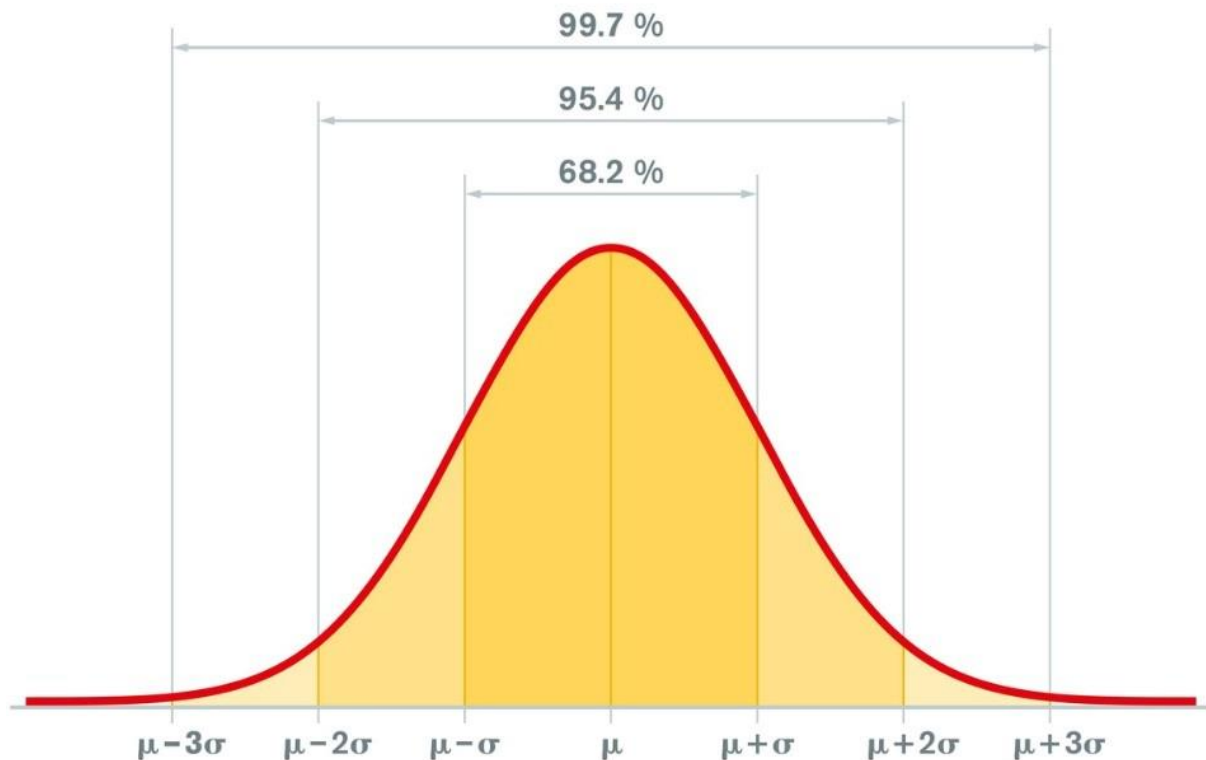
- Mean = Median = Mode
- Total area under the curve is 1.
- The probability distribution function(PDF) of a random variable x of a Normal Distribution is given by:

Question16: What is the empirical rule?

Empirical Rule is often called the **68 – 95 – 99.7** rule or **Three Sigma Rule**. It states that on a Normal Distribution:

- 68% of the data will be within one Standard Deviation of the Mean

- 95% of the data will be within two Standard Deviations of the Mean
- 99.7 of the data will be within three Standard Deviations of the Mean



Question17: It is a measure of lack of symmetry i.e. it measures the deviation of the given distribution of a random variable from a symmetric distribution (like normal Distribution).

There are two types of skewness:

- Positive Skewness
- Negative Skewness
-

Question18: What is Sampling?

It is a process of selecting a group of observations from the population, to study the characteristics of the data to make conclusions about the population.

Example: Covaxin (a covid-19 vaccine) is tested over thousand of males and females before giving it to all the people of the country.

Question19: What is an outlier in any dataset?

An outlier is a value in the data set that is extremely distinct from most of the other values.

Example:

Let there are 5 children having weights of 30 kg, 35 kg, 40kg, 50 kg and 300 kg.

Then the student's weight having 300 kg is an outlier.

An outlier in the data is due to

- Variability in the data
- Experimental Error
- Heavy skewness in data
- Missing values

Question20: What are the different methods to detect outliers in a dataset?

There are mainly 3 ways to detect outliers in a dataset:

- **Box-Plot**
- **Inter Quartile Range**
- **Z-score**

In a normal distribution, any data point whose z-score is outside the 3rd standard deviation is an outlier.

Question21: What is Hypothesis Testing?

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

There are 3 steps in Hypothesis Testing:

- State Null and Alternate Hypothesis
- Perform Statistical Test
- Accept or reject the Null Hypothesis

Question22: What is the Null and Alternate Hypothesis?

A null and alternate hypothesis is used in statistical hypothesis testing.

Null Hypothesis

- It states that the population parameter is equal to the assumed value
- It is an initial claim based on previous analysis or experience

Alternate Hypothesis

- It states that population parameters are equal or different to the assumed value
- It is what you might believe to be true or want to prove true

Question23: What are a p-value and its role in Hypothesis Testing?

P-value is the probability that a random chance generated the data or something else that is equal or rare.

P-values are used in hypothesis testing to decide whether to reject the null hypothesis or not.

- $p\text{-value} < \alpha$ – value

Means results are not in favor of the null hypothesis, reject the null hypothesis

- $p\text{-value} > \alpha$ – value

Means results are in favor of the null hypothesis, accept the null hypothesis.

Question24: What is the difference between standard deviation and variance?

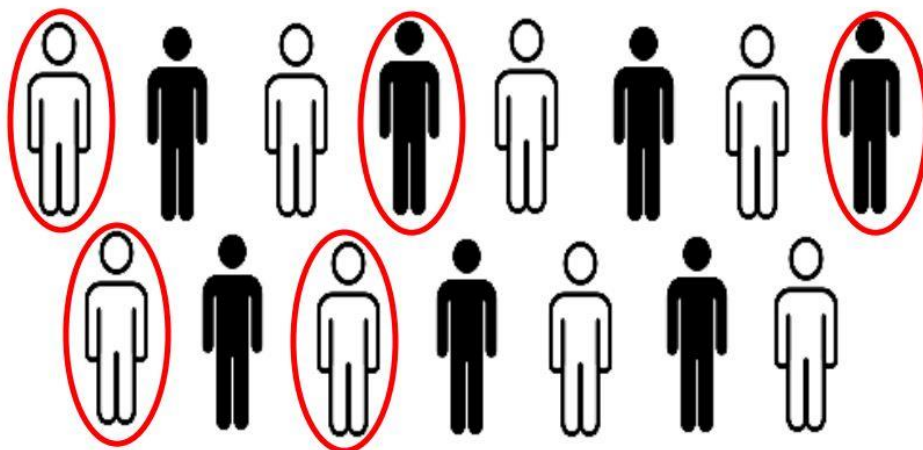
Variance and standard deviation are both measures of the spread or variability of a dataset. The variance is the average of the squared differences of each data point from the mean. The standard deviation is the square root of the variance. The main difference is that variance is measured in squared units of the original data, while standard deviation is measured in the same units as the data itself. Standard deviation is easier to interpret because it is expressed in the same units as the data.

Question25: What are the types of sampling in Statistics?

The four main types of data sampling in Statistics are:

Simple random sampling: This method involves pure random division. Each individual has the same probability of being chosen to be a part of the sample.

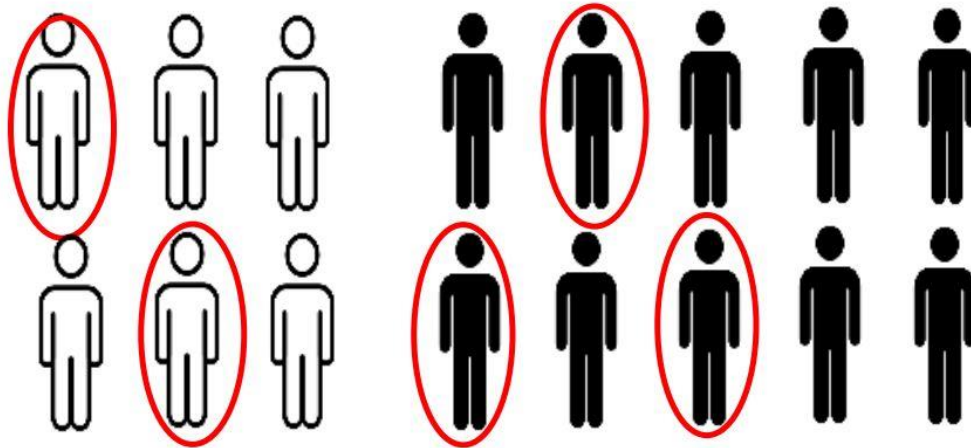
Simple random sampling



Cluster sampling: This method involves dividing the entire population into clusters. Clusters are identified and included in the sample based on demographic parameters like sex, age and location.

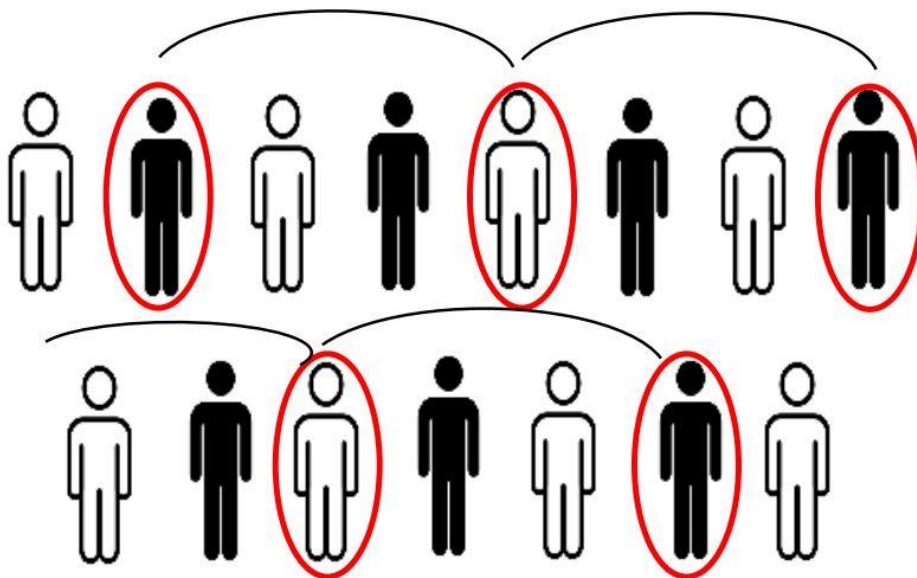
Stratified sampling: This method involves dividing the population into unique groups that represent the entire population. While sampling, these groups can be organized and then drawn a sample from each group separately.

Stratified sampling



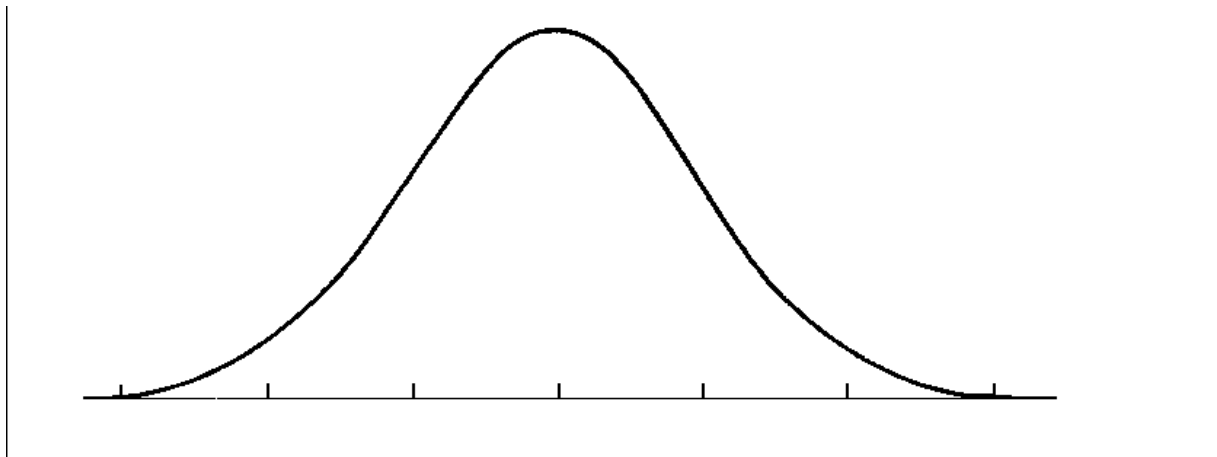
Systematic sampling: This sampling method involves choosing the sample members from a larger according to a random starting point but with a fixed, periodic interval called sampling interval. The sampling interval is calculated by dividing the population by the desired sample size. This type of sampling method has a predefined range, hence the least time-consuming.

Systematic sampling



Question26: What do you understand by the term Normal Distribution?

Normal distribution, also known as the Gaussian distribution, is a bell-shaped frequency distribution curve. Most of the data values in a normal distribution tend to cluster around the mean.



Question27: What is the assumption of normality?

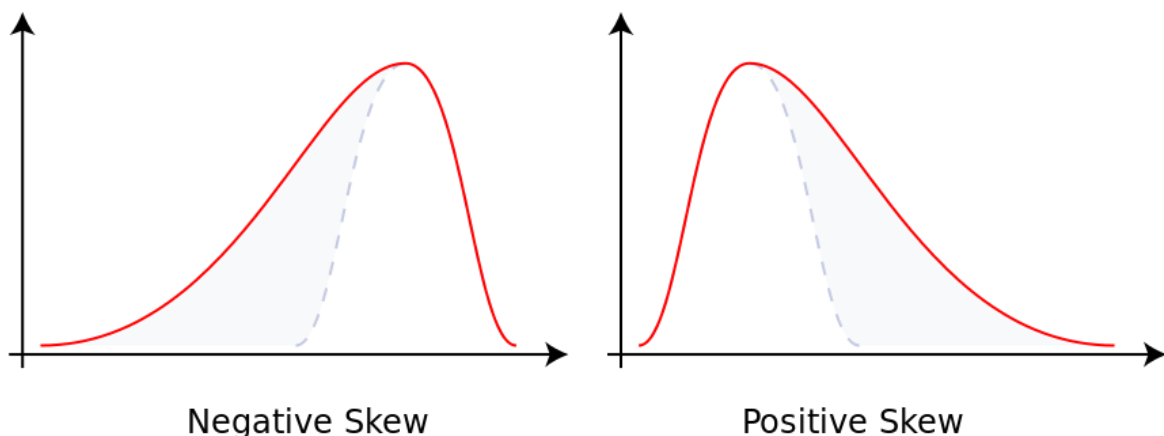
This assumption of normality dictates that if many independent random samples are collected from a population and some value of interest (like the sample mean) is calculated, and then a histogram is created to visualize the distribution of sample means, a normal distribution should be observed.

Question28: What are left-skewed distribution and right-skewed distribution?

Skewness is a way to describe the symmetry of a distribution.

A left-skewed (Negative Skew) distribution is one in which the left tail is longer than that of the right tail. For this distribution, $mean < median < mode$.

Similarly, right-skewed (Positively Skew) distribution is one in which the right tail is longer than the left one. For this distribution, $mean > median > mode$.



Question29: What are some of the properties of a normal distribution?

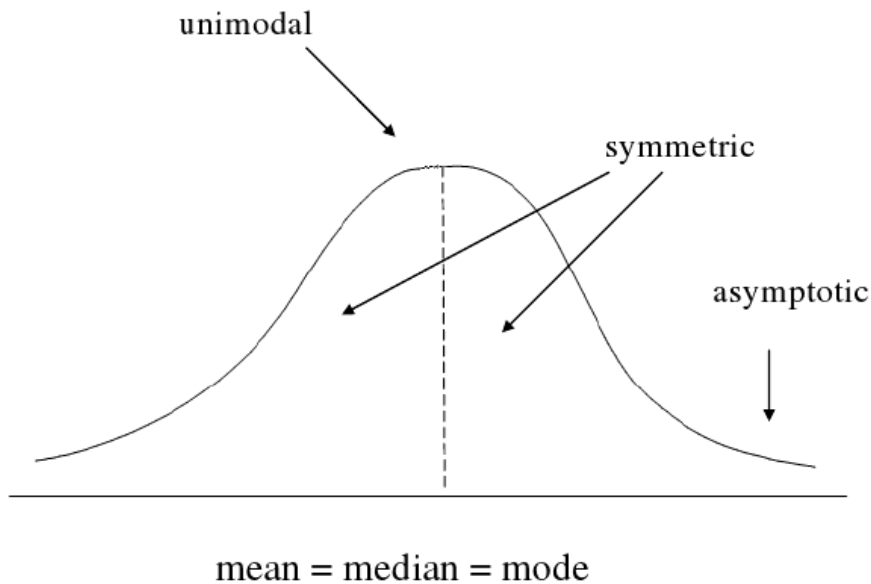
Some of the properties of a Normal Distribution are as follows:

Unimodal: normal distribution has only one peak. (i.e., one mode)

Symmetric: a normal distribution is perfectly symmetrical around its centre. (i.e., the right side of the centre is a mirror image of the left side)

The Mean, Mode, and Median are all located in the centre (i.e., are all equal)

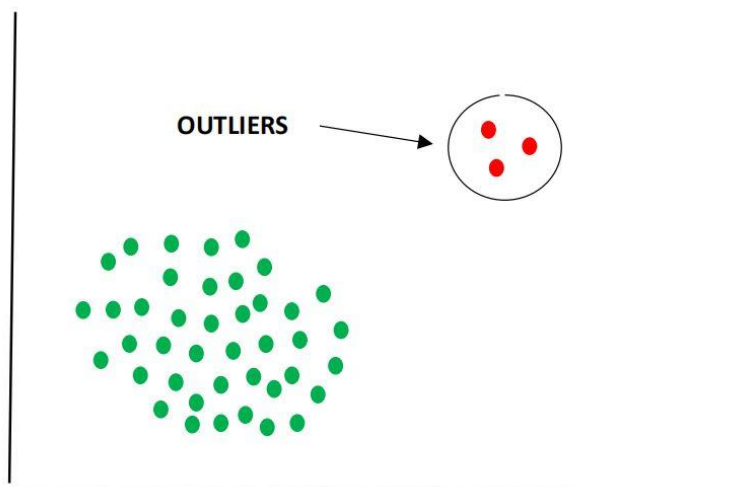
Asymptotic: normal distributions are continuous and have tails that are asymptotic. The curve approaches the x-axis, but it never touches.



Question30: What is an Outlier?

An outlier is a data point that differs significantly from other data points in a dataset. An outlier may be due to variability in measurement, or it may indicate an experimental error.

Outliers can greatly impact the statistical analyses and skew the results of any hypothesis tests.



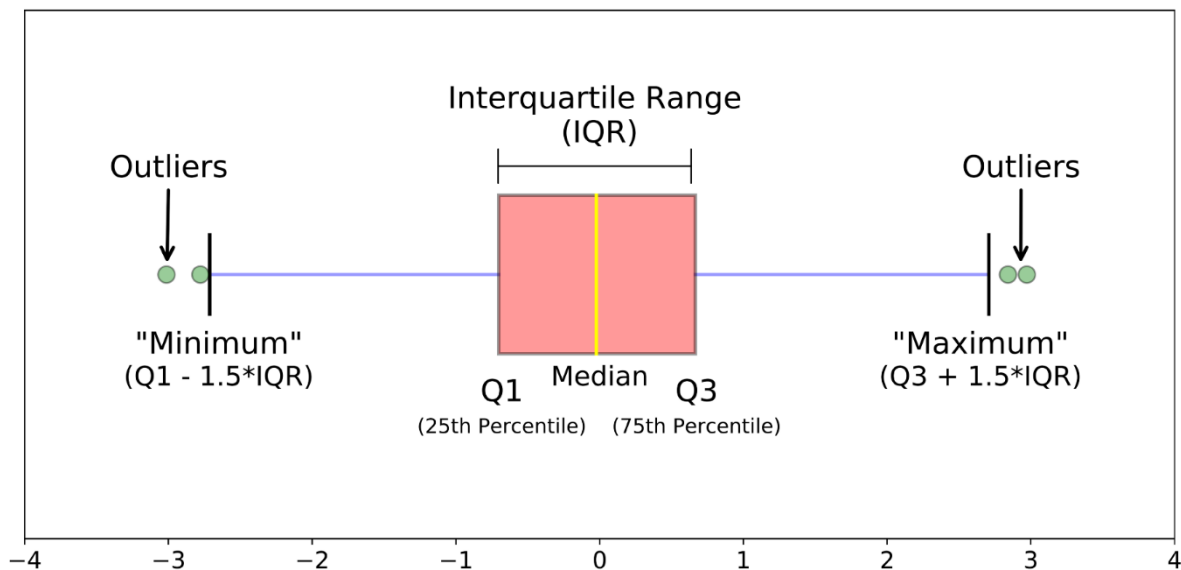
It is important to carefully identify potential outliers in the dataset and appropriately deal with them for accurate results.

Question31: Mention methods to screen for outliers in a dataset.

A simple way to check whether there is a need to investigate certain data points before using more sophisticated methods is the sorting method.

Values in the data can be sorted from low to high and then scanned for extremely low or extremely high values.

Visualization (e.g. box plot) is a useful way to see the data distribution at a glance and to detect outliers. This chart highlights statistical information like minimum and maximum values (the range), the median, and the interquartile range for the data. When reviewing a box plot, an outlier is a data point outside the box plot's whiskers.



A common method is the Interquartile range method. This method is helpful if there are few values on the extreme ends of the dataset, but you aren't sure whether any of them might count as outliers.

The interquartile range (IQR) also called midspread tells the range of the middle half of a dataset. The IQR can be used to create “fences” around the data then, the outliers can be defined as any values greater than the upper fence or less than the lower fence.

To use the IQR method:

1. Sort the data from low to high
2. Identify the first quartile (Q1), the median, and the third quartile (Q3).
3. Calculate the IQR; $IQR = Q3 - Q1$
4. Calculate the upper fence; $Q3 + (1.5 * IQR)$ and the lower fence; $Q1 - (1.5 * IQR)$
5. Use the fences to highlight any outliers (all values that fall outside your fences).

Another way to identify outliers is to use Z-score. The Z-score is just how many standard deviations away from the mean value that a certain data point is. To calculate z-score use the formula, $z = (x - \mu) / \sigma$

- If the z-score is positive, the data point is above average.

- If the z-score is negative, the data point is below average.
- If the z-score is close to zero, the data point is close to average.
- If the z-score is above or below 3 (assuming z-score = 3 is considered as a cut-off value to set the limit), it is an outlier and the data point is considered unusual.

Other methods to screen outliers include Isolation Forest and DBScan clustering.

Question32: What types of biases can you encounter while sampling?

Sampling bias occurs when a sample is not representative of a target population during an investigation or a survey. The three main that one can encounter while sampling is:

Selection bias: It involves the selection of individual or grouped data in a way that is not random.

Undercoverage bias: This type of bias occurs when some population members are inadequately represented in the sample.

Survivorship bias occurs when a sample concentrates on the 'surviving' or existing observations and ignores those that have already ceased to exist. This can lead to wrong conclusions in numerous different means.

Question33: What is the meaning of an inlier?





An **inlier** is a data value that lies within the general distribution of other observed values but is an error. Inliers are difficult to distinguish from good data values, therefore, they are sometimes difficult to find and correct.

An example of an inlier might be a value recorded in the wrong units.

Question34: What is the difference between type I vs. type II errors?

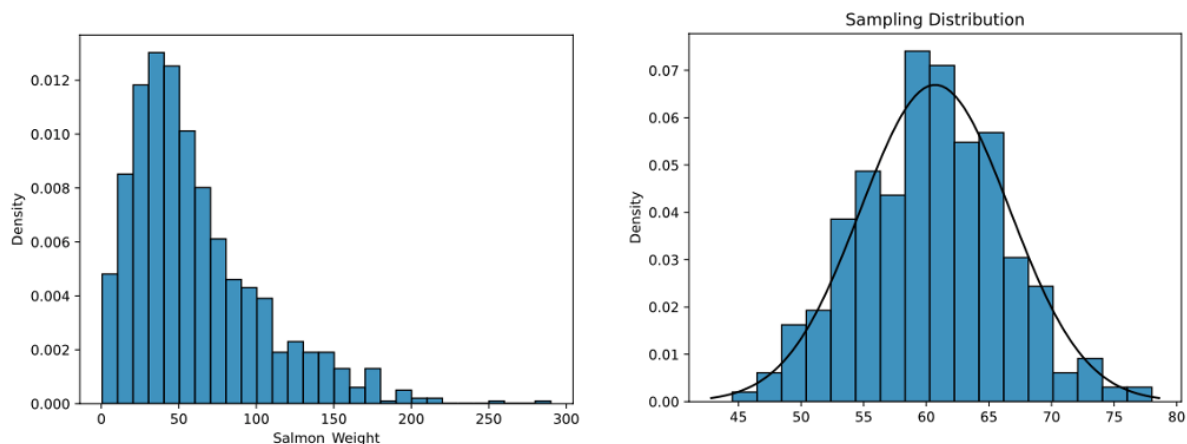
A type I error occurs when the null hypothesis true in the population is rejected. It is also known as false-positive.

A type II error occurs when the null hypothesis that is false in the population fails to get rejected. It is also known as a false-negative.

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	 Type I Error (False positive)	 Correct Outcome! (True positive)
Fail to reject null hypothesis	 Correct Outcome! (True negative)	 Type II Error (False negative)

Question35: What is the Central Limit Theorem?

The Central Limit Theorem (CLT) states that, given a sufficiently large sample size from a population with a finite level of variance, the sampling distribution of the mean will be normally distributed regardless of if the population is normally distributed.



Question36: What general conditions must be satisfied for the central limit theorem to hold?

The central limit theorem states that the sampling distribution of the mean will always follow a normal distribution under the following conditions:

The sample size is sufficiently large (i.e., the sample size is $n \geq 30$).

The samples are independent and identically distributed random variables.

The population's distribution has finite variance.

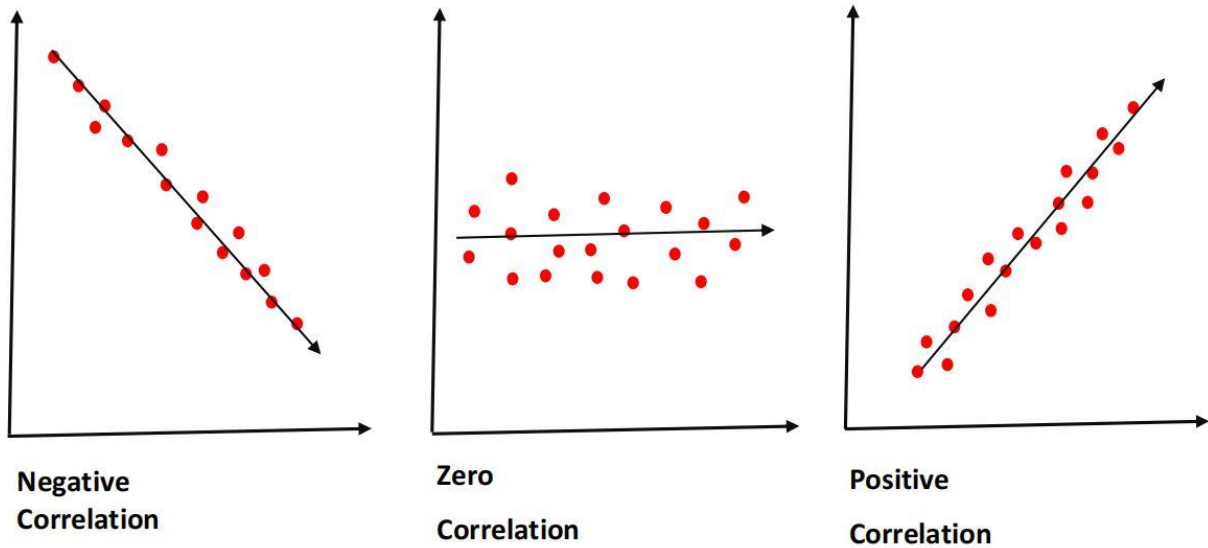
Question37: What are correlation and covariance in statistics?

Correlation indicates how strongly two variables are related. The value of correlation between two variables ranges from -1 to $+1$.

The -1 value represents a high negative correlation, i.e., if the value in one variable increases, then the value in the other variable will decrease. Similarly, $+1$ means a positive correlation, i.e., an increase in one variable leads to an increase in the other.

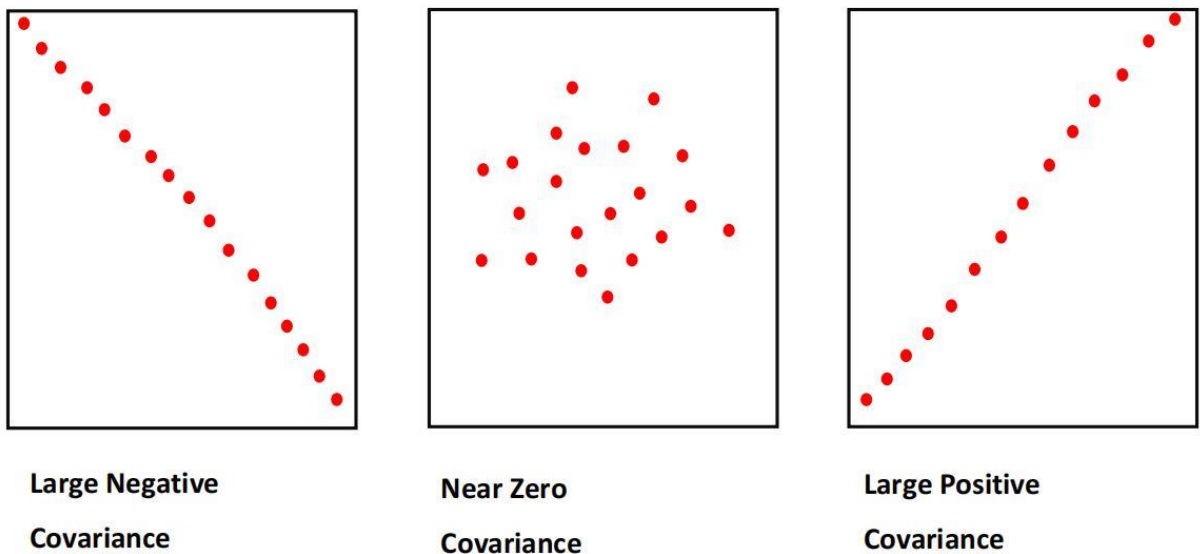
Whereas 0 means there is no correlation.

CORRELATION



Covariance, on the other hand, is a measure that indicates the extent to which a pair of random variables vary with each other. A higher number denotes a higher dependency.

COVARIANCE



Question38: How would you define Kurtosis?

Kurtosis is the extent to which the values of a distribution's tails differ from the centre of the distribution.

Outliers are detected in a data distribution using kurtosis. The higher the kurtosis, the higher the number of outliers in the data.

Question39: What is the goal of A/B testing?

A/B testing is statistical hypothesis testing. It is an analytical method for making decisions that estimate population parameters based on sample statistics.

The goal is usually to identify any changes to a web page to maximize or increase the outcome of interest. A/B testing is a fantastic method to figure out the best online promotional and marketing strategies for your business.

Question40: What is the main usage of long-tailed distributions? Where are they mainly used?

The long-tailed distributions are the type of distribution where the tail gradually drops off toward the curve's end. They are most widely used in classification and regression problems. The Pareto principle and the product sales distribution are good examples of using long-tailed distributions.

Question41: What do you understand by Hypothesis Testing?

In Statistics, Hypothesis Testing is mainly used to see if a certain experiment generates meaningful results. It helps assess the statistical significance of insight by finding the odds of the results occurring by chance. In Hypothesis Testing, the first thing is to know the null hypothesis and then specify it. After that, the p-value is calculated, and if the null hypothesis is true, the other values are also determined. The alpha value specifies the significance, and you can adjust it accordingly.

If the p-value is less than the alpha value, the null hypothesis is rejected, but the null hypothesis is accepted if the p-value is greater than the alpha value. If the null hypothesis is rejected, it indicates that the results obtained are statistically significant.

Question42: What is an exploratory data analysis in Statistics?

In Statistics, an exploratory data analysis is the process of performing investigations on data to understand the data better. In this process, the initial investigations are done to determine patterns, spot abnormalities, test hypotheses, and check if the assumptions are correct.

Question43: What do you understand by KPI in Statistics?

KPI is an acronym that stands for Key Performance Indicator. A KPI is a quantifiable measure to understand if we can achieve the goal or not. KPI is a reliable metric that is generally used to measure the performance level of an organization or individual for the objectives. An example of KPI in an organization is the expense ratio.

Question44: What do you understand by Covariance?

Covariance is a measure that specifies how much two random variables vary together. It indicates how two variables move in sync with each other. It also specifies the direction of the

relationship between two variables. There are two types of Covariance: positive and negative Covariance. The positive Covariance specifies that both variables tend to be high or low simultaneously. On the other hand, the negative Covariance specifies that the other tends to be below when one variable is high.

Question45: What is the Pareto principle used in Statistics?

The Pareto principle used in Statistics is also called the 80/20 principle or 80/20 rule. This principle specifies that 80 per cent of the results are obtained from 20 per cent of the causes in an experiment.

For example, you will have observed in your real life that 80 per cent of the wheat comes from the 20 per cent of the wheat plants on a farm.

Question46: What type of data does not have a log-normal or Gaussian distribution?

The exponential distributions types of data do not have a log-normal distribution or a Gaussian distribution. Any type of categorized data will not have these distributions as well.

For example, duration of a phone call, time until the next earthquake, etc.

Question47: What is IQR in Statistics? How can you calculate the IQR?

IQR is an acronym that stands for interquartile range. It is a measurement of the "**middle fifty**" in a data set. The IQR describes the middle 50% of values when ordered from lowest to highest.

Follow the steps given below to find the interquartile range (IQR) in Statistics:

- First, find the median (middle value) of the lower and upper half of the data.
- These values are quartile 1 (Q1) and quartile 3 (Q3).
- The IQR is the difference between Q3 and Q1.

$$\text{IQR} = \text{Q3} - \text{Q1}$$

Q3 is the third quartile (75 percentile), and Q1 is the first quartile (25 percentile).

Question48: What do you understand by the five-number summary in Statistics?

In Statistics, the five-number summary is used to measure five entities covering the entire data range. It is mainly used in descriptive analysis or during the preliminary investigation of a large data set.

The five-number summary contains the following five values:

- Low extreme (Min)
- The first quartile (Q1)
- Median
- Upper quartile (Q3)
- High extreme (Max)

Question49: What is the difference between the 1st quartile, the 2nd quartile, and the 3rd quartile?

In Statistics, quartiles are used to describe data distribution by dividing the data into three equal portions. In this partition of the data, the boundary or edge of these portions is called quartiles.

There are three types of quartile:

- **The lower quartile (Q1)** specifies the 25th percentile of the data.
- **The middle quartile (Q2):** It is also called the median and specifies the 50th percentile of the data.
- **The upper quartile (Q3)** specifies the 75th percentile of the data.
-

Question50: What do you understand by skewness?

Skewness can be described as a distortion or asymmetry that deviates from a data set's symmetrical bell curve or normal distribution. You can assume it as a degree of asymmetry observed in a probability distribution.

Depending on the varying degrees, skewness can be of two types, i.e. the right (positive) skewness and the left (negative) skewness. Skewness is centred on the mean. If skewness is negative, the data is spread more on the left of the mean than the right. If skewness is positive, the data moves more to the right. A normal distribution (bell curve) shows zero skewness.

Question51: What is the difference between a left-skewed distribution and right-skewed distribution?

The key difference between the left-skewed distribution and the right-skewed distribution is that the left tail is longer than the right side in the left-skewed distribution. Here, $\text{mean} < \text{median} < \text{mode}$. On the other hand, the right tail is longer than the right side in the right-skewed distribution. Here, $\text{mode} < \text{median} < \text{mean}$.

Question52: What is the relationship between the significance level and the confidence level in Statistics?

In Statistics, the significance level is the probability of getting a completely different result from the condition where the null hypothesis is true. On the other hand, the confidence level is used as a range of similar values in a population.

We can specify the similarity between the significance level and the confidence level by the following formula:

$\text{Significance level} = 1 - \text{Confidence level}$

Question53: What are observational and experimental data in statistics?

Observational data is derived from the observation of certain variables from observational studies. The variables are observed to determine any correlation between them.

Experimental data is derived from those experimental studies where certain variables are kept constant to determine any discrepancy or causality.

Question54: What does autocorrelation mean?

Autocorrelation is a representation of the degree of correlation between the two variables in a given time series. It means that the data is correlated in a way that future outcomes are linked to past outcomes. Autocorrelation makes a model less accurate because even errors follow a sequential pattern.

Question55: What does Design of Experiments mean?

The Design of Experiments or DOE is a systematic method that explains the relationship between the factors affecting a process and its output. It is used to infer and predict an outcome by changing the input variables.

Question56: What is the benefit of using box plots?

Boxplot is a visually effective representation of two or more data sets and facilitates quick comparison between a group of histograms.

Question57: What is a bell-curve distribution?

A bell-curve distribution is represented by the shape of a bell and indicates normal distribution. It occurs naturally in many situations especially while analyzing financial data. The top of the curve shows the mode, mean and median of the data and is perfectly symmetrical. The key characteristics of a bell-shaped curve are –

- The empirical rule says that approximately 68% of data lies within one standard deviation of the mean in either of the directions.
- Around 95% of data falls within two standard deviations and
- Around 99.7% of data fall within three standard deviations in either direction.
-

Question58: How do you assess the statistical significance of an insight?

You would perform hypothesis testing to determine statistical significance. First, you would state the null hypothesis and alternative hypothesis. Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

Question59: How do you handle missing data? What imputation techniques do you recommend?

There are several ways to handle missing data:

- Delete rows with missing data
- Mean/Median/Mode imputation
- Assigning a unique value
- Predicting the missing values
- Using an algorithm which supports missing values, like random forests

The best method is to delete rows with missing data as it ensures that no bias or variance is added or removed, and ultimately results in a robust and accurate model. However, this is only recommended if there's a lot of data to start with and the percentage of missing values is low.

Question60: How would you describe what a 'p-value' is to a non-technical person?

The best way to describe the p-value in simple terms is with an example. In practice, if the p-value is less than the alpha, say of 0.05, then we're saying that there's a probability of less than 5% that the result could have happened by chance. Similarly, a p-value of 0.05 is the same as saying "5% of the time, we would see this by chance."