1. What is standard deviation, and what is the use of it?

Standard Deviation: It measures the amount of variation or dispersion in a dataset.

Use: It quantifies how much the data points deviate from the mean. A low standard deviation indicates that the data points are close to the mean, while a high standard deviation indicates that the data points are spread out.

2. What is the difference between Normal Distribution and Skewed Distribution?

Normal Distribution: Symmetrical distribution where the mean, median, and mode are equal. The data is evenly distributed around the mean.

Skewed Distribution: Asymmetrical distribution where the tail on one side is longer or fatter than the other. It can be positively skewed (right-tailed) or negatively skewed (left-tailed).

3. What are various ways to check whether the data is normally distributed or not?

Visual Methods:

Histogram: Check for a bell-shaped curve.

Q-Q Plot: Compare the data distribution to a normal distribution.

Statistical Tests:

Shapiro-Wilk Test

Kolmogorov-Smirnov Test

Descriptive Statistics:

Check if mean ≈ median ≈ mode.

Check skewness (≈ 0) and kurtosis (≈ 3).

4. What is the difference between Normal Distribution and Standard Normal Distribution?

Normal Distribution: A general bell-shaped distribution with any mean (μ) and standard deviation (σ).

Standard Normal Distribution: A specific normal distribution with a mean of 0 and a standard deviation of 1. It is used to calculate z-scores.

5. What is a z-score? What is the use of it?

Z-Score: It measures how many standard deviations a data point is from the mean.

Use: It standardizes data, allowing comparison across different datasets. It is also used in hypothesis testing and confidence intervals.

6. What is the difference between Frequency Distribution and Probability Distribution?

Frequency Distribution: Shows how often each value occurs in a dataset (counts or percentages).

Probability Distribution: Describes the likelihood of different outcomes in a random experiment (e.g., normal distribution, binomial distribution).

7. What is the probability of a single value under the curve?

In a continuous probability distribution (e.g., normal distribution), the probability of a single value is 0. Probabilities are calculated over intervals, not single points.

8. Explain the Central Limit Theorem. What is the difference between Central Limit Theorem and Confidence Interval?

Central Limit Theorem (CLT): States that the sampling distribution of the mean of any independent, random variable will be approximately normal if the sample size is large enough (n ≥ 30).

Confidence Interval: A range of values that is likely to contain the population parameter with a certain level of confidence (e.g., 95%).

Difference: CLT explains the behavior of sample means, while confidence intervals provide a range for estimating population parameters.

9. What is the difference between Null Hypothesis and Alternate Hypothesis?

Null Hypothesis (H₀): A statement that there is no effect or no difference (default assumption).

Alternate Hypothesis (H₁): A statement that contradicts the null hypothesis, suggesting there is an effect or difference.

10. What is the difference between Type I error and Type II error?

Type I Error (False Positive): Rejecting the null hypothesis when it is true.

Type II Error (False Negative): Failing to reject the null hypothesis when it is false.

11. What is the difference between Z-test and T-test?

Z-Test: Used when the sample size is large (n ≥ 30) and the population variance is known.

T-Test: Used when the sample size is small (n < 30) and the population variance is unknown.

12. What is ANOVA Test? When do we apply it? What is the null hypothesis for it?

ANOVA (Analysis of Variance): Tests whether the means of three or more groups are equal.

When to Apply: When comparing means across multiple groups.

Null Hypothesis: All group means are equal.

13. What is the Chi-Square Test? When do we apply it? What is the null hypothesis for it?

Chi-Square Test: Tests the association between categorical variables.

When to Apply: For goodness-of-fit tests or testing independence in contingency tables.

Null Hypothesis: No association between the variables.

14. How to treat missing values in the given dataset?

Methods:

Remove rows/columns with missing values.

Impute with mean, median, or mode.

Use advanced techniques like KNN imputation or regression imputation.

15. What is an outlier? How to identify it? What are various ways to treat outliers?

Outlier: A data point significantly different from other observations.

Identification:

Visual: Boxplot, scatterplot.

Statistical: Z-score, IQR (Interquartile Range).

Treatment:

Remove outliers.

Transform data (e.g., log transformation).

Cap/floor outliers.

16. What is the difference between Nominal Encoding and Ordinal Encoding?

Nominal Encoding: Used for categorical variables with no order (e.g., one-hot encoding).

Ordinal Encoding: Used for categorical variables with a meaningful order (e.g., low, medium, high).

17. What is feature scaling? When do we apply it?

Feature Scaling: Standardizing or normalizing features to a specific range.

When to Apply: Before using algorithms sensitive to feature magnitude (e.g., KNN, SVM, gradient descent-based algorithms).

18. What is the difference between Standardization and Normalization?

Standardization: Rescales data to have a mean of 0 and a standard deviation of 1.

Normalization: Rescales data to a fixed range (e.g., 0 to 1).

19. Explain Discretization. What is the use of it?

Discretization: Converting continuous data into discrete bins or intervals.

Use: Simplifies data, reduces noise, and improves interpretability.

20. How to convert skewed distribution data to normally distributed data?

Methods:

Log transformation.

Square root transformation.

Box-Cox transformation.

Remove outliers.

1. Why do we consider n−1 to calculate the sample standard deviation?

Reason: The use of n−1 (instead of n) in the denominator is called Bessel's correction.

Purpose: It corrects the bias in the estimation of the population variance and standard deviation when using a sample.

Explanation: When calculating the sample variance,n−1 is used to account for the fact that the sample mean is used instead of the true population mean. This adjustment ensures that the sample variance is an unbiased estimator of the population variance.

2. Explain the Interquartile Range (IQR)?

Definition: The IQR is the range between the first quartile (Q1, 25th percentile) and the third quartile (Q3, 75th percentile) in a dataset.

Formula:

IQR=Q3−Q1

Use: It measures the spread of the middle 50% of the data and is used to identify outliers (data points outside 1.5×IQR below Q1 or above Q3).

3. What is the difference between Normal Distribution and Skewed Distribution?

Normal Distribution:

Symmetrical, bell-shaped curve.

Mean = Median = Mode.

Tails on both sides are equal.

Skewed Distribution:

Asymmetrical distribution.

Positively Skewed (Right-Skewed): Tail on the right side is longer; Mean > Median > Mode.

Negatively Skewed (Left-Skewed): Tail on the left side is longer; Mean < Median < Mode.

4. What is the probability of a single value under the curve? Why?

Probability of a Single Value: In a continuous probability distribution (e.g., normal distribution), the probability of a single value is 0.

Reason: Continuous distributions describe probabilities over intervals, not single points. The area under the curve for a single point is infinitesimally small, so the probability is effectively zero.

5. What is the difference between correlation and covariance in statistics?

Covariance:

Measures the direction of the linear relationship between two variables.

Can take any value between

−∞ and +∞.

Units are the product of the units of the two variables.

Correlation:

Measures both the direction and strength of the linear relationship between two variables.

Ranges between

−1 and +1.

Unitless (standardized measure).

6. Explain the Central Limit Theorem (CLT). Where do we apply it, and why?

Central Limit Theorem (CLT):

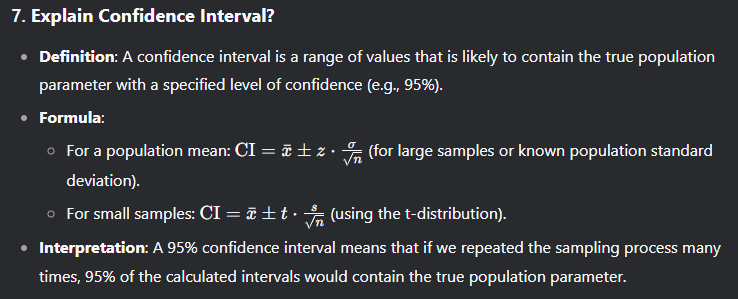
States that the sampling distribution of the mean of any independent, random variable will be approximately normal (bell-shaped) if the sample size is large enough (typically n≥30), regardless of the shape of the population distribution.

Application:

Used in hypothesis testing, confidence intervals, and statistical inference.

Why:

It allows us to make inferences about population parameters using sample statistics, even when the population distribution is not normal.



1. Explain Null Hypothesis and Alternate Hypothesis

Null Hypothesis (H₀):A statement that assumes no effect, no difference, or no relationship between variables.

Example: "There is no difference in the mean scores of two groups."

Alternate Hypothesis (H₁):A statement that contradicts the null hypothesis, suggesting there is an effect, difference, or relationship.

Example: "There is a significant difference in the mean scores of two groups."

2. What is the difference between Parametric and Non-Parametric Tests?

Parametric Tests:Assume the data follows a specific distribution (usually normal distribution).

Require assumptions about population parameters (e.g., mean, variance).

Examples: Z-test, T-test, ANOVA, Pearson correlation.

Non-Parametric Tests:Do not assume any specific distribution for the data.

Used when data does not meet the assumptions of parametric tests.

Examples: Mann-Whitney U test, Wilcoxon signed-rank test, Kruskal-Wallis test, Spearman correlation.

3. How to convert a skewed distribution to normally distributed data?

Methods:Log Transformation: Apply

log⁡(x)

log(x) to reduce right skewness.

Square Root Transformation: Apply √x for moderate skewness.

Box-Cox Transformation: A power transformation that stabilizes variance and makes the data more normal.

Remove Outliers: Outliers can cause skewness; removing them can help.

Binning: Group data into bins to reduce skewness.

4. What is the difference between Type I and Type II error?

Type I Error (False Positive):Rejecting the null hypothesis when it is true.

Significance level (α) controls the probability of Type I error.

Type II Error (False Negative):Failing to reject the null hypothesis when it is false.

Power (1−β) controls the probability of avoiding Type II error.

5. What is the difference between Z-test and T-test?

Z-Test:Used when the sample size is large (n≥30) and the population variance is known.

Assumes data is normally distributed.

T-Test:Used when the sample size is small (n<30) and the population variance is unknown.

More robust for small samples and does not strictly require normality.

6. What is ANOVA Test? When do we apply it? What is the null hypothesis for it?

ANOVA (Analysis of Variance):A statistical test used to compare the means of three or more groups.

When to Apply:When testing for significant differences between group means.

Example: Comparing the effectiveness of three different drugs.

Null Hypothesis (H₀):All group means are equal (μ1=μ2=μ3 ).

7. What is Chi-Square Test? When do we apply it? What is the null hypothesis for it?

Chi-Square Test:A statistical test used to determine if there is a significant association between categorical variables.

When to Apply:

For goodness-of-fit tests (comparing observed vs expected frequencies).

For testing independence in contingency tables.

Null Hypothesis (H₀):No association between the variables (they are independent).

**1. Meaning of Data Exploration:**

* Data exploration is the initial stage of the data analysis process.
* It involves examining and summarizing the key characteristics of your dataset to gain insights and understand its underlying patterns and trends.
* This includes tasks such as:
  + Summary statistics: Calculating measures like mean, median, standard deviation, min, max, etc.
  + Data visualization: Creating histograms, scatter plots, box plots, and other visualizations to identify patterns, outliers, and relationships between variables.
  + Univariate and bivariate analysis: Analyzing individual variables and the relationships between pairs of variables.
  + Identifying potential issues: Identifying missing values, outliers, inconsistencies, and other data quality problems.

**2. Difference between Data Cleaning & Data Exploration:**

* Data Exploration: Focuses on understanding the data, identifying patterns, and gaining insights. It's about discovering what the data tells you.
* **Data Cleaning:** Involves correcting and improving the quality of the data to ensure its accuracy and reliability for further analysis. It's about preparing the data for modeling.

**3. Steps involved in Data Cleaning:**

* **Handling Missing Values:**
  + Imputation: Replacing missing values with appropriate values (e.g., mean, median, mode, imputation methods).
  + **Removal:** Removing rows or columns with missing values (if the percentage of missing values is low).
* **Identifying and Handling Outliers:**
  + Detection: Using techniques like box plots, Z-score, IQR method.
  + Treatment: Removing outliers, transforming data (e.g., log transformation), or using robust statistical methods.
* **Handling Duplicate Data:**
  + Identification: Identifying and removing duplicate rows.
* **Correcting Inconsistent Data:**
  + Standardization: Ensuring consistent units and formats (e.g., date formats, currency).
  + Error Correction: Correcting typos, inconsistencies, and errors in data entry.
* **Data Transformation:**
  + **Scaling:** Scaling features (e.g., normalization, standardization) to improve model performance.
  + **Feature Engineering:** Creating new features from existing ones to improve model accuracy.

**4. Identifying and Rectifying Wrong Data:**

* **Identification:**
  + Visual inspection: Examining data visualizations (histograms, scatter plots) to identify unusual patterns or outliers.
  + **Domain knowledge:** Using expertise in the subject matter to identify values that are impossible or unlikely.
  + Data validation: Comparing data against known standards or external sources.
  + Cross-field validation: Checking for inconsistencies between related fields.
* **Rectification:**
  + Manual correction: Correcting errors manually based on available information.
  + Data imputation: Replacing incorrect values with estimated values.
  + **Removal:** Removing rows or columns with incorrect data if the impact is minimal.

**5. Treating Missing Values:**

* **Imputation:**
  + Mean/Median/Mode: Replacing missing values with the mean, median, or mode of the respective column.
  + K-Nearest Neighbors (KNN): Predicting missing values based on the values of nearest neighbors.
  + Regression: Predicting missing values using regression models.
  + Multiple Imputation: Creating multiple imputed datasets to account for uncertainty in the imputation process.
* **Removal:**
  + Listwise deletion: Removing rows with missing values.
  + Pairwise deletion: Excluding rows with missing values only for the specific analysis where those variables are used.

**6. Outliers:**

* **(i) Finding Outliers:**
  + Box plots: Visually identify outliers as data points outside the whiskers of the box plot.
  + **Z-score:** Calculate the Z-score for each data point. Values with a Z-score greater than 3 or less than -3 are often considered outliers.
  + Interquartile Range (IQR): Calculate the IQR and identify data points that fall below Q1 - 1.5IQR or above Q3 + 1.5IQR as outliers.
* **(ii) Extracting Outliers:**
  + Use boolean indexing to select rows with outlier values.
  + Example: outliers = df[df['column\_name'] > upper\_bound]
* **(iii) Treating Outliers:**
  + **Removal:** Remove outlier observations if they are deemed to be errors or have a significant impact on the analysis.
  + Transformation: Apply transformations like log transformation to reduce the influence of outliers.
  + Winsorization: Replace extreme values with less extreme values (e.g., replace the highest value with the 95th percentile).
  + **Robust methods:** Use statistical methods that are less sensitive to outliers (e.g., robust regression).

**1. What is meant by type casting in Python? What is the use of it?**

* **Type Casting:** Type casting (also known as type conversion) is the process of converting a variable from one data type to another.
* **Use Cases:**
  + **Data Compatibility:** Ensure data is in the correct format for operations (e.g., converting a string to an integer for mathematical calculations).
  + **Interoperability:** Allow data to be used with functions or libraries that expect specific data types.
  + **Data Manipulation:** Perform operations that require specific data types (e.g., string concatenation).

**2. What is the difference between mutable & immutable objects?**

* **Mutable Objects:** The value of a mutable object can be changed after it's created.
  + Examples: lists, dictionaries, sets.
* **Immutable Objects:** The value of an immutable object cannot be changed after it's created.
  + Examples: numbers (integers, floats), strings, tuples.

**3. What is the difference between list & tuple?**

* **List:**
  + Ordered, mutable collection of elements.
  + Enclosed in square brackets [].
  + Can be modified (elements can be added, removed, or changed).
* **Tuple:**
  + Ordered, immutable collection of elements.
  + Enclosed in parentheses ().
  + Cannot be modified after creation.

**4. What is the difference between list.append() & list.extend()?**

* **list.append()**: Adds a single element to the end of the list.
* **list.extend()**: Adds all elements of an iterable (like another list) to the end of the list.

**5. What is the difference between shallow copy & deep copy?**

* **Shallow Copy:** Creates a new object that shares the same memory location for the internal data with the original object. Changes made to the original object will also reflect in the copy.
* **Deep Copy:** Creates a new object with a completely independent copy of the data. Changes made to the original object will not affect the copy.

**6. Explain slicing concept in Python**

* Slicing allows you to extract a portion of a sequence (like a list, string, or tuple) using a specific syntax: [start:stop:step].
  + start: The index of the first element to include (inclusive, default is 0).
  + stop: The index of the element after the last element to include (exclusive, default is the end of the sequence).
  + step: The step size (default is 1).

**7. How to print the reverse of a string in one line of code?**

Python

string = "hello"

print(string[::-1])

**8. If tuple t=(1,2,3,[40,50,60],4,5,6). In this example, whether can I replace 50 with 100? If Yes, How?**

* Yes, you can replace 50 with 100 because the tuple contains a list as an element.
* Since lists are mutable, you can modify the list within the tuple:

Python

t = (1, 2, 3, [40, 50, 60], 4, 5, 6)

t[3][1] = 100

print(t) # Output: (1, 2, 3, [40, 100, 60], 4, 5, 6)

**9. What are different operators in Python?**

* **Arithmetic Operators:** +, -, \*, /, //, %, \*\* (addition, subtraction, multiplication, division, floor division, modulus, exponentiation)
* Comparison Operators: ==, !=, >, <, >=, <= (equal to, not equal to, greater than, less than, greater than or equal to, less than or equal to)
* Logical Operators: and, or, not
* Bitwise Operators: &, |, ^, ~, <<, >> (bitwise AND, OR, XOR, NOT, left shift, right shift)
* **Assignment Operators:** =, +=, -=, \*=, /= (assignment, addition assignment, subtraction assignment, etc.)
* **Identity Operators:** is, is not
* **Membership Operators:** in, not in

**10. What is the difference between comparison operators & logical operators?**

* **Comparison Operators:** Compare values and return a Boolean result (True or False).
* **Logical Operators:** Combine Boolean values to create more complex conditions.

**11. What is the difference between identity operators & membership operators?**

* **Identity Operators:** Check if two objects are the same object in memory (is, is not).
* **Membership Operators:** Check if a value is a member of a sequence (e.g., list, tuple, string) (in, not in).

**12. What is the difference between Logical OR & Logical AND?**

* **Logical OR (**or**)**: Returns True if at least one of the conditions is True.
* **Logical AND (**and**)**: Returns True only if all conditions are True.

**13. What are different conditional statements available in Python? What is the difference between only if & if with else?**

* **Conditional Statements:**
  + if
  + if-else
  + if-elif-else
* **Difference between** if **and** if-else**:**
  + if statement executes a block of code only if the condition is True.
  + if-else statement executes one block of code if the condition is True and another block of code if the condition is False.

**14. What is the difference between for loop & while loop?**

* **for loop:** Iterates over a sequence (e.g., list, tuple, string) or a range of numbers.
* **while loop:** Repeats a block of code as long as a given condition is True.

**15. What is the difference between break & continue?**

* **break:** Exits the loop entirely.
* **continue:** Skips the current iteration of the loop and moves to the next iteration.

**16. What is the difference between if condition vs while loop?**

* **if condition:** Executes a block of code once if the condition is True.
* **while loop:** Executes a block of code repeatedly as long as the condition remains True.

**17. What is meant by function? What is the difference between inbuilt & user-defined function? What is the purpose of creating a function?**

* **Function:** A reusable block of code that performs a specific task.
* **Inbuilt Functions:** Functions that are pre-defined in Python (e.g., print(), len(), max(), min()).
* **User-Defined Functions:** Functions created by the programmer to perform specific tasks within their code.
* **Purpose of Creating a Function:**
  + **Code Reusability:** Avoid writing the same code multiple times.
  + **Modularity:** Break down complex problems into smaller, more manageable units.
  + **Readability:** Improve code readability and maintainability.

**18. What is the difference between function & Method?**

* **Function:** A standalone block of code that performs a specific task.
* **Method:** A function that is associated with an object (e.g., a method of a class).

**19. What is the difference between positional arguments, default arguments & keyword arguments?**

* **Positional Arguments:** Arguments that must be provided in the same order as they are defined in the function definition.
* **Default Arguments:** Arguments that have default values assigned to them. If no value is provided for the argument during the function call, the default value is used.
* **Keyword Arguments:** Arguments that are passed to a function by explicitly specifying their names. The order of keyword arguments does not matter.

**20. What is the difference between module function & inbuilt function?**

* **Inbuilt Functions:** Pre-defined functions that are part of the Python language itself (e.g., print(), len(), max()).
* **Module Functions:** Functions that are defined within a specific module (e.g., math.sqrt(), os.path.exists()). Modules are libraries of pre-written code that provide additional functionality.

****1. What is the difference between .loc & .iloc in pandas?****

****.loc:****

* + Accesses data by ****label**** (row and column labels).
  + Can use labels directly (e.g., df.loc['row\_label'], df.loc[:, 'column\_label']).
  + Supports boolean indexing.

****.iloc:****

* + Accesses data by ****integer location**** (row and column numbers).
  + Uses integer-based indexing (e.g., df.iloc[0], df.iloc[:, 1]).

****2. What is the syntax for .iloc in pandas?****

df.iloc[row\_indexer, column\_indexer]

* + row\_indexer: Can be an integer, a list of integers, a slice object, or a boolean array.
  + column\_indexer: Can be an integer, a list of integers, a slice object, or a boolean array.

****3. How can we extract the data from the dataset based on requirements using Pandas?****

****Using .loc:****

* + Select rows by label: df.loc['row\_label']
  + Select columns by label: df.loc[:, 'column\_label']
  + Select rows and columns by label: df.loc['row\_label', 'column\_label']
  + Boolean indexing: df.loc[df['column\_name'] > value]

****Using .iloc:****

* + Select rows by integer location: df.iloc[0] (first row)
  + Select columns by integer location: df.iloc[:, 0] (first column)
  + Select rows and columns by integer location: df.iloc[0, 0] (first row, first column)
  + Slicing: df.iloc[0:3, :] (first three rows, all columns)

****Using boolean indexing:****

* + Create a boolean mask and use it to select rows: df[df['column\_name'] > value]

****4. What is the syntax for writing multiple conditions in pandas in order to extract the data from the given dataset?****

* Use the bitwise operators & (AND) and | (OR) to combine multiple conditions:

Python

df[(df['column1'] > value1) & (df['column2'] < value2)] # Select rows where column1 is greater than value1 AND column2 is less than value2

df[(df['column1'] > value1) | (df['column2'] < value2)] # Select rows where column1 is greater than value1 OR column2 is less than value2

****5. What is meant by crosstab() function in pandas. Where do we use it?****

****crosstab()**** creates a pivot table that computes two-way frequency tables of the count occurrences of the combination of two (or more) variables.

****Use Cases:****

* + Analyze the relationship between two categorical variables.
  + Explore how different categories of one variable are distributed across categories of another variable.
  + Identify potential associations or dependencies between variables.

****6. What is meant by groupby() function in pandas. Where do we use it?****

* ****groupby()**** splits the DataFrame into groups based on one or more columns.
* ****Use Cases:****
  + Calculate summary statistics (mean, sum, count, etc.) for each group.
  + Apply aggregate functions (e.g., mean(), sum(), count()) to each group.
  + Perform group-wise transformations.
  + Analyze data by different categories or groups.

****7. What is the syntax to fill the missing values in pandas?****

* df.fillna(value): Fills missing values with a specified value.
* df.fillna(method='ffill'): Fills missing values forward (using the previous value).
* df.fillna(method='bfill'): Fills missing values backward (using the next value).
* df.fillna(df.mean()): Fills missing values with the mean of the column.
* df.fillna(df.median()): Fills missing values with the median of the column.

****8. What is the difference between Population & Sample Data?****

* ****Population:**** The entire set of individuals or objects of interest.
* ****Sample:**** A subset of the population selected for analysis.

****9. What is the difference between descriptive statistics vs inferential statistics****

* ****Descriptive Statistics:**** Summarizes and describes the main features of a dataset (e.g., mean, median, standard deviation, frequency distributions).
* ****Inferential Statistics:**** Uses sample data to make inferences or draw conclusions about the population. (e.g., hypothesis testing, confidence intervals).

****10. Explain about Mean, Median, Mode.****

* ****Mean:**** The average value of a dataset.
* ****Median:**** The middle value when the data is sorted in ascending order.
* ****Mode:**** The most frequent value in the dataset.

****11. What is meant by Measures of Dispersion?****

* Measures of dispersion quantify the spread or variability of data points in a dataset.
* They provide information about how much the data points deviate from the central tendency (mean, median).

****12. What is the difference between Mean Deviation (vs) Mean Absolute Deviation (vs) Mean Square Deviation?****

* ****Mean Deviation:**** Average of the absolute differences between each data point and the mean.
* ****Mean Absolute Deviation (MAD):**** Same as Mean Deviation.
* ****Mean Square Deviation:**** Average of the squared differences between each data point and the mean.

****13. What is the difference between Variance vs Standard Deviation?****

* ****Variance:**** The average of the squared deviations from the mean.
* ****Standard Deviation:**** The square root of the variance. It provides a measure of dispersion in the same units as the original data.

****14. Why do we have to consider the n-1, in order to calculate the sample standard deviation?****

* When calculating the sample standard deviation, we use (n-1) in the denominator instead of n to provide an unbiased estimate of the population standard deviation. Using n would underestimate the population standard deviation.

****15. Explain about the Inter Quartile Range?****

* ****Interquartile Range (IQR):**** The difference between the third quartile (Q3) and the first quartile (Q1) of a dataset.
* It represents the middle 50% of the data.
* It is a measure of dispersion that is less sensitive to outliers than the standard deviation.

****1. What is meant by Database Schema & what is its importance?****

* **Database Schema:** A blueprint that defines the structure of a database.1 It specifies the tables, columns, data types, constraints, and relationships between them.2
* ****Importance:****
  + Ensures data consistency and integrity.3
  + Defines how data is organized and stored.4
  + Facilitates efficient data retrieval and manipulation.5
  + Enforces data validation rules.6

****2. What is the difference between M.S.Excel table vs SQL table?****

* ****Excel Table:****
  + Spreadsheet application for storing and manipulating data.7
  + Limited data types (text, numbers, dates, etc.).
  + Less suitable for large datasets or complex data relationships.
* ****SQL Table:****
  + Structured format for storing and managing data in a relational database.
  + Variety of data types (e.g., integer, string, date, Boolean).
  + Supports complex queries and data manipulation.8
  + Designed for efficient data retrieval and management for multiple users.9

****3. What are the different data types available in SQL?****

SQL offers various data types to represent different kinds of data:10

* ****Numeric:****
  + INTEGER: Whole numbers (e.g., 123, -45)11
  + DECIMAL/NUMERIC: Numbers with decimal precision (e.g., 3.14159, 123.45)12
  + FLOAT: Single-precision floating-point numbers (e.g., 3.14, 1.23e+5)
* ****Character:****
  + CHAR(n): Fixed-length character string (e.g., CHAR(10) for 10 characters)13
  + VARCHAR(n): Variable-length character string (up to a specified maximum)14
* ****Date & Time:****
  + DATE: Stores date information (e.g., 2023-02-02)
  + TIME: Stores time information (e.g., 15:31:00)
  + DATETIME: Stores both date and time information.15
* ****Logical:****
  + BOOLEAN: True or False values.16
* ****Others:****
  + BLOB: Binary large objects (e.g., images, audio)17
  + CLOB: Character large objects (e.g., long text)18

****4. What is the difference between CHAR and VARCHAR data types?****

* ****CHAR(n):****
  + Fixed-length character string.
  + Allocates n bytes of storage regardless of the actual data length.
  + Padding with spaces to ensure the full n characters are used.
  + ****Example:**** CHAR(10) for a 10-character name, even if some names are shorter.
* ****VARCHAR(n):****
  + Variable-length character string.19
  + Allocates only the required bytes for the actual data.
  + More storage-efficient for datasets with varying string lengths.20
  + ****Example:**** VARCHAR(50) for names, allowing them to be any length up to 50 characters.

****5. What is the difference between primary keys, Foreign keys, unique keys?****

* ****Primary Key:****
  + Uniquely identifies each row in a table.21
  + A table can only have one primary key.22
  + Enforces data integrity by ensuring no duplicate values exist for the primary key column(s).23
* ****Foreign Key:****
  + A column (or set of columns) that references the primary key of another table.
  + Creates a link between related tables.24
  + Ensures data consistency by referencing valid existing data in the referenced table.
* ****Unique Key:****
  + A column (or set of columns) that must have unique values within a table.
  + Can be multiple unique keys per table (unlike primary key).25
  + Enforces data integrity by preventing duplicate values for the specified column(s).

****6. What are constraints in SQL? Name some commonly used constraints****

* ****Constraints:**** Rules that enforce data integrity and consistency within a database.
* ****Commonly Used Constraints:****
  + ****NOT NULL:**** Ensures a column cannot have null values.
  + ****UNIQUE:**** Enforces uniqueness of values within a column (or set of columns).26
  + **PRIMARY KEY:** Combination of NOT NULL and UNIQUE, uniquely identifying each row.27
  + **FOREIGN KEY:** References the primary key of another table, establishing a relationship.28
  + ****CHECK:**** Defines a specific condition that column values must satisfy.

****7. What is the difference between WHERE and HAVING?****

* ****WHERE:**** Used in the SELECT statement to filter rows based on conditions applied to columns within the selected table itself.
  + Filters data before any aggregation occurs.29
* ****HAVING:**** Used in the GROUP BY clause to filter groups based on conditions applied to aggregate functions (e.g., COUNT(), SUM(), AVG()).
  + Filters data after aggregation is performed.30
* **Left Join:** Returns all rows from the left table and matching rows from the right table.33 If there's no match in the right table, null values are returned for34 the right table's columns.

8. Various Types of Joins:

- \*\*INNER JOIN\*\*: Returns only matching records from both tables.

- \*\*LEFT JOIN\*\* (or LEFT OUTER JOIN): Returns all records from the left table and matching records from the right table; fills NULLs if no match is found.

- \*\*RIGHT JOIN\*\* (or RIGHT OUTER JOIN): Returns all records from the right table and matching records from the left table; fills NULLs if no match is found.

- \*\*FULL JOIN\*\* (or FULL OUTER JOIN): Returns all records when there is a match in either table, filling NULLs where necessary.

- \*\*CROSS JOIN\*\*: Returns the Cartesian product of both tables (every row from the first table joins with every row from the second table).

- \*\*SELF JOIN\*\*: A join where a table is joined with itself.

9. SQL query to find all missing values in a particular column:

SELECT \* FROM table\_name WHERE column\_name IS NULL;

10. SQL query to find all employees with salary above 500000:

SELECT \* FROM Employees WHERE salary > 500000;

11. SQL query to retrieve all records from a table named Employees:

SELECT \* FROM Employees;

12. SQL query to retrieve unique departments from the Employees table:

SELECT DISTINCT department FROM Employees;

13. SQL query to count the number of employees in each department:

SELECT department, COUNT(\*) AS employee\_count FROM Employees GROUP BY department;

14. SQL query to find the employee with the highest salary:

SELECT \* FROM Employees ORDER BY salary DESC LIMIT 1;

15. SQL query to find the average salary of employees in each department:

SELECT department, AVG(salary) AS avg\_salary FROM Employees GROUP BY department;

16. SQL query to find all employees who joined after the year 2020:

SELECT \* FROM Employees WHERE YEAR(joining\_date) > 2020;

17. SQL query to find the second highest salary from the Employees table:

SELECT DISTINCT salary FROM Employees ORDER BY salary DESC LIMIT 1 OFFSET 1;

18. SQL query to find all employees whose name starts with 'A':

SELECT \* FROM Employees WHERE name LIKE 'A%';

19. SQL query to update the salary of an employee whose ID is 101:

UPDATE Employees SET salary = new\_salary WHERE employee\_id = 101;

20. SQL query to delete a record from the Employees table:

DELETE FROM Employees WHERE employee\_id = some\_id;

21. SQL query to fetch the list of employees who are also managers:

SELECT \* FROM Employees WHERE role = 'Manager';

22. SQL query to fetch common records between two tables:

SELECT \* FROM table1 INNER JOIN table2 ON table1.common\_column = table2.common\_column;

23. SQL query to select all employees who have not been assigned to any department:

SELECT \* FROM Employees WHERE department IS NULL;

24. SQL query to fetch the names of employees who have a salary greater than the average salary:

SELECT name FROM Employees WHERE salary > (SELECT AVG(salary) FROM Employees);

25. SQL query to fetch the details of employees whose salary is within a certain range:

SELECT \* FROM Employees WHERE salary BETWEEN lower\_limit AND upper\_limit;