Python, Probability and Statistics

Interview Questions and Answers

Python is the most sought-after programming language skill. In this Python Interview Preparation document, the most frequently asked Python interview questions have been compiled.

Let us first begin with some Basic Python Interview Questions.

Q1. What is the difference between list and tuples in Python?

Answer: The differences between list and tuples are as follows:

List	Tuples
Lists are mutable i.e they can be edited.	Tuples are immutable (tuples are lists which can't be edited).
Lists are slower than tuples.	Tuples are faster than list.
Syntax: list_1 = [10, 'Chelsea', 20]	Syntax: tup_1 = (10, 'Chelsea' , 20)

Q2. What are the key features of Python?

Answer: The key features of Python can be listed as:

- Python is an **interpreted** language. That means that, unlike languages like C and its variants, Python does not need to be compiled before it is run. Other interpreted languages include PHP and Ruby.
- Python is **dynamically typed**, this means that you don't need to state the types of variables when you declare them or anything like that. You can do things like x=111 and then x="I'm a string" without error

- Python is well suited to **object orientated programming** in that it allows the definition of classes along with composition and inheritance. Python does not have access specifiers (like C++'s public, private).
- In Python, **functions** are **first-class objects**. This means that they can be assigned to variables, returned from other functions and passed into functions. Classes are also first-class objects
- **Writing Python code is quick** but running it is often slower than compiled languages. Fortunately, Python allows the inclusion of C-based extensions so bottlenecks can be optimized away and often are.The **numpy** package is a good example of this, it's really quite quick because a lot of the number-crunching it does isn't actually done by Python
- Python finds **use in many spheres** – web applications, automation, scientific modeling, big data applications and many more. It's also often used as "glue" code to get other languages and components to play nice.

Q3. What type of language is python? Programming or scripting?

Answer: Python is capable of scripting, but in general sense, it is considered asa general-purpose programming language.

Q4. Python an interpreted language. Explain.

Answer: An interpreted language is any programming language which is not in machine-level code before runtime. Therefore, Python is an interpreted language.

Q5. What is pep 8?

Answer: PEP stands for **Python Enhancement Proposal.** It is a set of rules that specify how to format Python code for maximum readability.

Q6. What are the benefits of using Python?

Answer: The benefits of using python are:

- **Easy to use**– Python is a high-level programming language that is easy touse, read, write and learn.
- **Interpreted language**– Since python is interpreted language, it executesthe code line by line and stops if an error occurs in any line.
- **Dynamically typed**– the developer does not assign data types to variables at the time of coding. It automatically gets assigned duringexecution.
- **Free and open-source**– Python is free to use and distribute. It is opensource.
- **Extensive support for libraries**– Python has vast libraries that containalmost any function needed. It also further provides the facility to import other packages using Python Package Manager(pip).
- **Portable**– Python programs can run on any platform without requiringany change.
- The data structures used in python are user friendly.
- It provides more functionality with less coding.

Q7. What are Python namespaces?

Answer: A namespace in python refers to the name which is assigned to each object in python. The objects are variables and functions. As each object is created, its name along with space (the address of the outer function in which the object is), gets created. The namespaces are maintained in python like a dictionary where the key is the namespace and value is the address of the object. There are 4 types of namespace in python-

- **Built-in namespace**– These namespaces contain all the built-in objects inpython and are available whenever python is running.
- **Global namespace**– These are namespaces for all the objects created atthe level of the main program.
- **Enclosing namespaces**– These namespaces are at the higher level or

outer function.

- **Local namespaces**– These namespaces are at the local or inner function.

Q8. What are decorators in Python?

Answer: Decorators are used to add some design patterns to a function without changing its structure. Decorators generally are defined before the function they are enhancing. To apply a decorator, we first define the decorator function. Then we write the function it is applied to and simply add the decorator function above the function it has to be applied to. For this, we use the @ symbol before the decorator.

Q9. What are Dict and List comprehensions?

Answer: Dictionary and list comprehensions are just another concise way to define dictionaries and lists.

Example of list comprehension is-

```
x= [i for i in range(5)]
```

The above code creates a list as below-[0,1,2,3,4]

Example of dictionary comprehension is-

```
x= [i : i+2 for i in range(5)]
```

The above code creates a list as below-

[0: 2, 1: 3, 2: 4, 3: 5, 4: 6]

Q10.What are the common built-in data types in Python?

Answer: The common built-in data types in python are-

- **Numbers**– They include integers, floating-point numbers, and complex numbers.

- Example: 1, 7.9,3+4i
- **List**– An ordered sequence of items is called a list. The elements of a list may belong to different data types.
- Example: [5,'market',2.4]
- **Tuple**– It is also an ordered sequence of elements. Unlike lists , tuples are immutable, which means they can't be changed.
- Example: (3,'tool',1)
- **String**– A sequence of characters is called a string. They are declared within single or double quotes.
- Example: "Sana", 'She is going to the market', etc.
- **Set**– Sets are a collection of unique items that are not in order. Example: {7,6,8}
- **Dictionary**– A dictionary stores values in key and value pairs where each value can be accessed through its key. The order of items is not important.
- Example: {1:'apple',2:'mango}
- **Boolean**– There are 2 boolean values- **True** and **False**.

Q11.What is the difference between .py and .pyc files?

Answer: The .py files are the python source code files. While the .pyc files contain the bytecode of the python files. .pyc files are created when the codeis imported from some other source. The interpreter converts the source .py files to .pyc files which helps by saving time.

Q12.What is slicing in Python?

Answer: Slicing is used to access parts of sequences like lists, tuples, and strings. The syntax of slicing is-[start:end:step]. The step can be omitted as well.When we write [start:end] this returns all the elements of the sequence from the start (inclusive) till the end-1 element. If the start or end element is negative i, it means the ith element from the end. The step indicates the jump or how many elements have to be skipped. Eg. if there is a list- [1,2,3,4,5,6,7,8].Then [-1:2:2] will return elements starting from the last element till the third element by printing every second element. i.e. [8,6,4].

Q13.What are Keywords in Python?

Answer: Keywords in python are reserved words that have special meaning. They are generally used to define type of variables. Keywords cannot be used for variable or function names. There are following 33 keywords in python-

- And
- Or
- Not
- If
- Elif
- Else
- For
- While
- Break
- As
- Def
- Lambda
- Pass
- Return
- True
- False
- Try
- With
- Assert

- Class
- Continue
- Del
- Except
- Finally
- From
- Global
- Import
- In
- Is
- None
- Nonlocal
- Raise
- Yield

Q14. What are Literals in Python and explain about different Literals.

Answer: A literal in python source code represents a fixed value for primitive data types. There are 5 types of literals in python-

1. **String literals**– A string literal is created by assigning some text enclosed in single or double quotes to a variable. To create multiline literals, assign the multiline text enclosed in triple quotes. Eg.name=”Tanya”

2. **A character literal**– It is created by assigning a single character enclosed in double quotes. Eg. a=’t’

3. **Numeric literals** include numeric values that can be either integer, floating point value, or a complex number. Eg. a=50

4. **Boolean literals**– These can be 2 values- either True or False.

5. **Literal Collections**– These are of 4 types-

   a) List collections-Eg. a=[1,2,3,'Amit']

   b) Tuple literals- Eg. a=(5,6,7,8)

   c) Dictionary literals- Eg. dict={1: 'apple', 2: 'mango, 3: 'banana`'}

   d) Set literals- Eg. {"Tanya", "Rohit", "Mohan"}

6. **Special literal**- Python has 1 special literal None which is used to return a null variable.

Q15.How to combine dataframes in pandas?

Answer: The dataframes in python can be combined in the following ways-

1. Concatenating them by stacking the 2 dataframes vertically.
2. Concatenating them by stacking the 2 dataframes horizontally.
3. Combining them on a common column. This is referred to as joining.

The concat() function is used to concatenate two dataframes. Its syntax is- pd.concat([dataframe1, dataframe2]).

Dataframes are joined together on a common column called a key. When we combine all the rows in dataframe it is union and the join used is outer join. While, when we combine the common rows or intersection, the join used is the inner join. Its syntax is- pd.concat([dataframe1, dataframe2], axis='axis', join='type_of_join')

Q16.What are the new features added in Python 3.9.0.0 version?

Answer: The new features in Python 3.9.0.0 version are-

a. New Dictionary functions Merge(|) and Update(|=)
b. New String Methods to Remove Prefixes and Suffixes
c. Type Hinting Generics in Standard Collections

d. New Parser based on PEG rather than LL1
e. New modules like zoneinfo and graphlib
f. Improved Modules like ast, asyncio, etc.
g. **Optimizations such as optimized idiom for assignment, signal handling, optimized python built ins, etc.**
h. Deprecated functions and commands such as deprecated parser andsymbol modules, deprecated functions, etc.
i. Removal of erroneous methods, functions, etc.

Q17. How is memory managed in Python?

Answer: Memory is managed in Python in the following ways:

1. Memory management in python is managed by **Python private heap space**. All Python objects and data structures are located in a private heap. The programmer does not have access to this private heap. The python interpreter takes care of this instead.
2. The allocation of heap space for Python objects is done by Python's memory manager. The core API gives access to some tools for the programmer to code.
3. Python also has an inbuilt garbage collector, which recycles all the unused memory and so that it can be made available to the heap space.

Q18. What is namespace in Python?

Answer: A namespace is a naming system used to make sure that names are unique to avoid naming conflicts.

Q19. What is PYTHONPATH?

Answer: It is an environment variable which is used when a module isimported. Whenever a module is imported, PYTHONPATH is also looked up to

check for the presence of the imported modules in various directories. The interpreter uses it to determine which module to load.

Q20. What are python modules? Name some commonly used built-in modules in Python?

Answer: Python modules are files containing Python code. This code can eitherbe functions classes or variables. A Python module is a .py file containing executable code.

Some of the commonly used built-in modules are:

- os
- sys
- math
- random
- data time
- JSON

Q21.What are local variables and global variables in Python?

Answer: Global Variables:

Variables declared outside a function or in global space are called globalvariables. These variables can be accessed by any function in the program.

Local Variables:

Any variable declared inside a function is known as a local variable. This variable is present in the local space and not in the global space.

Example:

```
a=2

def add():b=3
```

```
c=a+b

print(c)

add()
```

Output: 5

When you try to access the local variable outside the function add(), it will throw an error.

Q22. Is python case sensitive?

Answer: Yes. Python is a case sensitive language.

Q23.What is type conversion in Python?

Answer: Type conversion refers to the conversion of one data type into another.

int() – converts any data type into integer type

float() – converts any data type into float type

ord() – converts characters into integer

hex() – converts integers to hexadecimal

oct() – converts integer to octal

tuple() – This function is used to convert to a tuple.

set() – This function returns the type after converting to set.

list() – This function is used to convert any data type to a list type.

dict() – This function is used to convert a tuple of order (key, value) into a dictionary.

str() – Used to convert integer into a string.

complex(real,imag) – This function converts real numbers to complex(real,imag) number.

Q24. How to install Python on Windows and set path variable?

Answer: To install Python on Windows, follow the below steps:

- Install python from this link: https://www.python.org/downloads/
- After this, install it on your PC. Look for the location where PYTHON hasbeen installed on your PC using the following command on your command prompt: cmd python.
- Then go to advanced system settings and add a new variable and nameit as PYTHON_NAME and paste the copied path.
- Look for the path variable, select its value and select 'edit'.
- Add a semicolon towards the end of the value if it's not present andthen type %PYTHON_HOME%

Q25. Is indentation required in python?

Answer: Indentation is necessary for Python. It specifies a block of code. All code within loops, classes, functions, etc is specified within an indented block. It is usually done using four space characters. If your code is not indented necessarily, it will not execute accurately and will throw errors as well.

Q26. What is the difference between Python Arrays and lists?

Answer: Arrays and lists, in Python, have the same way of storing data. But, arrays can hold only a single data type elements whereas lists can hold anydata type elements.

Example:

```
import array as arr My_Array=arr.array('i',[1,2,3,4])
```

```
My_list=[1,'abc',1.20]

print(My_Array) print(My_list)
```

Output:

array('i', [1, 2, 3, 4]) [1, 'abc', 1.2]

Q27. What are functions in Python?

Answer: A function is a block of code which is executed only when it is called.To define a Python function, the **def** keyword is used.

Example:

```
def Newfunc():

print("Hi, Welcome to Edureka") Newfunc();

#calling the function
```

Output: Hi, Welcome to Edureka

Q28.What is_____init__?

Answer: init is a method or constructor in Python. This method is automatically called to allocate memory when a new object/ instance of a classis created. All classes have the_init_method.

Here is an example of how to use it.

```
class Employee:

def_____init_____(self, name, age,salary):self.name =

name

self.age = age
```

```
self.salary = 20000

E1 = Employee("XYZ", 23, 20000)

# E1 is the instance of class Employee.#__init__allocates

memory for E1. print(E1.name)

print(E1.age)

print(E1.salary
```

Output:

XYZ

23

20000

Q29.What is a lambda function?

Answer: An anonymous function is known as a lambda function. This functioncan have any number of parameters but can have just one statement.

Example:

```
a = lambda x,y : x+yprint(a(5,

6))
```

Output: 11

Q30. What is self in Python?

Answer: Self is an instance or an object of a class. In Python, this is explicitly included as the first parameter. However, this is not the case in Java where it's

optional. It helps to differentiate between the methods and attributes of aclass with local variables.

The self variable in the init method refers to the newly created object while in other methods, it refers to the object whose method was called.

Q31. How does break, continue and pass work?

Answer:

Break	Allows loop termination when some condition is met and the control is transferred to the next statement.
Continue	Allows skipping some part of a loop when some specific condition is met and the control is transferred to the beginning of the loop
Pass	Used when you need some block of code syntactically, but you want to skip its execution. This is basically a null operation. Nothing happens when this is executed.

Q32. What does [::-1} do?

Answer: [::-1] is used to reverse the order of an array or a sequence.

For example:

```
import array as arr

My_Array=arr.array('i',[1,2,3,4,5])My_Array[::-1]
```

Output: array('i', [5, 4, 3, 2, 1])

[::-1] reprints a reversed copy of ordered data structures such as an array or a list. the original array or list remains unchanged.

Q33. How can you randomize the items of a list in place in Python?

Answer: Consider the example shown below:

```
from random import shuffle

x = ['Keep', 'The', 'Blue', 'Flag', 'Flying', 'High']shuffle(x)

print(x)
```

The output of the following code is as below.

['Flying', 'Keep', 'Blue', 'High', 'The', 'Flag']

Q34. What are python iterators?

Answer: Iterators are objects which can be traversed though or iterated upon.

Q35. How can you generate random numbers in Python?

Answer: Random module is the standard module that is used to generate a random number. The method is defined as:

```
import random

random.random
```

The statement random.random() method return the floating-point numberthat is in the range of [0, 1). The function generates random float numbers.The methods that are used with the random class are the bound methods of

the hidden instances. The instances of the Random can be done to show the multi-threading programs that creates a different instance of individual threads. The other random generators that are used in this are:

1. **randrange(a, b)**: it chooses an integer and define the range in-between [a, b). It returns the elements by selecting it randomly from the range that is specified. It doesn't build a range object.

2. **uniform(a, b)**: it chooses a floating point number that is defined in the range of [a,b).Iyt returns the floating point number

3. **normalvariate(mean, sdev)**: it is used for the normal distribution where the mu is a mean and the sdev is a sigma that is used for standard deviation.

4. **The Random class** that is used and instantiated creates independent multiple random number generators.

Q36. What is the difference between range & xrange?

Answer: For the most part, xrange and range are the exact same in terms of functionality. They both provide a way to generate a list of integers for you to use, however you please. The only difference is that range returns a Python listobject and x range returns an xrange object.

This means that xrange doesn't actually generate a static list at run-time like range does. It creates the values as you need them with a special technique called yielding. This technique is used with a type of object known as generators. That means that if you have a really gigantic range you'd like to generate a list for, say one billion, xrange is the function to use.

This is especially true if you have a really memory sensitive system such as a cell phone that you are working with, as range will use as much memory as it can to create your array of integers, which can result in a Memory Error and crash your program. It's a memory hungry beast.

Q37. How do you write comments in python?

Answer: Comments in Python start with a # character. However, alternatively at times, commenting is done using docstrings(strings enclosed within triple quotes).

Example:

```
<span data-mce-type="bookmark" style="display: inline-block; width:
0px; overflow: hidden; line-height: 0;"
class="mce_SELRES_end"></span>

<pre><span>#Comments in Python start like thisprint("Comments

in Python start with a #")
```

Output: Comments in Python start with a #

Q38. What is pickling and unpickling?

Answer: Pickle module accepts any Python object and converts it into a string representation and dumps it into a file by using dump function, this process is called pickling. While the process of retrieving original Python objects from the stored string representation is called unpickling.

Q39. What are the generators in python?

Answer: Functions that return an iterable set of items are called generators.

Q40. How will you capitalize the first letter of string?

Answer: In Python, the capitalize() method capitalizes the first letter of a string.If the string already consists of a capital letter at the beginning, then, it returns the original string.

Q41. How will you convert a string to all lowercase?

Answer: To convert a string to lowercase, lower() function can be used.

Example:

```
stg='ABCD'

print(stg.lower())
```

Output: abcd

Q42. How to comment multiple lines in python?

Answer: Multi-line comments appear in more than one line. All the lines to be commented are to be prefixed by a #. You can also a very good **shortcut method to comment multiple lines**. All you need to do is hold the ctrl keyand **left click** in every place wherever you want to include a # character and type a # just once. This will comment all the lines where you introduced your cursor.

Q43.What are docstrings in Python?

Answer: Docstrings are not actually comments, but they are **documentation strings**. These docstrings are within triple quotes. They are not assigned to any variable and therefore, at times, serve the purpose of comments as well.

Example:

```
"""

Using docstring as a comment.This code

divides 2 numbers """

x=8

y=4

z=x/y

print(z)
```

Output: 2.0

Q44. What is the purpose of 'is', 'not' and 'in' operators?

Answer: Operators are special functions. They take one or more values and produce a corresponding result.

is: returns true when 2 operands are true (Example: "a" is 'a')

not: returns the inverse of the boolean value

in: checks if some element is present in some sequence

Q45. What is the usage of help() and dir() function in Python?

Answer: Help() and dir() both functions are accessible from the Python interpreter and used for viewing a consolidated dump of built-in functions.

1. **Help() function**: The help() function is used to display the documentation string and also facilitates you to see the help related to modules, keywords, attributes, etc.
2. **Dir() function**: The dir() function is used to display the defined symbols.

Q46. Whenever Python exits, why isn't all the memory de-allocated?

Answer:

1. Whenever Python exits, especially those Python modules which are having circular references to other objects or the objects that are referenced from the global namespaces are not always de-allocated or freed.
2. It is impossible to de-allocate those portions of memory that are reserved by the C library.
3. On exit, because of having its own efficient clean up mechanism, Python would try to de-allocate/destroy every other object.

Q47. What is a dictionary in Python?

Answer: The built-in datatypes in Python is called dictionary. It defines one-to- one relationship between keys and values. Dictionaries contain pair of keys and their corresponding values. Dictionaries are indexed by keys.

Let's take an example:

The following example contains some keys. Country, Capital & PM. Their corresponding values are India, Delhi and Modi respectively.

```
dict={'Country':'India','Capital':'Delhi','PM':'Modi'}print dict[Country]
```

Output:India

```
print dict[Capital]
```

Output:Delhi

```
print dict[PM]
```

Output:Modi

Q48. How can the ternary operators be used in python?

Answer: The Ternary operator is the operator that is used to show the conditional statements. This consists of the true or false values with a statement that has to be evaluated for it.

Syntax:

The Ternary operator will be given as:

```
[on_true] if [expression] else [on_false]x, y = 25, 50big = x if x < y else y
```

Example:

The expression gets evaluated like if x<y else y, in this case if x<y is true then the value is returned as big=x and if it is incorrect then big=y will be sent as a result.

Q49. What does this mean: *args, **kwargs? And why would we use it?

Answer: We use *args when we aren't sure how many arguments are going to be passed to a function, or if we want to pass a stored list or tuple of arguments to a function. **kwargs is used when we don't know how many keyword arguments will be passed to a function, or it can be used to pass the values of a dictionary as keyword arguments. The identifiers args and kwargs are a convention, you could also use *bob and **billy but that would not be wise.

Q50. What does len() do?

Answer: It is used to determine the length of a string, a list, an array, etc.

Example:

```
stg='ABCD'
len(stg)
```

Output:4

Q51. Explain split(), sub(), subn() methods of "re" module in Python.

Answer: To modify the strings, Python's "re" module is providing 3 methods. They are:

- **split()** – uses a regex pattern to "split" a given string into a list.
- **sub()** – finds all substrings where the regex pattern matches and then

replace them with a different string

- **subn()** – it is similar to sub() and also returns the new string along withthe no. of replacements.

Q52. What are negative indexes and why are they used?

Answer: The sequences in Python are indexed and it consists of the positive as well as negative numbers. The numbers that are positive uses '0' that is uses as first index and '1' as the second index and the process goes on like that.

The index for the negative number starts from '-1' that represents the last index in the sequence and '-2' as the penultimate index and the sequence carries forward like the positive number.

The negative index is used to remove any new-line spaces from the string and allow the string to except the last character that is given as S[:-1]. The negative index is also used to show the index to represent the string in correct order.

Q53. What are Python packages?

Answer: Python packages are namespaces containing multiple modules.

Q54.How can files be deleted in Python?

Answer: To delete a file in Python, you need to import the OS Module. After that, you need to use the os.remove() function.

Example:

```
import os

os.remove("xyz.txt")
```

Q55. What are the built-in types of Python?

Answer: Built-in types in Python are as follows –

- Integers
- Floating-point
- Complex numbers
- Strings
- Boolean
- Built-in functions

Q56. What advantages do NumPy arrays offer over (nested) Python lists?

Answer:

1. Python's lists are efficient general-purpose containers. They support (fairly) efficient insertion, deletion, appending, and concatenation, and Python's list comprehensions make them easy to construct and manipulate.
2. They have certain limitations: they don't support "vectorized" operations like elementwise addition and multiplication, and the fact that they can contain objects of differing types mean that Python must store type information for every element and must execute type dispatching code when operating on each element.
3. NumPy is not just more efficient; it is also more convenient. You get a lot of vector and matrix operations for free, which sometimes allow one to avoid unnecessary work. And they are also efficiently implemented.
4. NumPy array is faster and You get a lot built in with NumPy, FFTs, convolutions, fast searching, basic statistics, linear algebra, histograms, etc.

Q57. How to add values to a python array?

Answer: Elements can be added to an array using the **append()**, **extend()** andthe **insert (i,x)** functions.

Example:

```
a=arr.array('d', [1.1 , 2.1 ,3.1] )a.append(3.4)

print(a) a.extend([4.5,6.3,6.8])

print(a) a.insert(2,3.8) print(a)
```

Output:

```
array('d', [1.1, 2.1, 3.1, 3.4])

array('d', [1.1, 2.1, 3.1, 3.4, 4.5, 6.3, 6.8])

array('d', [1.1, 2.1, 3.8, 3.1, 3.4, 4.5, 6.3, 6.8])
```

Q58. How to remove values to a python array?

Answer: Array elements can be removed using **pop()** or **remove()** method. The difference between these two functions is that the former returns the deleted value whereas the latter does not.

Example:

```
a=arr.array('d', [1.1, 2.2, 3.8, 3.1, 3.7, 1.2, 4.6])

print(a.pop())

print(a.pop(3))

a.remove(1.1)

print(a)
```

Output:

4.6

3.1

array('d', [2.2, 3.8, 3.7, 1.2])

Q59. Does Python have OOps concepts?

Answer: Python is an object-oriented programming language. This means that any program can be solved in python by creating an object model. However, Python can be treated as a procedural as well as structural language.

Q60. What is the difference between deep and shallow copy?

Answer: Shallow copy is used when a new instance type gets created and it keeps the values that are copied in the new instance. Shallow copy is used to copy the reference pointers just like it copies the values. These referencespoint to the original objects and the changes made in any member of the class will also affect the original copy of it. Shallow copy allows faster execution of the program and it depends on the size of the data that is used.

Deep copy is used to store the values that are already copied. Deep copy doesn't copy the reference pointers to the objects. It makes the reference to an object and the new object that is pointed by some other object gets stored. The changes made in the original copy won't affect any other copy that uses the object. Deep copy makes execution of the program slower due to making certain copies for each object that is been called.

Q61. How is Multithreading achieved in Python?

Answer:

1. Python has a multi-threading package but if you want to multi-thread to speed your code up, then it's usually not a good idea to use it.

2. Python has a construct called the Global Interpreter Lock (GIL). The GIL makes sure that only one of your 'threads' can execute at any one time. A thread acquires the GIL, does a little work, then passes the GIL onto the next thread.

3. This happens very quickly so to the human eye it may seem like your threads are executing in parallel, but they are really just taking turns using the same CPU core.

4. All this GIL passing adds overhead to execution. This means that if you want to make your code run faster then using the threading package often isn't a good idea.

Q62. What is the process of compilation and linking in python?

Answer: The compiling and linking allow the new extensions to be compiled properly without any error and the linking can be done only when it passes the compiled procedure. If the dynamic loading is used then it depends on the style that is being provided with the system. The python interpreter can be used to provide the dynamic loading of the configuration setup files and will rebuild the interpreter.

The steps that are required in this as:

1. Create a file with any name and in any language that is supported by the compiler of your system. For example file.c or file.cpp

2. Place this file in the Modules/ directory of the distribution which isgetting used.

3. Add a line in the file Setup.local that is present in the Modules/ directory.

4. Run the file using spam file.o

5. After a successful run of this rebuild the interpreter by using the make command on the top-level directory.

6. If the file is changed then run rebuildMakefile by using the command as 'make Makefile'.

Q63. What are Python libraries? Name a few of them.

Answer: Python libraries are a collection of Python packages. Some of the majorly used python libraries are – Numpy, Pandas, Matplotlib, Scikit-learn and many more.

Q64. What is split used for?

Answer: The split() method is used to separate a given String in Python.

Example:

```
a="edureka python"

print(a.split())
```

Output: ['edureka', 'python']

Q65. How to import modules in python?

Answer: Modules can be imported using the **import** keyword. You can import modules in three ways-

Example:

```
import array                    #importing using the originalmodule

                                #name

import array as arr             # importing using an aliasname

from array import *             #imports everything present inthe

                                #array module
```

Next, in this Python Interview Questions document, let's have a look at Object Oriented Concepts in Python.

OOPS in Python Interview Preparation

Q66. Explain Inheritance in Python with an example.

Answer: Inheritance allows One class to gain all the members (say attributes and methods) of another class. Inheritance provides code reusability and makes it easier to create and maintain an application. The class from which we are inheriting is called super-class and the class that is inherited is called a derived / child class.

They are different types of inheritance supported by Python:

1. **Single Inheritance** – where a derived class acquires the members of a single superclass.
2. **Multi-level inheritance** – a derived class d1 is inherited from base class base1, and d2 is inherited from base2.
3. **Hierarchical inheritance** – from one base class you can inherit any number of child classes
4. **Multiple inheritance** – a derived class is inherited from more than one base class.

Q67. How are classes created in Python?

Answer: Class in Python is created using the **class** keyword.

Example:

```
class Employee:

def_____init_____(self, name):

self.name = name
```

```
E1=Employee("abc")

print(E1.name)
```

Q68. What is monkey patching in Python?

Answer: In Python, the term monkey patch only refers to dynamic modifications of a class or module at run-time.

Consider the below example:

```
# m.py

class MyClass:

def f(self):

print "f()"
```

We can then run the monkey-patch testing like this:

```
import m

def monkey_f(self):print

"monkey_f()"


m.MyClass.f = monkey_fobj =

m.MyClass() obj.f()
```

The output will be as below:

```
monkey_f()
```

As we can see, we did make some changes in the behavior of *f()* in *MyClass* using the function we defined, *monkey_f()*, outside of the module *m*.

Q69. Does python support multiple inheritance?

Answer: Multiple inheritance means that a class can be derived from more than one parent class. Python does support multiple inheritance, unlike Java.

Q70. What is Polymorphism in Python?

Answer: Polymorphism means the ability to take multiple forms. So,for instance, if the parent class has a method named ABC then the childclass also can have a method with the same name ABC having itsparameters and variables. Python allows polymorphism.

Q71. Define encapsulation in Python.

Answer: Encapsulation means binding the code and the data together. A Python class is an example of encapsulation.

Q72. How do you do data abstraction in Python?

Answer: Data Abstraction is providing only the required details and hides the implementation from the world. It can be achieved in Python by using interfaces and abstract classes.

Q73.Does python make use of access specifiers?

Answer: Python does not deprive access to an instance variable or function. Python lays down the concept of prefixing the name of the variable, function, or method with a single or double underscore to imitate the behavior of protected and private access specifiers.

Q74. How to create an empty class in Python?

Answer: An empty class is a class that does not have any code defined within its block. It can be created using the pass keyword. However,you can create objects of this class outside the class itself.

IN PYTHON THE PASS command does nothing when it's executed. it's a null statement.

For example:

```
class a:
    pass obj=a()
obj.name="xyz"
print("Name = ",obj.name)
```

Output:

```
Name =  xyz
```


Q75. What does an object() do?

Answer: It returns a featureless object that is a base for all classes. Also,it does not take any parameters.

Next, let us have a look at some Basic Python Programs in these Python Interview Questions.

Basic Python Programs – Interview Preparation

Q76. Write a program in Python to execute the Bubble sort algorithm.

Answer:

```
def bs(a):
# a = name of listb=len(a)-
    1nbsp;
# minus 1 because we always compare 2 adjacent values
    for x in range(b):
        for y in range(b-x):
            a[y]=a[y+1]
a=[32,5,3,6,7,54,87]
bs(a)
```

Output: [3, 5, 6, 7, 32, 54, 87]

Q77. Write a program in Python to produce a Star triangle.

Answer:

```
def pyfunc(r):
    for x in range(r):
```

```
            print(' '*(r-x-1)+'*'*(2*x+1))pyfunc(9)
```

Output:

```
                *
               ***
              *****
             *******
            *********
           ***********
          *************
         ***************
        *****************
```


Q78. Write a program to produce the Fibonacci series in Python.

Answer:

```
# Enter number of terms neededbnbsp;#0,1,1,2,3,5....
a=int(input("Enter the terms"))f=0;#first
element of series s=1#second element    of
seriesif a=0:
    print("The requested series is",f)
else:
    print(f,s,end=" ")
    for x in range(2,a):
```

```
            print(next,end=" ")f=s
            s=next
```

Output: Enter the terms 5 0 1 1 2 3

Q79. Write a program in Python to check if a number is prime.

Answer:

```
a=int(input("enter number"))
if a=1:
    for x in range(2,a):
        if(a%x)==0: print("not
          prime")
    break
    else:
        print("Prime")
else:
    print("not prime")
```

Output:

enter number 3

Prime

Q80. Write a program in Python to check if a sequence is a Palindrome.

Answer:

```
a=input("enter sequence")b=a[::-1]
if a==b: print("palindrome")
else:
    print("Not a Palindrome")
```

Output:

```
enter sequence 323 palindrome
```

Q81. Write a one-liner that will count the number of capital letters in a file. Your code should work even if the file is too big to fit in memory.

Answer: Let us first write a multiple-line solution and then convert it to one-liner code.

```
with open(SOME_LARGE_FILE) as fh:
count = 0
text = fh.read()
for character in text:
    if character.isupper():count += 1
```

We will now try to transform this into a single line.

```
count sum(1 for line in fh for character in lineif character.isupper())
```

Q82. Write a sorting algorithm for a numerical dataset in Python.

Answer: The following code can be used to sort a list in Python:

```
list = ["1", "4", "0", "6", "9"]

list = [int(i) for i in list]list.sort()

print (list)
```

Q83. Looking at the below code, write down the final values of A0, A1, …An.

Answer:

```
A0 = dict(zip(('a','b','c','d','e'),(1,2,3,4,5)))

A1 = range(10)A2 = sorted([i for i in A1 if i in A0])A3 = sorted([A0[s] for s in
A0])

A4 = [i for i in A1 if i in A3]A5 = {i:i*i for i in
A1}

A6 = [[i,i*i] for i in A1]

print(A0,A1,A2,A3,A4,A5,A6)
```

The following will be the final outputs of A0, A1, … A6

A0 = {'a': 1, 'c': 3, 'b': 2, 'e': 5, 'd': 4} # the order may vary

A1 = range(0, 10)

A2 = []

A3 = [1, 2, 3, 4, 5]

A4 = [1, 2, 3, 4, 5]

A5 = {0: 0, 1: 1, 2: 4, 3: 9, 4: 16, 5: 25, 6: 36, 7: 49, 8: 64, 9: 81}

A6 = [[0, 0], [1, 1], [2, 4], [3, 9], [4, 16], [5, 25], [6, 36], [7, 49], [8, 64], [9, 81]]

Web - Scraping Python Interview Questions

Q84. How To Save An Image Locally Using Python Whose URL AddressI Already Know?

Answer: We will use the following code to save an image locally froman URL address

```
import urllib.request

urllib.request.urlretrieve("URL", "local-filename.jpg")
```

Q85. How can you get the Google cache age of any URL or web page?

Answer: Use the following URL format:

http://webcache.googleusercontent.com/search?q=cache:URLGOESHE RE

Be sure to replace "URLGOESHERE" with the proper web address of the page or site whose cache you want to retrieve and see the time for. For example, to check the Google Web cache age of edureka.co you'd use the following URL:

http://webcache.googleusercontent.com/search?q=cache:edureka.co

Q86. You are required to scrap data from IMDb's top 250 movies page.It should only have fields like movie name, year, and rating.

Answer: We will use the following lines of code:

```
from bs4 import BeautifulSoup

import requests

import sys


url = '<a href="http://www.imdb.com/chart/top">http://www.
imdb.com/chart/top</a>'

response = requests.get(url)

soup = BeautifulSoup(response.text)tr =

soup.findChildren("tr")

tr = iter(tr)next(tr)


for movie in tr:

title = movie.find('td', {'class': 'titleColumn'}
).find('a').contents[0]

year = movie.find('td', {'class': 'titleColumn'} ).find('span',
{'class':'secondaryInfo'}).contents[0]

rating = movie.find('td', {'class':'ratingColumn
imdbRating'} ).find('strong').contents[0]

row = title + ' - ' + year + ' ' + ' ' + rating


print(row)
```

The above code will help scrape data from IMDb's top 250 list

Data Analysis – Python Interview Questions

Q87. What is map function in Python?

Answer: *map* function executes the function given as the first argument on all the elements of the iterable given as the second argument. If the function given takes in more than 1 arguments, then many iterables are given. #Followthe link to know more similar functions.

Q88. Is python numpy better than lists?

Answer: We use python numpy array instead of a list because of the below three reasons:

1. Less Memory
2. Fast
3. Convenient

Q89. How to get indices of N maximum values in a NumPy array?

Answer: We can get the indices of N maximum values in a NumPy array using the below code:

```
import numpy as np

arr = np.array([1, 3, 2, 4, 5])

print(arr.argsort()[-3:][::-1])
```

Output

[4 3 1]

Q90. How do you calculate percentiles with Python/ NumPy?

Answer: We can calculate percentiles with the following code:

```
import numpy as np

a = np.array([1,2,3,4,5])

p = np.percentile(a, 50) #Returns 50th percentile,
e.g. median

print(p)
```

Output:3

Q91. What is the difference between NumPy and SciPy?

Answer:

NumPy	SciPy
It refers to Numerical python.	It refers to Scientific python.
It has fewer new scientific computing features.	Most new scientific computing features belong in SciPy.
It contains less linear algebra functions.	It has more fully-featured versions of the linear algebra modules, as well as many other numerical algorithms.

NumPy has a faster processing speed.	SciPy on the other hand has slower computational speed.

Q92. How do you make 3D plots/visualizations using NumPy/SciPy?

Answer: Like 2D plotting, 3D graphics is beyond the scope of NumPy and SciPy, but just as in the 2D case, packages exist that integrate with NumPy. Matplotlib provides basic 3D plotting in the mplot3d subpackage, whereas Mayavi provides a wide range of high-quality 3D visualization features,utilizing the powerful VTK engine.

Next in this Python Interview Questions blog, let's have a look at some MCQs

Multiple Choice Questions (MCQ) –

Interview Preparation

Q93. Which of the following statements create a dictionary? (Multiple Correct Answers Possible)

a) d = {}
b) d = {“john”:40, “peter”:45}
c) d = {40:”john”, 45:”peter”}
d) d = (40:”john”, 45:”50”)

Answer: b, c & d.

Dictionaries are created by specifying keys and values.

Q94. Which one of these is floor division?

a) /
b) //
c) %
d) None of the mentioned

Answer: b) //

When both of the operands are integer then python chops out the fraction part and gives you the round off value, to get the accurate answer use floor division. For ex, 5/2 = 2.5 but both of the operands are integer so answer of this expression in python is 2. To get the 2.5 as the answer, use floor division using //. So, 5//2 = 2.5

Q95. What is the maximum possible length of an identifier?

a) 31 characters
b) 63 characters
c) 79 characters
d) None of the above

Answer: d) None of the above

Identifiers can be of any length.

Q96. Why are local variable names beginning with an underscore discouraged?

a) they are used to indicate a private variable of a class
b) they confuse the interpreter
c) they are used to indicate global variables
d) they slow down execution

Answer: a) they are used to indicate a private variable of a class

As Python has no concept of private variables, leading underscores are used to indicate variables that must not be accessed from outside the class.

Q97. Which of the following is an invalid statement?

a) abc = 1,000,000
b) a b c = 1000 2000 3000
c) a,b,c = 1000, 2000, 3000
d) a_b_c = 1,000,000

Answer: b) a b c = 1000 2000 3000 Spaces

are not allowed in variable names.

Q98. What is the output of the following?

```
try:
        if '1' != 1:
                raise "someError"
        else:
                print("someError has not occured")
except "someError":
        print ("someError has occured")
```

a) someError has occured
b) someError has not occured
c) invalid code
d) none of the above

Answer: c) invalid code

A new exception class must inherit from a BaseException. There is no such inheritance here.

Q99. Suppose list1 is [2, 33, 222, 14, 25], What is list1[-1] ?

a) Error
b) None
c) 25
d) 2

Answer: c) 25

The index -1 corresponds to the last index in the list.

Q100. To open a file c:scores.txt for writing, we use

a) outfile = open(“c:scores.txt”, “r”)
b) outfile = open(“c:scores.txt”, “w”)
c) outfile = open(file = “c:scores.txt”, “r”)
d) outfile = open(file = “c:scores.txt”, “o”)

Answer: b) The location contains double slashes ( ) and w is used to indicate that file is being written to.

Q101. What is the output of the following?

```
f = None
for i in range (5):
with open("data.txt", "w") as f:
if (i > 2):
break
print f.closed
```

a) True
b) False
c) None
d) Error

Answer: a) True

The WITH statement when used with open file guarantees that the file objectis closed when the with block exits.

Q102. When will the else part of try-except-else be executed?

a) always
b) when an exception occurs
c) when no exception occurs
d) when an exception occurs into except block

Answer: c) when no exception occurs

The else part is executed when no exception occurs.

This set of Python Interview Questions will help you in preparing for yourinterviews. All the best!

Probability and Statistics Interview Questions

In today's world of data-based analytics, skills like statistical analysis are not only necessary but also an important requirement for the job role of data scientists and analysts. A data scientist must be fluent with probabilistic measures and statistics to generate the most useful insights from the data. Following are the most frequently asked interview questions in this domain.

Q1. Walkthrough the probability fundamentals

Answer: The possibility of the occurrence of an event, among all the possible outcomes, is known as its probability. The probability of an event always lies between(including) 0 and 1.

$$P(A) = \frac{(Number\ of\ times\ event\ A\ occurred)}{(Number\ of\ times\ experiment\ performed)}$$

Factorial - it is used to find the total number of ways n number ofthings can be arranged in n places without repetition. Its value is n multiplied by all natural numbers to n-1.

For example 5!=5X4X3X2X1=120

Permutation - It is used when replacement is not allowed, and the order of items is important. Its formula is:

$$_nP_r = \frac{n!}{(n-r)!}$$

Where,

n is the total number of items

R is the number of ways items are being selected

Combination - It is used when replacement is not allowed, and theorder of items is not important.

Its formula is:

$$_nC_r = \frac{n!}{(n-r)!r!}$$

Some probability rules are:

1. **Addition Rule**: P(A or B)= P(A) + P(B) – P(A and B)
2. **Conditional probability** It is the probability of event B occurring, assuming that event A has already occurred.

P(A and B)= P(A) . P(B|A)

3. **Central Limit theorem:** It states that when we draw random samples from a large population and take the mean of these

samples, they form a normal distribution.

Q2. What is Selection Bias?

Answer: Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn't random. It is sometimes referred to as the selection effect. It is the distortion of statistical analysis, resulting from the method of collecting samples. If the selection bias is not considered, then some conclusions of the studymay not be accurate.

The types of selection bias include:

1. **Sampling bias**: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
2. **Time interval**: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.
3. **Data**: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
4. **Attrition**: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

Q3. What is a bias-variance trade-off?

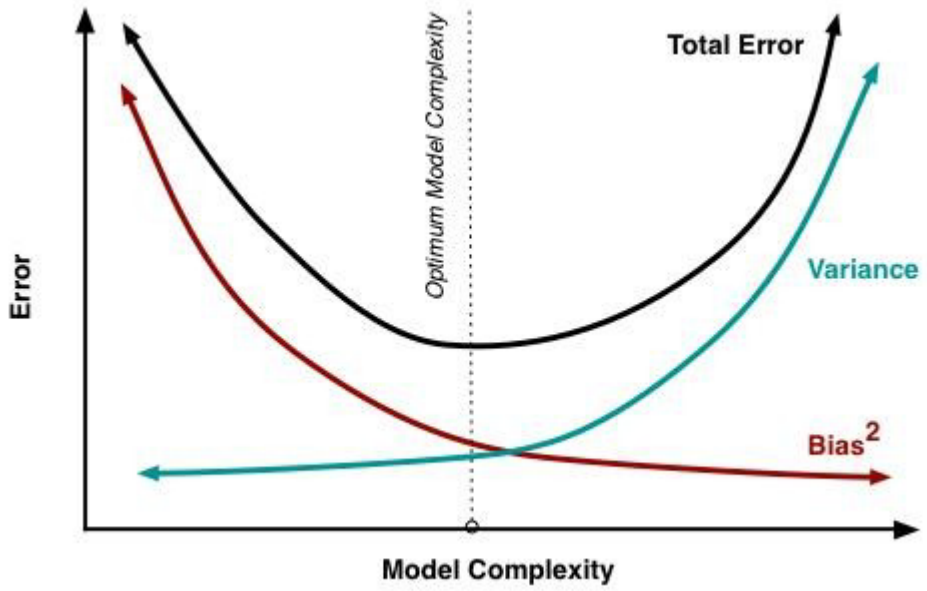
Answer: Bias is an error introduced in your model due to the oversimplification of the machine learning algorithm. It can lead to underfitting. When you train your model at that time model makes simplified assumptions to make the target function easier to understand.

Low bias machine learning algorithms — Decision Trees, k-NN, and SVM
High bias machine learning algorithms — Linear Regression, Logistic Regression

Variance: Variance is the error introduced in your model due to a complex machine learning algorithm, your model learns noise also fromthe training data set and performs badlythe on test data set. It can lead to high

sensitivity and overfitting.

Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens until a particular point. As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.



Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.

1. The k-nearest neighbor algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of nneighborsthat contribute to the prediction and in turn increases the bias of the model.

2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.

Q4. Describe different regularization methods, such as L1 and L2 regularization

Answer: There are 3 important regularization methods as follows:

L2 regularization-(Ridge regression): In L2 regularization, we add the sum of the squares of all the weights, multiplied by a value lambda, to the loss function. The formula for Ridge regression is as follows:

$$\text{Loss} = \sum_{j=1}^{m}\left(Yi - Wo - \sum_{i=1}^{n} Wi\,Xji\right)^2 + \lambda \sum_{i=1}^{n} Wi^2$$

As you can see, if the value of the weights is multiplied by the data value for a particular data point and the feature becomes very large, the original loss will become small. But the added value of lambda multiplied by the sum of squares of weights will become large as well. Similarly, if the original loss value becomes very large, the added value will become small. Thus, it will control the final value from becoming too large or too small.

L1 Regularization-(Lasso regression): In L1 regularization, we add the sum of the absolute values of all the weights, multiplied by a value lambda, to the loss function. The formula for Lasso regression is as follows:

$$\text{Loss} = \sum_{j=1}^{m}\left(Yi - Wo - \sum_{i=1}^{n} Wi\,Xji\right)^2 + \lambda \sum_{i=1}^{n} |Wi|$$

The loss function along with the optimization algorithm brings parameters near to zero but not zero, while lasso eliminates less important features and sets respective weight values to zero.

Dropout

This is used for regularization in neural networks. Fully connected layers are more prone to overfitting. Dropout leaves out some neurons with a 1-p probability in neural networks. Dropout reduces overfitting, improves training speed, and makes the model more robust.

Q5. How should you maintain a deployed model?

Answer: After a model has been deployed, it needs to be maintained. The data being fed may change over time. For example, in the case of a model predicting house prices, the prices of houses may rise over time or fluctuate due to some other factor. The accuracy of the model on new data can be recorded. Some common ways to ensure accuracy include-

1. The model should be frequently checked by feeding negative test data. If the model gives low accuracy with negative test data, it is fine.

2. An Auto Encoder should be built that Uses anomaly detection techniques, the AE model will calculate the Reconstruction error value. If the Reconstruction error value is high, it means the new data does not follow the old pattern learned by the model.

If the model shows good prediction accuracy with new data, it means that the new data follows the pattern or the generalization learned by the model on old data. So, the model can be retrained on the new data. If the accuracy of new data is not that good, the model can be retrained on the new data with feature engineering on the data features along with the old data.

If the accuracy is not good, the model may need to be trained from scratch.

Q6. Write the equation and calculate the precision and recall rate.

Answer: Precision quantifies the number of correct positive predictions made. Precision is calculated as the number of true positives divided by the total number of true positives and false positives.

Precision = True Positives / (True Positives + False Positives)

Precision is defined as the number of correct positive predictions made out of all positive predictions that could have been made. The recall is calculated as the number of true positives divided by the total number of true positives and false negatives.

Recall = True Positives / (True Positives + False Negatives)

Q7. Why do we use the summary function?

Answer: Summary functions are used to give the summary of all the numeric values in a dataframe.

For Example, The describe() function can be used to provide the summary of all the data values given to it.

column_name.describe() will give the following values of all the numeric

data in the column:

1. Count
2. Mean
3. Std-Standard deviation
4. Min-Minimum
5. 25%
6. 50%
7. 75%
8. max-Maximum

Q8. How will you measure the Euclidean distance between the two arrays in NumPy?

Answer: The Euclidean distance between 2 arrays A[1,2,3,] and B[8,9,10] can be calculated by taking the Euclidean distance of each point respectively. The built-in function numpy.linalg.norm()can be used as follows:


```
import numpy as np
```

```
[10] A=np.array([1,2,3])
     B=np.array([8,9,10])
```

```
[11] numpy.linalg.norm(A-B)

     12.12435565298214
```

Q9. What is the difference between an error and a residual error?

Answer: An error refers to the difference between the predicted value and the actual value. The most popular means for calculating errors in data science are Mean Absolute Error (MAE), Mean Squared Error (MSE), and

Root Mean Squared Error (RMSE). While residual is thedifference between a group of values observed and their arithmetical mean. An error is generally unobservable while a residual error can be visualized on a graph. Error represents how observed data differs from the actual population. While a residual represents the way observed data differs from the sample population data.

Q10. Difference between Normalisation and Standardization?

Answer: Normalization, also known as min-max scaling, is a technique where all the data values are converted such that they lie between 0 and 1.

The formula for Normalization is:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where,

X_max is the maximum value of feature.

X_min is the minimum value of feature.

Standardization refers to converting our data such that the data has a normal distribution with its mean as 0 and standard deviation as 1.

Statistics Interview Questions

Q11. What is the difference between “long” and “wide” format data?

Answer: In the **wide format**, a subject’s repeated responses will be in a single row, and each response is in a separate column. In the **long format**, each row is a one-time point per subject. You can recognize data in wide format by the fact that columns generally represent groups.

Name	Height	Weight
John	160	67
Christopher	182	78

Figure: *Wide Format*

Name	Attribute	Value
John	Height	160
John	Weight	67
Christopher	Height	182
Christopher	Weight	78

Figure: *Long Format*

Q12. What do you understand by the term Normal Distribution?

Answer: Data is usually distributed in different ways with a bias to the left or the right or it can all be jumbled up.

However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

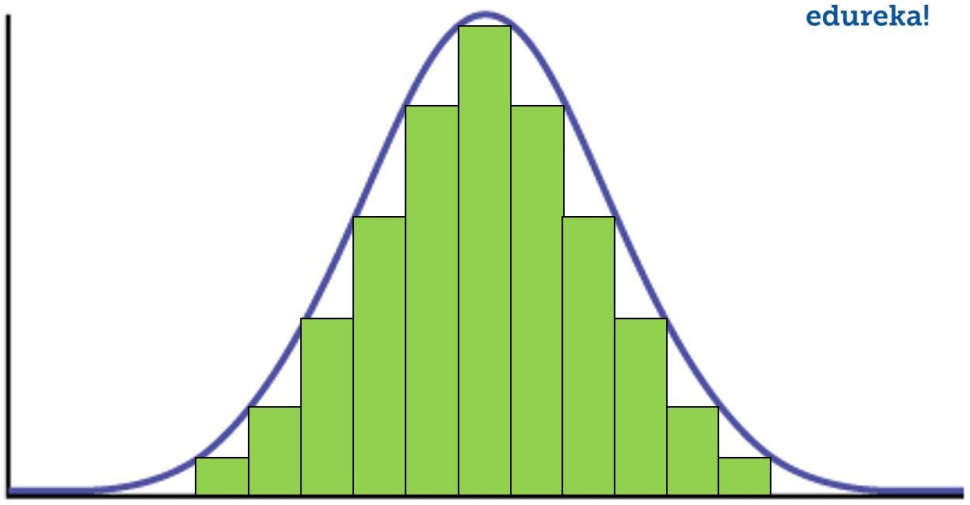


Figure: *Normal distribution in a bell curve*

The random variables are distributed in the form of a symmetrical, bell-shaped curve.

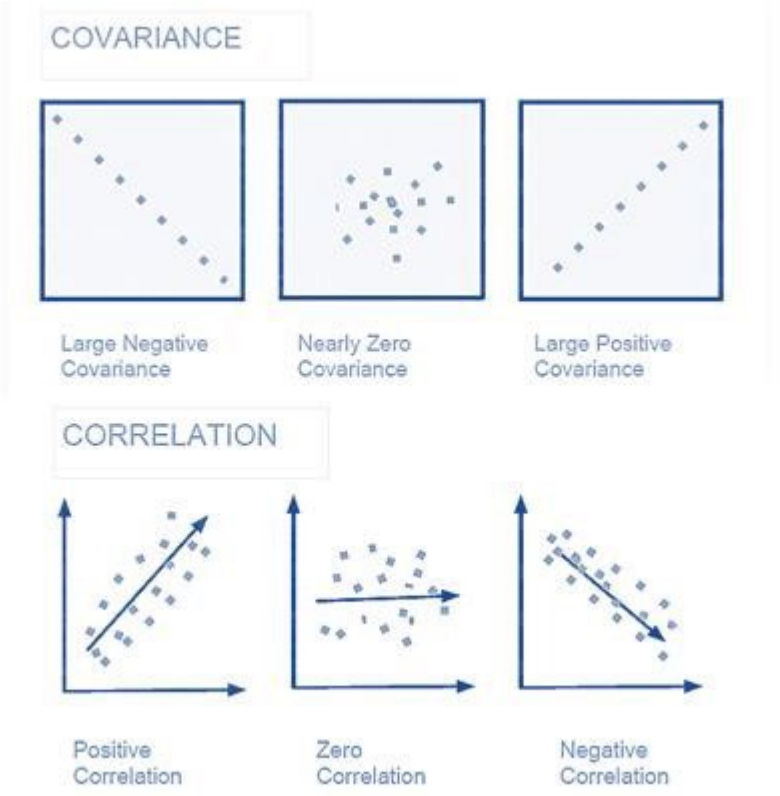
Properties of Normal Distribution are as follows:

1. **Unimodal** -one mode

2. **Symmetrical** -left and right halves are mirror images
3. **Bell-shaped** -maximum height (mode) at the mean
4. Mean, Mode, and Median are all located in the center
5. Asymptotic

Q13. What are correlation and covariance in statistics?

Answer: Covariance and Correlation are two mathematical concepts; these two approaches are widely used in statistics. Both Correlation and Covariance establish the relationship and measure the dependency between two random variables. Though the work is similar between these two in mathematical terms, they are different from each other.



Correlation: Correlation is considered or described as the best technique for measuring and estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.

Covariance: In covariance two items vary together and it's a measure that indicates the extent to which two random variables change in the cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable are reciprocal to a

corresponding change in another variable.

Q14. What is the difference between Point Estimates and Confidence Interval

Answer: Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.

A confidence interval gives us a range of values that is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and is represented by 1 — alpha, where alpha is the level of significance.

Q15. What is the goal of A/B Testing?

Answer: It is a hypothesis testing for a randomized experiment with two variables A and B.

The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business. It can be used to test everything from website copy to sales emails to search ads.

An example of this could be identifying the click-through rate for a banner ad.

Q16. What is a p-value?

Answer: When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. A p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called the Null Hypothesis.

A low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. A high p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the null

Hypothesis p-value of 0.05 indicates the Hypothesis could go either way. To put it another way,

High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null.

Q17. 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?

Answer: The probability of not seeing any shooting star in 15 minutes is:

= 1 – P(Seeing one shooting star)
= 1 – 0.2 = 0.8

Probability of not seeing any shooting star in the period of one hour:

= (0.8) ^ 4 = 0.4096

Probability of seeing at least one shooting star in one hour:

= 1 – P(Not seeing any star)
= 1 – 0.4096 = 0.5904

Q18. How can you generate a random number between 1 – 7 withonly a die?

Answer:

- Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes.
- To get our 7 equal outcomes we must reduce this 36 to a number divisible by 7. We can thus consider only 35 outcomes and exclude the other one.
- A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice.

- All the remaining combinations from (1,1) till (6,5) can be divided into 7 parts of 5 each. This way all the seven sets of outcomes are equally likely.

Q19. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

Answer: In the case of two children, there are 4 equally likely possibilities: **BB, BG, GB, and GG.**

where **B** = Boy and **G** = Girl and the first letter denotes the first child.

From the question, we can exclude the first case of BB. Thus, from the remaining 3 possibilities of **BG**, **GB, and BB**, we must find the probability of the case with two girls.

Thus, P (Having two girls given one girl) = **1 / 3**

Q20. 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 see10 heads, what is the probability that the next toss of that coin is also a head?

Answer: 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 a fair coin = 999/1000 = **0.999**
Probability of selecting an 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 = 0.753

Q21. What do you understand by statistical power of sensitivity and how do you calculate it?

Answer: Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest, etc.).

Sensitivity is nothing but "Predicted True events/ Total events". True events here are the true events, and the model also predicted them as true.

Calculation of seasonality is straightforward:

Seasonality = (True Positives) / (Positives in Actual Dependent Variable)

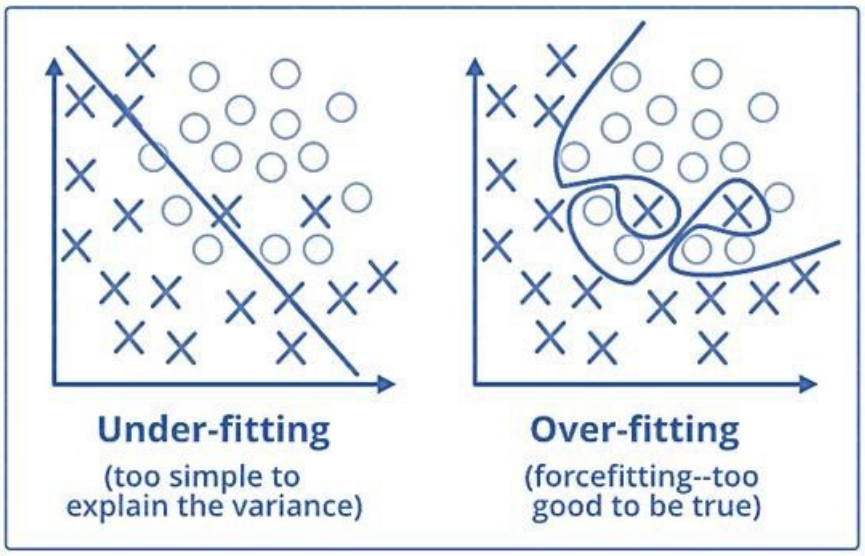
Q22. Why Is Re-sampling Done?

Answer: Resampling is done in any of these cases:

- Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points
- Substituting labels on data points when performing significance tests
- Validating models by using random subsets (bootstrapping, cross-validation)

Q23. What are the differences between over-fitting and under-fitting?

Answer: In statistics and machine learning, one of the most common tasks is to fit a *model* to a set of training data, to be able to make reliable predictions on general untrained data.



In **overfitting**, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfitted, has poor predictive performance, as it overreacts to minor fluctuations in the training data.

Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

Q24. How do combat Overfitting and Underfitting?

Answer: To combat overfitting and underfitting, you can resample the data to estimate the model accuracy (k-fold cross-validation) and by having a validation dataset to evaluate the model.

Q25. What is regularisation? Why is it useful?

Answer: Regularisation is the process of adding tuning parameters to a model to induce smoothness to prevent overfitting. This is most often done by adding a constant multiple to an existing weight vector. This constant is often the L1(Lasso) or L2(ridge). The model predictions should then minimize the loss function calculated on the regularized training set.

Q26. What Is the Law of Large Numbers?

Answer: It is a theorem that describes the result of performing the same experiment many times. This theorem forms the basisof **frequency-style** thinking. It says that the sample means, the sample variance, and the sample standard deviation converge to what they are trying to estimate.

Q27. What Are Confounding Variables?

Answer: In statistics, a confounder is a variable that influences both the dependent variable and the independent variable.

For example, if you are researching whether a lack of exercise leads to weight gain,

lack of exercise = independent variable

weight gain = dependent variable.

A confounding variable here would be any other variable that affectsboth variables, such as the **age of the subject**.

Q28. What Are the Types of Biases That Can Occur During Sampling?

Answer:

- Selection bias
- Under coverage bias
- Survivorship bias

Q29. What is Survivorship Bias?

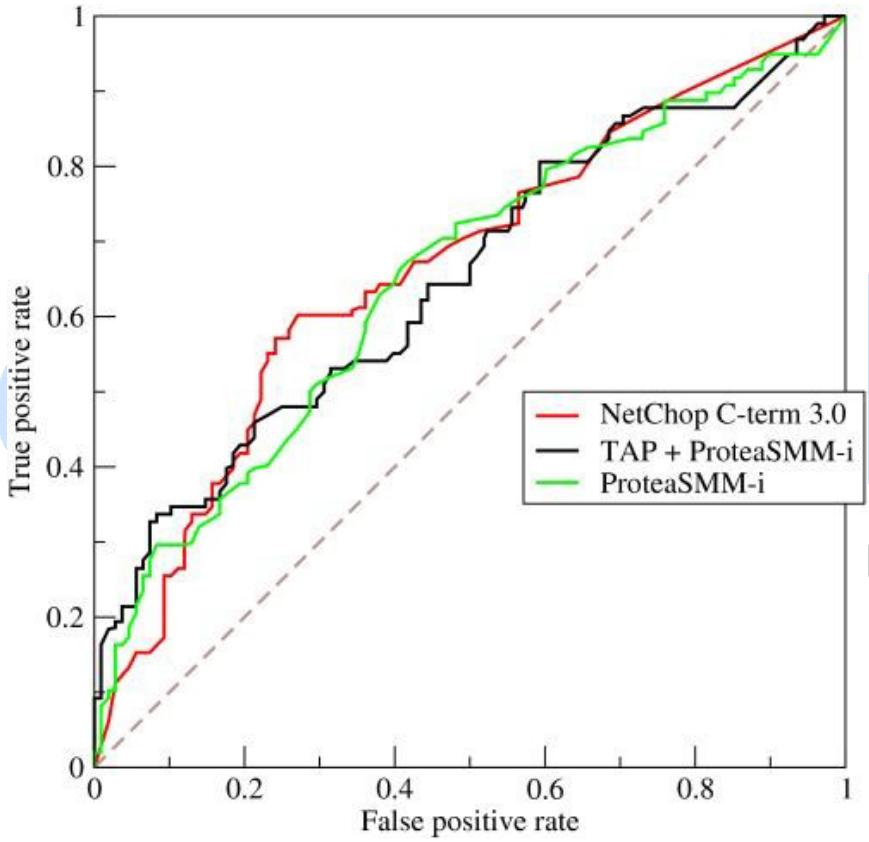
Answer: It is the logical error of focusing on aspects that support surviving some process and casually overlooking those that did notwork because of their lack of prominence. This can lead to wrong conclusions in numerous different means.

Q30. What is selection Bias?

Answer: A Selection bias occurs when the sample obtained is not representative of the population intended to be analyzed.

Q31. Explain how a ROC curve works.

Answer: The **ROC** curve is a graphical representation of the contrast between true positive rates and false-positive rates at various thresholds. It is often used as a proxy for the trade-off between the sensitivity (true positive rate) and the false-positive rate.



Q32. What is TF-IDF vectorization

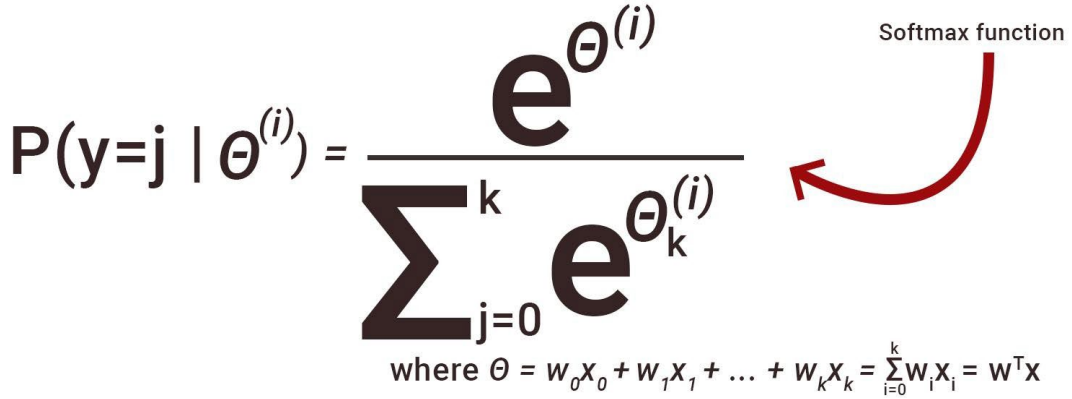
Answer: TF-IDF is short for term frequency-inverse document frequency, which is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Q33. Why do we generally use the Softmax non-linearity function as the last operation in-network?

Answer: It is because it takes in a vector of real numbers and returns a probability distribution. Its definition is as follows. Let x be a vector of real numbers (positive, negative, whatever, there are no constraints).

Then the i^{th} component of Softmax(*x) is:*



It should be clear that the output is a probability distribution: each element is non-negative and the sum over all components is 1.

The formula for Standardization is:

$$X' = \frac{X - \mu}{\sigma}$$

So, while normalization rescales the data into the range from 0 to 1 only, standardization ensures data follows the standard normal distribution.

Q34. There are two types of coins: one fair (heads) and one unfair (tails) (both sides tails). You choose one at random, flip it five times, and note that it lands on tails each time. How likely is it that you are tossing an unfair coin?

Answer: Here, the Bayes Theorem can be used. Let U stand for the scenario in which we flip an unfair coin and F for the scenario in which we flip a fair coin. We are aware that P(U) = P(F) = 0.5 since the coin is picked at random. Let 5T stand for the scenario in which weconsistently flip 5 heads. After that, assuming that we saw 5 tails in a row, we are interested in finding a solution for P(U|5T), or the likelihood that we are tossing an unfair coin.

Since the unjust coin will always land on heads, we know P(5T|U) = 1.

Furthermore, we are aware that P(5T|F) = 1/25 = 1/32 according to the concept of a fair coin. Using the Bayes Theorem, we can:

$$P(U|5T) = \frac{P\left(\frac{5T}{U}\right) \times P(U)}{P\left(\frac{5T}{U}\right) \times P(U) + P\left(\frac{T5}{F}\right) \times P(F)} = \frac{0.5}{0.5 + 0.5 \times \frac{1}{32}}$$

= 0.97

Therefore, the probability we picked the unfair coin is about 97%.

Q35. What is Chi-Square test?

Answer: Chi square is a statistical test used to compare the observed data with the data that we would expect to obtain according to aspecific hypothesis.

Formula for the chi square test is:

chisq.test performs chi-squared contingency table tests and goodness-of-fit tests.

Usage:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{n_{i.}n_{.j}}{n}$$

$$df = (r-1)(c-1)$$

chisq.test(x, y = NULL, correct = TRUE,

p = rep(1/length(x), length(x)), rescale.p =

FALSE,simulate.p.value = FALSE, B = 2000)

Q35. What is the central limit theorem? How is a normal distributiondifferent from chi-square distribution?

Answer: Central limit theorem states that the distribution of an average will tend to be Normal as the sample size increases, regardless of the distribution from which the average is taken except when the moments of the parent distribution do not exist. All practical distributions in statistical engineering have defined moments, and thus the CLT applies.

Chi-square distribution uses standard normal variates which are a part of normal distribution. In statistical terms:

If X is normally distributed with mean μ and variance σ2 > 0, then:

$$V = \left(\frac{X-\mu}{\sigma}\right)^2 = Z^2$$

is distributed as a chi-square random variable with 1 degree of freedom.

Q36. What is F test?

Answer: The **F-test** is designed to test if two population variances are equal. It does this by comparing the ratio of two variances. So, if the variances are equal, the ratio of the variances will be 1.

$$F = \frac{s_1^2}{s_2^2}$$

Usage:

var.test(x, …)

Default S3 method:

var.test(x, y, ratio = 1,

alternative = c(“two.sided”, “less”,

“greater”),conf.level = 0.95, …)

S3 method for class ‘formula’

var.test(formula, data, subset, na.action,

…)

Q37. What is a Z-test and T-Test?

Answer: Z-test is a statistical test where normal distribution is applied

and is basically used for dealing with problems related to large samples when n (sample size) ≥ 30.

It is used to determine whether two population means are different when the variances are known, and the sample size is large. The test statistic is assumed to have a normal distribution and parameters such as standard deviation should be known in order for z-test to be performed.

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

A one-sample location test, two-sample location test, paired difference test and maximum likelihood estimate are examples of tests that can be conducted as z-tests. Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size. Also, t-tests assume that the standard deviation is unknown, while z-tests assume that it is known. If the standard deviation of the population is unknown, the assumption that the sample variance equals the population variance is made.

It implements a z-test similar to the t.test

function.Usage:

simple.z.test(x, sigma, conf.level=0.95)

T-test assesses whether the means of two groups are statistically different from each other

A two-sample t-test examines whether two samples are different and iscommonly used when the variances of two normal distributions are unknown and when an experiment uses a small sample size

For example, a t-test could be used to compare the average floor routine score of the U.S. women's Olympic gymnastic team to the average floor routine score of China's women's team

$$t_{obt} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

t_{obt} = obtained t

$\overline{X}_1$ and $\overline{X}_2$ = means for the two groups

s_1^2 and s_2^2 = variances of the two groups

n_1 and n_2 = number of participants in each of the two groups

It performs one and two sample t-tests on vectors of data.

Usage:

t.test(x, …)

Default S3 method:

t.test(x, y = NULL,

alternative = c("two.sided", "less", "greater"),
mu = 0, paired = FALSE, var.equal =

FALSE,conf.level = 0.95, …)

S3 method for class 'formula'

t.test(formula, data, subset, na.action, …)

Q38. Differentiate between univariate, bivariate, and multivariate analysis.

Answer: **Univariate** *analyses* are descriptive statistical analysis techniques that can be differentiated based on the number of variables involved at a given point in time. For example, the pie charts of sales based on territory involve only one variable and the analysis can be referred to as univariate analysis.

The **bivariate** *analysis* attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sales and spending can be considered as an example of

bivariate analysis.

The multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.

Q39. Explain Star Schema.

Answer: It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.

Q40. What is Cluster Sampling?

Answer: Cluster sampling is a technique used when it becomes difficultto study the target population spread across a wide area and simple random sampling cannot be applied. A cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For Example, A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Q41. What is Systematic Sampling?

Answer: *Systematic sampling* is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example of systematic sampling is the equal probability method.

Q42. What are Eigenvectors and Eigenvalues?

Answer: *Eigenvectors* are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing, or stretching.

Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

Q43. Can you cite some examples where a false positive is more important than a false negative?

Answer: Let us first understand what false positives and false negatives are:

- **False Positives** are the cases where you wrongly classified a non-event as an event a.k.a Type I error.
- **False Negatives** are the cases where you wrongly classify events
- as non-events, a.k.a Type II error.
- **Example 1:** In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn't have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.
- **Example 2:** Let's say an e-commerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least
- $10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above $10,000. Now the issue is if we send the $1000 gift vouchers to customers who have not actually purchased anything but are marked as having made
- $10,000 worth of purchase.

Q44. Can you cite some examples where a false negative important than a false positive?

Answer: Example 1: Assume there is an airport 'A' which has received high-security threats and based on certain characteristics they identify

whether a particular passenger can be a threat or not. Due toa shortage of staff, they decide to scan passengers being predicted as risk positives by their predictive model. What will happen if a true threat customer is flagged as non-threat by the airport model?

Example 2: What if the Jury or judge decides to make a criminal go free?

Example 3: What if you rejected marrying a very good person based on your predictive model and you happen to meet him/her after a few years and realize that you had a false negative?

Q45. Can you cite some examples where both false positives and false negatives are equally important?

Answer: In the **Banking** industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

Q46. Can you explain the difference between a Validation Set and a Test Set?

Answer: A **Validation set** can be considered as a part of the training set as it is used for parameter selection and to avoid overfitting of the model being built.

On the other hand, a **Test Set** is used for testing or evaluating the performance of a trained machine learning model.

In simple terms, the differences can be summarized as; training set is to fit the parameters that is, weights and test set is to assess the performance of the model i.e. evaluating the predictive power and generalization.

Q47. Explain cross-validation.

Answer: Cross-validation is a model validation technique for evaluating how the outcomes of statistical analysis will **generalize** to an **independent dataset**. Mainly used in backgrounds where the objective

is forecast, and one wants to estimate how accurately a model will accomplish in practice.

The goal of cross-validation is to term a data set to test the model in the training phase (that is, validation data set) to limit problems like overfitting and get an insight on how the model willgeneralize to an independent data set.

Q48. What is Hypothesis Testing?

Answer: Hypothesis testing is the backbone behind statistical inference and can be broken down into a couple of topics. The first is the Central Limit Theorem, which plays an important role in studying large samples of data. Other core elements of hypothesis testing are sampling distributions, p-values, confidence intervals, and type I and II errors. Lastly, it is worth looking at various tests involving proportions, and other hypothesis tests.

Most of these concepts play a crucial role in A/B testing, which is a commonly asked topic during interviews at consumer-tech companies like Facebook, Amazon, and Uber. It's useful to not only understand the technical details but also conceptually how A/B testing operates, what the assumptions are, possible pitfalls, and applications to real-life products.

Q49. P and Q are playing a game where P has n+1 coins, Q has n coins, and they each flip all of their coins. What is the probability that P will have more heads than Q?

Answer: Consider the first n coins that P flips, versus the n coins that Q flips.

There are three possible scenarios:

1. P has more heads than Q
2. P and Q have an equal amount of heads
3. P has fewer heads than Q

Notice that in scenario 1, P will always win (irrespective of coin n+1), and in scenario 3, P will always lose (irrespective of coin n+1). By

symmetry, these two scenarios have an equal probability of occurring.

Denote the probability of either scenario as x, or the probability of scenario 2 as y.

We know that 2x + y = 1 since these 3 scenarios are the only possible outcomes. Now let's consider coin n+1. If the flip results in heads, with probability 0.5, then P will have won after scenario 2 (which happens with probability y). Therefore, P's total chances of winning the game are increased by 0.5y.

Thus, the probability that P will win the game is:

$$x + \frac{1}{2}y = \frac{1}{2}(1 - 2x) = \frac{1}{2}$$

Q50. What does the value of R-squared signify?

Answer: The value of R-squared tells us the amount of change in the dependent variable (Y) that is explained by the independent variable (X). The R-squared value can range from 0 to 1.