Department of Information and Technology

Gauhati University Institute of Science & Technology

Gauhati University, GNB Nagar, Ghy-14

Assignment

Data Analytics - IT 936

Name **Manikangkan Das**

Roll No **190102020**

Semester **8th**

Branch **Information Technology**

PPT 1 - INTRODUCTION

Q1. What is the smallest and the largest unit of measuring size of data?

Ans - A bit is the smallest unit used to gauge data size. In computing and digital communications, a bit is the fundamental unit of information and has only two potential values: 0 or 1.

The yottabyte, or 1024 bytes, is the largest commonly used unit for determining the size of data.

Q2. How big a Quintillion measure is?

Ans - A quintillion is a very large number, equivalent to 10^18 bytes. It is used to represent huge quantities.

Q3. Give the examples of a smallest and the largest entities of data.

Ans -

Examples of data's smallest units -

- A single bit that can either represent a value of 0 or 1 in a computer's memory.
- A single pixel that corresponds to a certain colour and brightness level in a digital image.
- Any one letter or integer that appears in a text document as a single unit.

Largest data entities examples -

- The vast amount of digital stuff on the internet, including billions of web pages, pictures, movies, and other documents.
- Petabytes of data can be found in large scientific databases, such as those produced by telescopes or particle accelerators.
- The more than 3 billion base pair human genome, which is made up of DNA.
- Massive databases with millions or billions of records that are utilised by enterprises or governmental institutions.

Q4. Give FIVE parameters with which data can be categorized as

i) simple,
ii) Moderately complex and
iii) complex?

Ans - Five parameters that can be used to categorize data as simple, moderately complex, or complex are as follows -

1. Volume: The volume of data refers to the size of the data set, which can range from small to very large. Simple data sets typically have a low volume, while complex data sets have a high volume.
2. Structure: The structure of the data refers to how the data is organized or formatted. Simple data sets are often unstructured or have a simple structure, while complex data sets have a highly structured format with many interrelated tables or entities.
3. Relationships: The relationships within the data refer to how the data entities are related to each other. Simple data sets usually have few or no relationships between entities, while complex data sets have many interrelated entities with complex relationships.
4. Variability: Variability refers to the extent to which the data values are distributed across the dataset. Simple data sets usually have low variability with a narrow range of values, while complex data sets have high variability with a wide range of values.
5. Complexity of Analysis: The complexity of analysis refers to the level of expertise required to analyze the data. Simple data sets can usually be analyzed with basic statistical or analytical methods, while complex data sets require more advanced methods, such as machine learning or artificial intelligence.

Q5. What type of data are involved in the following applications?

1) Weather forecasting
2) Mobile usage of all customers of a service provider
3) Anomaly (e.g. fraud) detection in a bank organization
4) Person categorization, that is, identifying a human
5) Air traffic control in an airport
6) Streaming data from all flying aircrafts of Boeing

Ans -

1. Weather forecasting: Processing a significant amount of data gathered from numerous sources, including weather stations, satellites, and radar systems, is required for weather forecasting. Among other things, this information includes information on the temperature, humidity, wind speed, and precipitation.

2. Mobile usage of all customers of a service provider: Mobile usage data of all customers of a service provider includes information such as call logs, text messages,

data usage, location data, and other related metrics. This data is collected and processed to analyze customer behavior, identify trends, and improve services.

3. Finding anomalies (such as fraud) in a financial organisation: Processing financial transaction data to find odd or suspicious activity that could be a sign of fraud is the process of anomaly detection in a bank. This information consists of transaction summaries, account standings, IP addresses, timestamps, and other pertinent metrics. To examine this data and find trends that point to fraud, machine learning and statistical methods are frequently used.

4. Person categorization, that is, identifying a human: Person categorization, also known as human identification, involves processing data related to the visual characteristics of humans, such as facial features, body structure, and movements. This data includes images or videos that are captured using cameras or other sensors.

5. Air traffic control in an airport: Air traffic control involves processing data related to the movement of aircraft, such as their position, altitude, speed, and destination. This data is typically collected from radar systems, GPS, and other navigation tools. It also includes information about the flight plan, communication between pilots and air traffic controllers, and weather conditions.

6. Streaming data from all flying aircrafts of Boeing: Processing real-time data pertaining to the performance and well-being of every flying Boeing aircraft entails streaming data from all of the company's flying aircraft. These parameters, which are sent in real-time from the aircraft to the ground systems, include altitude, speed, engine temperature, fuel usage, and other relevant metrics.

PPT 2 - DATA CATEGORIZATION

Q1. Consider an image as an entity.

- What are the attributes you should think to represent an image?
- Categorize each attribute according to the NOIR data classification.
- Suppose, two images are given. Give an idea to check if two images are identical or not.

Ans - The following attributes can be used to represent an image:

1. Pixel values - The values of individual pixels that make up the image.
2. Image size - The dimensions of the image, such as height and width in pixels.
3. Color space - The color space used to represent the image, such as RGB or CMYK.
4. Image format - The file format used to store the image, such as JPEG or PNG.
5. Compression type - The type of compression used to reduce the size of the image file, such as lossy or lossless compression.
6. Metadata - Information about the image, such as the date and time it was taken, location, camera model, and exposure settings.

The categorization of each attribute according to the NOIR data classification is shown below:

- Pixel values: Ratio data
- Image size: Interval data
- Color space: Nominal data
- Image format: Nominal data
- Compression type: Nominal data
- Metadata: Nominal/Ordinal data (depending on the type of metadata, some could be nominal and some could be ordinal)

To check if two images are identical or not, we can follow the below approach -

1. Check if the size of both images is the same or not. If the size is not the same, the images are not identical.
2. If the size is the same, then we can compare the pixel values of both images. For this, you can iterate through all the pixels in both images and compare the pixel values at each location. If the pixel values at each location are the same, the images are identical; otherwise, they are not identical.
3. If the images have different formats, we may need to convert them to a common format before comparing pixel values.

Q2. How you can convert a data of interval type to ordinal type?

Give an example. What are the issues of such transformation? Whether the reverse is possible or not? Justify you answer.

Ans - In order to convert data of interval type to ordinal type, one approach is to use a technique called discretization. This involves dividing the range of values into intervals of equal size and assigning each interval a unique ordinal value.

For instance, let's consider a dataset of temperatures measured in degrees Celsius, ranging from -10 to 40. To convert this data to ordinal type, we can divide the temperature range into intervals of 10 degrees each: [-10, 0), [0, 10), [10, 20), [20, 30), and [30, 40). We can then assign an ordinal value to each interval, such as 1 for the first interval, 2 for the second interval, and so on, up to 5 for the last interval.

After discretization, the original interval data is transformed into ordinal data, which represents the rank order of the data values within each interval. However, it's important to note that the ordinal values do not represent the exact values within each interval, but only their order or rank.

Discretizing interval data into ordinal data can have several issues. One issue is that the choice of interval size can significantly impact the results. If the intervals are too large, important details may be lost, while if they are too small, the resulting ordinal data may not be meaningful or informative.

Another issue is that discretization can introduce bias or distortion in the data. For instance, if the original data is skewed, dividing it into equal intervals may not be appropriate. Moreover, discretization can result in the loss of information since the exact values within each interval are not preserved.

It's generally not possible to reverse the transformation from ordinal to interval data since the ordinal values do not provide sufficient information to reconstruct the original data values. This is because the original data values are continuous, while the ordinal values are discrete and represent only the rank order of the data values within each interval. Therefore, while the ordinal data can be useful for certain analyses, it's generally not advisable to use it to reconstruct the original interval data.

Q3) What are the different properties used to categorize the data according to NOIR data categorization?

Ans - NOIR data is categorized based on four properties: Nominal, Ordinal, Interval, and Ratio.

Nominal Data: This type of data is characterized by categorical variables that have no inherent order or ranking. Nominal data is used to label variables, such as gender, race, or types of cars. Nominal data can be analyzed using frequency tables and chi-square tests.

Ordinal Data: Ordinal data is characterized by categorical variables that have an inherent order or ranking. The ranking reflects the relative magnitude of the categories but does not indicate the exact amount of difference between them. Examples of ordinal data include educational levels, income levels, and satisfaction ratings. Ordinal data can be analyzed using ranking methods, such as Spearman's rank correlation coefficient.

Interval Data: Interval data is characterized by numerical variables where the differences between values are meaningful and consistent, but there is no true zero point. This means that values can be added, subtracted, and averaged, but ratios cannot be calculated. Examples of interval data include temperature measured in Celsius or Fahrenheit and dates measured in years. Interval data can be analyzed using descriptive statistics and inferential statistics, such as t-tests and ANOVA.

Ratio Data: Ratio data is characterized by numerical variables where the differences between values are meaningful, consistent, and there is a true zero point. This means that ratios can be calculated, and meaningful comparisons can be made between values. Examples of ratio data include height, weight, and income. Ratio data can be analyzed using descriptive statistics, inferential statistics, and regression analysis.

Q4 - Given an entity say “STUDENT” with the following attributes. Identify the NOIR category to which each of them belongs.

Scholarship Amount	Name	Roll No.	DoB	Aadhar No.	Gender	Mobile No.	Email id

Ans - The NOIR category to which each of the attributes belongs is as follows:

Scholarship Amount	Name	Roll No.	DoB	Aadhar No.	Gender	Mobile No.	Email id
Ratio data	Nominal data	Nominal data	Interval data	Nominal data	Nominal data	Nominal data	Nominal data

Nominal: name, Aadhaar no, gender, mobile no, email id (these are categories or labels)
Ordinal: None of the attributes are ordinal as there is no specific order or ranking among them.
Interval: DoB (there is a consistent distance between values, but there is no true zero point.
Ratio: scholarship amount, roll no (there is a specific order, a consistent distance between values, and a true zero point)

Q5. Give the concept of data cube to represent hyper-dimensional data? Also, explain with suitable diagrams the following.

- Roll up
- Drill down
- Slice

Ans - Data cube is a multidimensional representation of data that allows for analysis of data from multiple perspectives. A data cube has one or more dimensions or attributes (such as time, geography, and product), and each dimension can have multiple levels or categories. The intersection of these dimensions forms a cell, which contains the data values for the specific combination of attribute values. Here is an example of a data cube for sales data, with three dimensions: time, product, and geography -

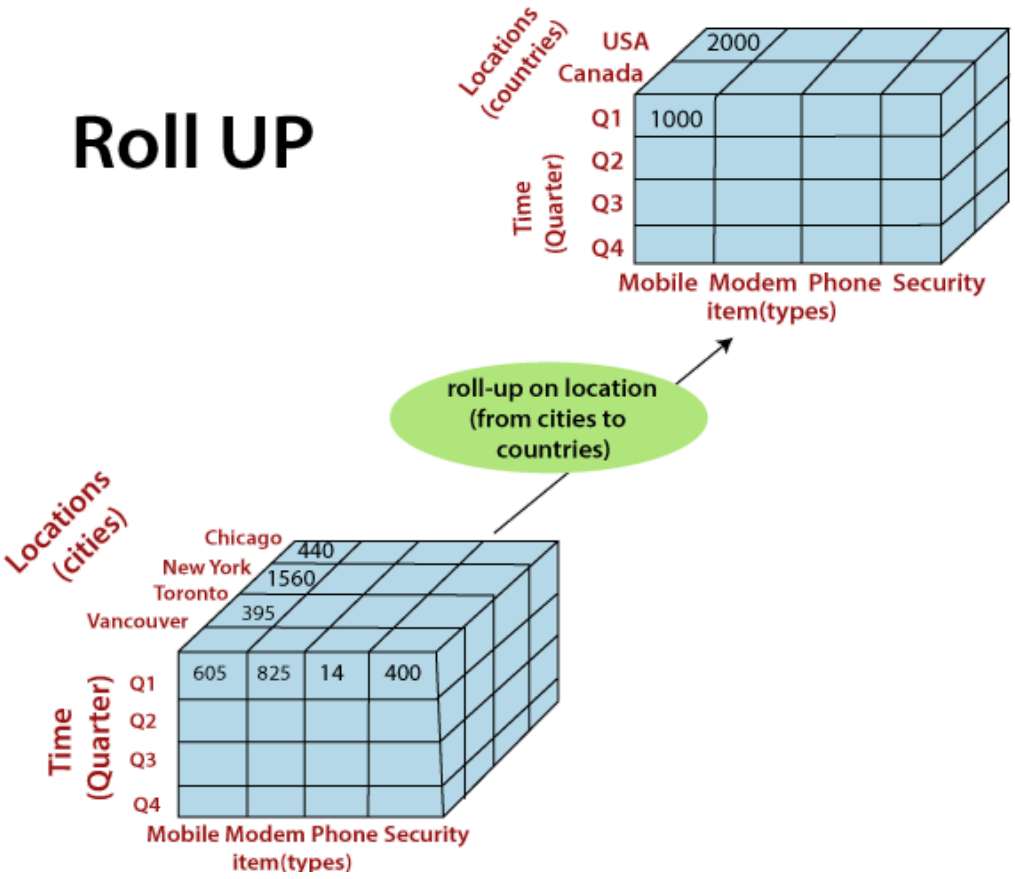
| Sales |
Time Product | USA | Canada | Mexico | Total |

Q1 2022 Product A | 100 | 50 | 75 | 225 |
Product B | 150 | 75 | 50 | 275 |
Total | 250 | 125 | 125 | 500 |

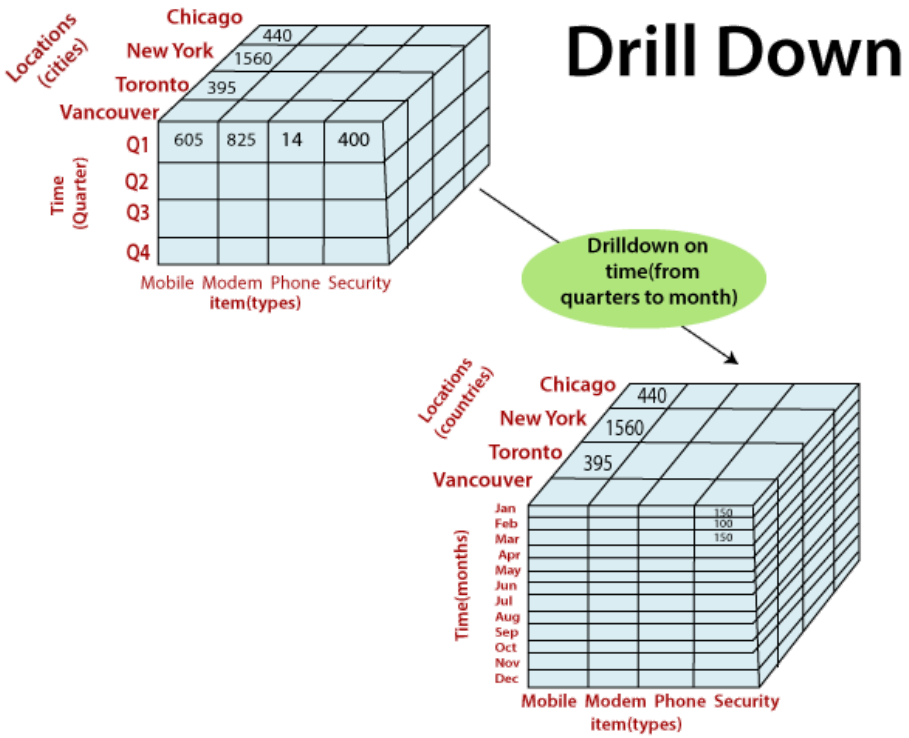
Q2 2022 Product A | 125 | 60 | 80 | 265 |

Product B | 100 | 80 | 70 | 250 |
Total | 225 | 140 | 150 | 515 |

- Roll up: Roll up is the process of summarizing or aggregating data from a lower-level to a higher-level in a data cube. For example, we can roll up the above data cube from product level to the total level, by summing up the sales values across all products for each time period and geography, as shown in the "Total" row of the table.

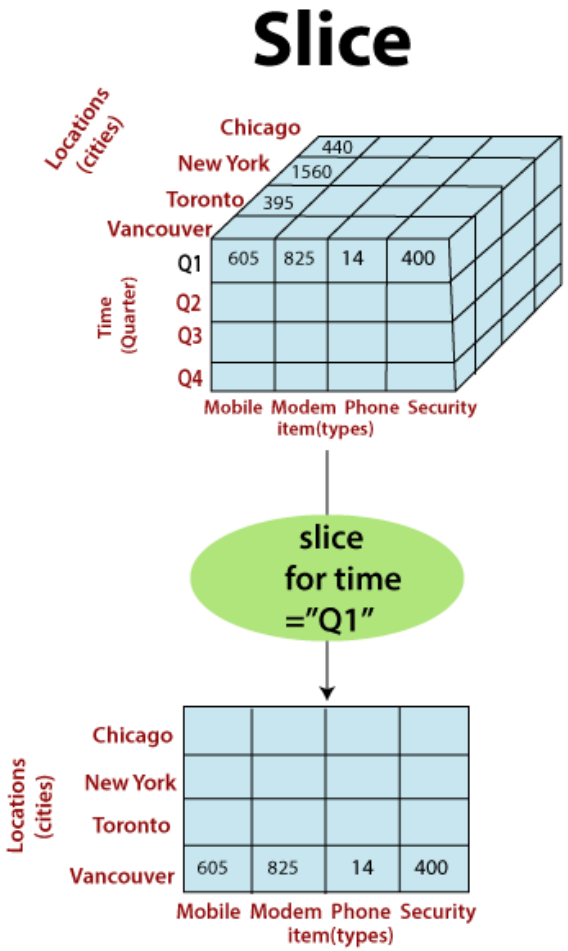


- Drill down: Drill down is the process of breaking down data from a higher-level to a lower-level in a data cube. For example, we can drill down the above data cube from time level to the product level, by showing the sales values for each product in each time period and geography. This would result in a more detailed table with more rows and columns.



- Slice is a subset of the cubes corresponding to a single value for one or more members of the dimension. For example, a slice operation is executed when the customer wants a selection on one dimension of a three-dimensional cube resulting in a two-dimensional site. So, the Slice operations perform a selection on one dimension of the given cube, thus resulting in a subcube.

For example, the following diagram illustrates how Slice works.



Here Slice is functioning for the dimensions "time" using the criterion time = "Q1".

Q6. Using the concept of data cube, how YouTube can archive videos of all type?

Ans - A data cube is a multidimensional database that allows data to be analyzed from different perspectives. In the case of YouTube, a data cube can be used to store and archive videos of all types.

The data cube can have multiple dimensions such as the date the video was uploaded, the category of the video, the language used in the video, the location where the video was filmed, and so on. These dimensions allow for a wide range of queries to be run on the data, such as finding all videos uploaded on a certain date or all videos in a specific category.

To store the videos themselves, YouTube can use a combination of file systems and cloud storage. The videos can be compressed and stored in a file system such as Hadoop Distributed File System (HDFS) or Amazon S3. These systems allow for scalable and reliable storage of large amounts of data.

To access and analyze the data cube, YouTube can use a variety of tools and technologies such as Apache Hive, Apache Spark, or Google BigQuery. These tools allow for querying and processing of large amounts of data in a distributed and parallel manner. Overall, by utilizing a data cube and scalable storage technologies, YouTube can effectively archive videos of all types while also enabling easy access and analysis of the data.

Q7. Give FOUR differences between data of types “interval” and “ratio-scale”

Ans - Four differences between data of types “interval” and “ratio-scale” are:

Interval Scale	Ratio Scale
Interval scale data has equal intervals between values, meaning that the difference between any two adjacent values is the same.	Ratio scale data is equally spaced with a true zero, meaning that the difference between any two adjacent values is the same, and there is a true zero point.
Interval scale data can be added and subtracted but cannot be multiplied or divided.	Ratio scale data can be added, subtracted, multiplied, and divided, allowing for more complex mathematical operations.
Interval scale data does not have an absolute zero point, meaning that zero does not indicate the complete absence of the measured quantity.	Ratio scale data has an absolute zero point, where zero indicates the complete absence of the measured quantity.
Interval scale data can have negative values.	Ratio scale data cannot have negative values.

PPT 3 - DESCRIPTIVE STATISTICS

Q1. Which of the following central tendency measurements allows distributive, algebraic and holistic measure?

- Mean
- Median
- Mode

Which measure may be faster than other? Why?

Ans - The only central tendency measurement that allows distributive, algebraic, and holistic measures is the mean.

The mean is distributive because it can be broken down into smaller parts and then added together. For example, the mean of a set of numbers can be calculated by adding all the numbers together and then dividing by the total number of numbers.
The mean is algebraic because it can be manipulated using mathematical operations. For example, if you add a constant value to every number in the set, the mean will also increase by the same constant value.
Finally, the mean is holistic because it takes into account all the values in the data set equally. Every value contributes to the mean, regardless of how many times it occurs or where it falls in the data set.

The mode may be the fastest to calculate since it only involves finding the most frequent value in the data set. In contrast, calculating the mean and median can be more time-consuming, particularly with large data sets or outliers. Mean requires adding all values and dividing by the total number, while the median involves ordering and identifying the middle value(s).

Q2) Give three situations where AM, GM, and HM are the right measures of central tendency?

Ans: Three situations where AM is the right measure of central tendency are:

- In situations where the data set follows a normal distribution, such as the heights of a population, the arithmetic mean is an appropriate measure of central tendency.
- The arithmetic mean is a suitable measure of central tendency when the data set contains values that are evenly distributed around a central value, such as grades on an exam.

- When the data set does not contain too many extreme outliers, such as ages of a group of people, the arithmetic mean provides a reliable measure of central tendency.

Three situations where GM is the right measure of central tendency are:

- The geometric mean is a useful measure of central tendency when the data set contains values that are related to each other multiplicatively, such as compound interest rates or population growth rates.
- In situations where the data set contains values that are logarithmically transformed and follow a normal distribution, such as stock prices, the geometric mean is an appropriate measure of central tendency.
- When the data set contains values that are skewed by extreme outliers, such as income levels in a country with a few extremely wealthy individuals, the geometric mean is a suitable alternative to the arithmetic mean.

Three situations where HM is the right measure of central tendency are:

- The harmonic mean is an appropriate measure of central tendency in situations where the data set contains values that are inversely proportional, such as speed and time or price and quantity.
- When the data set contains values that are skewed by extreme outliers, such as average income levels in a country with a few extremely wealthy individuals, the harmonic mean can provide a more robust measure of central tendency compared to the arithmetic mean.
- In situations where the data set contains values that are rates or ratios, such as miles per gallon or earnings per share in the stock market, the harmonic mean is a suitable measure of central tendency.

Q3) Given a sample of data, how to decide whether it is
a)Symmetric?
b)Skew-symmetric (positive or negative)?
c)Uniformly increasing (or decreasing)?
d)In-variate?

Ans: The following methods can be applied ways to decide the shape of the given sample data:

(a)Symmetric:

- The histogram should have a bell-shaped curve with the left half being a mirror image of the right half.
- The median of the boxplot should be in the center of the box with the whiskers being approximately equal in length.
- The skewness coefficient should be close to zero.
- The normal probability plot should show points approximately following a straight line.
- The run sequence plot should not show any increasing or decreasing pattern.

(b)Skew-symmetric (positive):

- The histogram should have a long right tail with the left half being shorter than the right half.
- The median of the boxplot should be shifted towards the lower end of the box with the lower whisker being longer than the upper whisker.
- The skewness coefficient should be positive.
- The normal probability plot should show points curving upwards.
- The run sequence plot should show an increasing pattern.

Skew-symmetric (negative):

- The histogram should have a long left tail with the right half being shorter than the left half.
- The median of the boxplot should be shifted towards the upper end of the box with the upper whisker being longer than the lower whisker.
- The skewness coefficient should be negative.
- The normal probability plot should show points curving downwards.
- The run sequence plot should show a decreasing pattern.

(c) Uniformly increasing:

- The histogram should show an increasing pattern.
- The median of the boxplot should not be in the center of the box with the lower whisker being longer than the upper whisker.
- There is no skewness coefficient for this type of data.
- There is no clear interpretation for the normal probability plot.
- The run sequence plot should show a clear increasing pattern.

Uniformly decreasing:

- The histogram should show a decreasing pattern.
- The median of the boxplot should not be in the center of the box with the upper whisker being longer than the lower whisker.
- There is no skewness coefficient for this type of data.

- There is no clear interpretation for the normal probability plot.
- The run sequence plot should show a clear decreasing pattern.

(d) In-variate:

- The histogram should not show any clear pattern.
- The median of the boxplot should not be in the center of the box with the whiskers being approximately equal in length.
- There is no skewness coefficient for this type of data.
- There is no clear interpretation for the normal probability plot.
- The run sequence plot should show points approximately evenly distributed around a horizontal line.

Q4) How the box-plots will look for the following types of samples?

(a) Symmetric
(b) Positively skew-symmetric
(c) Negatively skew-symmetric
(d) In-variate

Ans:

a) For a symmetric sample, the box plot will show a distribution that is evenly balanced around the median. The median line will be in the middle of the box, and the whiskers will extend to approximately the same length on both sides. The box will have similar length on both sides of the median, indicating that the data is equally dispersed.

b) For a positively skew-symmetric sample, the box plot will show a distribution that is skewed to the right, indicating that there are more values on the left side of the median. The median line will be closer to the bottom of the box, and the whisker on the right side will be longer than the one on the left. This suggests that the data is more spread out on the right side.

c) For a negatively skew-symmetric sample, the box plot will show a distribution that is skewed to the left, indicating that there are more values on the right side of the median. The median line will be closer to the top of the box, and the whisker on the left side will be longer than the one on the right. This suggests that the data is more spread out on the left side.

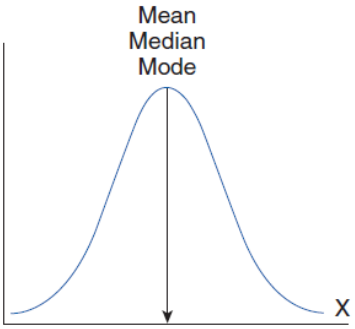
d) For an in-variate sample, the box plot will show a distribution that is evenly balanced around the median. The median line will be in the middle of the box, and the whiskers will extend to approximately the same length on both sides. The box will be of similar length

on both sides of the median, indicating that the data is equally dispersed. This suggests that the data is uniform and there is no skewness present in the sample.

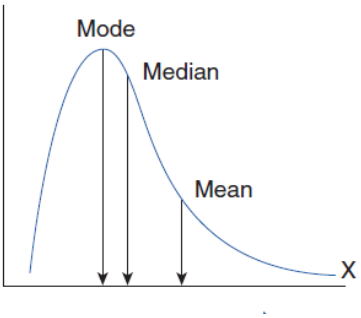
Q5) Draw the curves for the following types of distributions and clearly mark the likely locations of mean, median and mode in each of them. (a) Symmetric (b) Positively skew symmetric (c) Negatively skew symmetric

Ans: The curves for the given distribution types are shown :

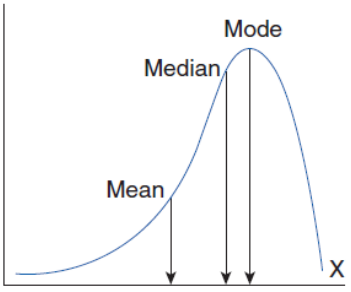
(a) Symmetric:



(b) Positively skew symmetric:



(c) Negatively skew symmetric:



Q6) The variance σ^2 of a sample X = {x1, x2, x3, ..., xn} of n data is defined as follows: σ^2 =1/(n-1) $\sum_{i=1}^{n}$ (xi - $\bar{x}$)^2, where $\bar{x}$ denotes the mean of the sample. Why (n-1) is in the denominator instead of n ?

Ans: When calculating the variance of a sample, we divide by (n-1) instead of n to account for the fact that we used the sample mean in our calculation. This is because we have one less degree of freedom, or piece of independent information, to estimate the population variance once we calculate the sample mean. Using (n-1) in the denominator gives an unbiased estimate of the population variance, which is important for making valid statistical inferences.

Q7) What are the degree of freedoms in each of the following cases?

a. A sample with a single data.
b. A sample with n data.
c. A sample of tabular data with n rows and m columns.

Ans: For a sample with a single data point, there are no degrees of freedom because there is only one value, and it cannot vary.

- For a sample with n data points, the degrees of freedom are n-1. This is because once n-1 data points are known, the value of the nth point is determined by the sum or mean of the previous n-1 data points.
- For a sample of tabular data with n rows and m columns, the degrees of freedom are (n-1)x(m-1). This is because once the means of each row and column are known, the value of each cell can be determined by the sum or mean of the other cells in the same row or column. Therefore, the degrees of freedom for each row and column are one less than the number of cells in that row or column. To get the total degrees of freedom, we need to subtract the degrees of freedom for each row and column (n-1+m-1) from the total number of cells (n x m), which gives us (n-1)x(m-1).

Q8) Calculate the Coefficient of variation (CV) for the following sample data.

(a) 10, -5, 20, 15, -5, 25, 30, 35, -25, 25

(b)

x	10	20	30	40	50
f(x)	0.2	0.4	0.1	0.2	0.1

Ans: (a) Mean (μ) = (10 - 5 + 20 + 15 - 5 + 25 + 30 + 35 - 25 + 25) / 10 = 12.5

Standard deviation (σ) = sqrt(((10 - 12.5)^2 + (-5 - 12.5)^2 + (20 - 12.5)^2 + (15 - 12.5)^2 + (-5 - 12.5)^2 + (25 - 12.5)^2 + (30 - 12.5)^2 + (35 - 12.5)^2 + (-25 - 12.5)^2 + (25 - 12.5)^2) / (10 - 1))
=> σ = sqrt(2950 / 9)
=> σ = 17.19

Coefficient of variation = (17.19 / 12.5) x 100% = 137.52%

(b) The mean (μ) for the given data is:

μ = Σ(x * f(x)) / Σf(x)

=> μ = (10*0.2 + 20*0.4 + 30*0.1 + 40*0.2 + 50*0.1) / (0.2 + 0.4 + 0.1 + 0.2 + 0.1)

=> μ = 23

The standard deviation (σ) is:

σ = sqrt[Σ(f(x) * (x - μ)^2) / Σf(x)]

=> σ = sqrt[(0.2*(10-23)^2 + 0.4*(20-23)^2 + 0.1*(30-23)^2 + 0.2*(40-23)^2 + 0.1*(50-23)^2) / (0.2 + 0.4 + 0.1 + 0.2 + 0.1)]

=> σ = 12.782

Therefore, the coefficient of variation (CV) is:

CV = (σ / μ) * 100%

=> CV = (12.782 / 23) * 100%

=> CV = 55.57%

PPT 4 - STATISTICAL METHODS

Q1) Give some examples of random variables? Also, tell the range of values and whether they are with continuous or discrete values.

Ans: Some examples of random variables are as follows:

I. The number of customers arriving at a store in a given hour. This is a discrete random variable, with possible values ranging from 0 to infinity.
II. The height of a randomly chosen person in a population. This is a continuous random variable, with possible values ranging from 0 to infinity.
III. The number of cars passing through a toll booth in a given hour. This is a discrete random variable, with possible values ranging from 0 to infinity.
IV. The amount of rainfall in a particular location during a given month. This is a continuous random variable, with possible values ranging from 0 to infinity.
V. The temperature in a given city on a given day. This is a continuous random variable, with possible values ranging from negative infinity to positive infinity.
VI. The number of heads obtained when flipping a coin 10 times. This is a discrete random variable, with possible values ranging from 0 to 10.

Q2) In the following cases, what are the probability distributions are likely to be followed. In each case, you should mention the random variable and the parameter(s) influencing the probability distribution function.

(a) In a retail source, how many counters should be opened at a given time period.
(b) Number of people who are suffering from cancers in a town?
(c) A missile will hit the enemy's aircraft.
(d) A student in the class will secure EX grade.
(e)Salary of a person in an enterprise.
(f)Accident made by cars in a city.
(g)People quit education after i) primary ii) secondary and iii) higher secondary educations.

Ans:

(a) The number of counters that should be opened at a given time period in a retail store would likely follow a discrete probability distribution because the number of counters is a whole number. The Poisson distribution is a possible probability distribution that could be used to model this scenario. The Poisson distribution

describes the number of events occurring in a fixed interval of time or space, given the average rate of occurrence of the event. In this case, the random variable would be the number of counters opened, and the parameter influencing the probability distribution function would be the average rate of customers arriving at the store per unit time.

(b) The number of people suffering from cancer in a town would also likely follow a discrete probability distribution because it is a whole number. The binomial distribution is a possible probability distribution that could be used to model this scenario. The binomial distribution describes the number of successes in a fixed number of independent trials, where each trial has the same probability of success. In this case, the random variable would be the number of people suffering from cancer, and the parameters influencing the probability distribution function would be the probability of a person having cancer in the town and the total population of the town.

(c) The probability distribution in this case would likely follow the binomial distribution since the event of hitting the enemy aircraft can only result in a success or a failure. The random variable in this case is the number of successes (hits) in a fixed number of trials (missile launches). The parameter(s) influencing the probability distribution function are the probability of a successful hit (p) and the number of trials (n).

(d) The probability distribution in this case would likely follow the Bernoulli distribution since the event of securing an EX grade can only result in a success or a failure. The random variable in this case is the binary outcome of success (EX grade) or failure (other grades). The parameter influencing the probability distribution function is the probability of a successful EX grade (p).

(e) The probability distribution in this case would likely follow a normal (Gaussian) distribution since salary is a continuous variable and tends to be distributed normally in large populations. The random variable in this case is the salary amount, and the parameters influencing the probability distribution function are the mean (μ) and standard deviation (σ) of the salaries in the enterprise.

(f) The probability distribution in this case would likely follow the Poisson distribution since accidents tend to occur randomly and independently over time, and the number of accidents in a fixed time period is countable. The random variable in this case is the number of accidents in a fixed time period, and the parameter influencing the probability distribution function is the average rate of accidents (λ) per unit time.

(g) The probability distribution in this case would likely follow a categorical (multinomial) distribution since the outcome can be classified into mutually exclusive categories (quit education after primary, secondary, or higher secondary education). The random variable in this case is the number of individuals in each category, and the parameters influencing the probability distribution function are the probabilities of quitting education after primary (p1), secondary (p2), and higher secondary (p3) educations.

Q3) How you can calculate the mean and standard deviation of a population if the population follows the following probability distribution functions with respect to an event.

(a) Binomial distribution function.
(b) Poisson's distribution function.
(c) Hypergeometric distribution function.
(d) Normal distribution function.
(e) Standard normal distribution function.

Ans: Here are the methods to calculate the mean and standard deviation for each of the mentioned probability distribution functions:

(a) Binomial Distribution Function:

For a binomial distribution function with parameters n (number of trials) and p (probability of success), the mean (μ) is given by: μ = n * p.
The standard deviation (σ) is given by: σ = sqrt(n * p * (1-p))

(b) Poisson's Distribution Function:

For a Poisson distribution function with parameter λ (mean number of events in a fixed interval), the mean (μ) and standard deviation (σ) are both given by: μ = σ = λ

(c) Hypergeometric Distribution Functionn:

For a hypergeometric distribution function with parameters N (total population size), (number of successes in the population), and n (sample size), the mean (μ) is given by: μ = n * K / N.
The standard deviation (σ) is given by:
σ = sqrt(n * K * (N-K) * (N-n) / (N^2 * (N-1)))

(d) Normal Distribution Function:

For a normal distribution function with parameters μ (mean) and σ (standard deviation), the mean (μ) and standard deviation (σ) are the same as the parameters of the distribution.

(e) Standard Normal Distribution Function:

For a standard normal distribution function (i.e., a normal distribution function with μ = 0 and σ = 1), the mean (μ) is 0, and the standard deviation (σ) is 1.

Q4) What are the degrees of freedom in the following cases.
Case 1: A single number.
Case 2: A list of n numbers.
Case 3: a table of data with m rows and n columns.
Case 4: a data cube with dimension m×n×p.

Ans: The degrees of freedom for the following cases are:
Case 1: A single number

- There are no degrees of freedom since there is only one value and no estimation is required.

Case 2: A list of n numbers

- There are n-1 degrees of freedom because one value can be freely chosen, but the remaining n-1 values must be chosen in a way that satisfies the constraint that the sum of the values is a fixed constant.

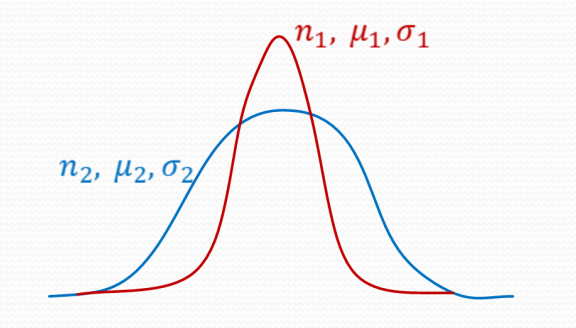
Case 3: A table of data with m rows and n columns

- The degrees of freedom depend on the statistical test being performed. For example, in a chi-squared test of independence between two categorical variables, the degrees of freedom are (m-1)(n-1), since the number of expected values can be estimated from the marginal totals of the rows and columns.

Case 4: A data cube with dimension m×n×p

- The degrees of freedom depend on the statistical test being performed. For example, in a three-way ANOVA with one factor having m levels, another factor having n levels, and a third factor having p levels, the degrees of freedom for the main effects are (m-1), (n-1), and (p-1), respectively, and the degrees of freedom for the interactions are (m-1)(n-1)(p-1).

Q5) In the following, two normal sampling distributions are shown with parameters n, μ and σ (all symbols bear their usual meanings).



What are the relations among the parameters in the two?

Ans: The image shows two normal sampling distributions with the same sample size n, that is, n1 = n2.
The means (μ1 and μ2) are equal, that is, μ1 = μ2.
But they have different standard deviations (σ1 and σ2), that is, σ1 < σ2.

Q6) Suppose, $\overline{X}$ and S denote the sample mean and standard deviation of a sample. Assume that population follows normal distribution with population mean μ and standard deviation σ. Write down the expression of z and t values with degree of freedom n.

Ans: The z-score is given by:
$z = (\overline{X} - \mu) / (\sigma / \sqrt{n})$
where $\overline{X}$ is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size.

The t-score is given by:
$t = (\overline{X} - \mu) / (S / \sqrt{n})$
where S is the sample standard deviation. The t-distribution has n-1 degrees of freedom.

PPT 5 - STATISTICAL LEARNING

Q1) In a hypothesis testing, suppose H0 is rejected. Does it mean that H1 is accepted? Justify your answer.

Ans -

No, rejecting the null hypothesis (H0) does not necessarily mean that the alternative hypothesis (H1) is accepted. The rejection of the null hypothesis simply means that there is enough evidence to suggest that the null hypothesis is not true and should be rejected. However, it does not prove that the alternative hypothesis is true. When conducting a hypothesis test, the null hypothesis is typically the default assumption or claim that is being tested, and the alternative hypothesis is the claim that we believe to be true if the null hypothesis is rejected. If the null hypothesis is rejected, it simply means that there is evidence against the null hypothesis and we should reject it. However, it is possible that the alternative hypothesis is not true, and we need further evidence to support it. Therefore, when the null hypothesis is rejected, we cannot automatically accept the alternative hypothesis. Instead, we may need to collect more data, perform additional tests, or use other methods to support the alternative hypothesis. Alternatively, we may need to revise the alternative hypothesis and formulate a new hypothesis that better fits the observed data.

Q2) Give the expressions for z, t and χ^2 in terms of population and sample parameters, whichever is applicable to each. Signifies these values in terms of the respective distributions.

Ans: The expressions for z, t, and χ^2 in terms of population and sample parameters are as follows:

1. Z-test: The z-test is a measure of how many standard deviations a data point is from the population mean.

The formula for the z-test is: $z = (x - \mu) / \sigma$ where x is the sample mean, μ is the population mean, and σ is the population standard deviation. In the case of a normal distribution, the z-test follows a standard normal distribution with mean 0 and standard deviation 1.

2. t-test: The t-test is similar to the z-test, but it is used when the population standard deviation is unknown and must be estimated from the sample.

The formula for the t-test is: t = (x - μ) / (s / sqrt(n)) where x is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size. In the case of a t-distribution, the t-test follows a t-distribution with n-1 degrees of freedom.

3. χ^2 (chi-squared): The chi-squared statistic is used to test the independence of two categorical variables or goodness of fit of a sample to a theoretical distribution.

The formula for χ^2 depends on the specific test being performed, but in general, it is calculated as: χ^2 = Σ ((O - E)^2 / E) where O is the observed frequency, E is the expected frequency, and the sum is taken over all categories. In the case of a chi-squared distribution, the χ^2 statistic follows a chi-squared distribution with degrees of freedom equal to the number of categories minus one.

Q3) How can you obtain the value say P (z = a)? What this value signifies?
Ans:
It is not possible to obtain the value of P (z = a) because the standard normal distribution, which is represented by the letter z, is a continuous probability distribution. This means that the probability of any specific point in the distribution is zero. However, it is possible to calculate the probability of an interval of values for z using the cumulative distribution function (CDF) of the standard normal distribution. The CDF gives the probability that a random variable fall below a certain value, and is denoted by Φ(z). Therefore, if we want to find the probability that a standard normal random variable is less than or equal to a particular value, we can use the CDF as follows: P (z ≤ a) = Φ(a) Alternatively, if we want to find the probability that a standard normal random variable is greater than a particular value, we can use the complement of the CDF as follows: P (z > a) = 1 - Φ(a) The value of P (z ≤ a) or P (z > a) can be obtained using standard statistical tables or statistical software. The value of P (z ≤ a) or P (z > a) signifies the probability of obtaining a z-score less than or equal to a particular value or greater than a particular value, respectively, under the standard normal distribution. In other words, it represents the likelihood of observing a certain outcome given the assumption that the population follows a standard normal distribution. The area under the standard normal distribution curve between two z-values represents the probability of observing a value within that range.

Q4) On what occasion, you should consider z-distribution but not t-distribution and vice-versa?

Ans:
The choice between using a z-distribution or a t-distribution depends on the sample size and whether the population standard deviation is known.

A z-distribution should be used when the sample size is large (typically $n > 30$) and the population standard deviation is known. In this case, we can use the z-distribution to calculate the probability of getting a sample mean that is a certain number of standard deviations away from the population mean. This is because the sample mean follows a normal distribution with a mean equal to the population mean and a standard deviation equal to the population standard deviation divided by the square root of the sample size.

On the other hand, a t-distribution should be used when the sample size is small (typically $n < 30$) or the population standard deviation is unknown. In this case, we cannot use the z-distribution because we do not know the population standard deviation. Instead, we use the t-distribution, which has a wider distribution than the z-distribution to account for the uncertainty in the sample standard deviation. The t-distribution has a larger variance than the z-distribution because the sample standard deviation tends to be more variable than the population standard deviation. As the sample size increases, the t-distribution becomes more like the z-distribution.

Q5) Give a situation when you should consider χ^2 distribution but neither z- nor t-distribution.

Ans:
The chi-square distribution should be considered when we need to test hypotheses about the variance or standard deviation of a population, and the population mean is unknown. This occurs in situations where we have a sample of independent and normally distributed random variables, and we need to estimate the population variance or test hypotheses about the population variance or standard deviation. One such situation could be testing the variability of scores in a test, where we are not interested in the mean score but only in the variance of scores. In this case, we can use the chi-square distribution to test hypotheses about the population variance.

Q. Plot all types of Probability distributions using Python.
Ans: A probability distribution is a mathematical function that describes the likelihood of different outcomes in a random event. The various types of probability distributions are:

(1) Discrete Probability Distributions: It is of the following types:
- Binomial Distribution
- Multinomial Distribution
- Poisson Distribution
- Hypergeometric Distribution

(2) Continuous Probability Distributions: It is of the following types:
- Normal Distribution
- Standard Normal Distribution
- Gamma Distribution
- Exponential Distribution
- Chi Square Distribution
- Lognormal Distribution
- Weibull Distribution

All the above distributions are plotted as shown:

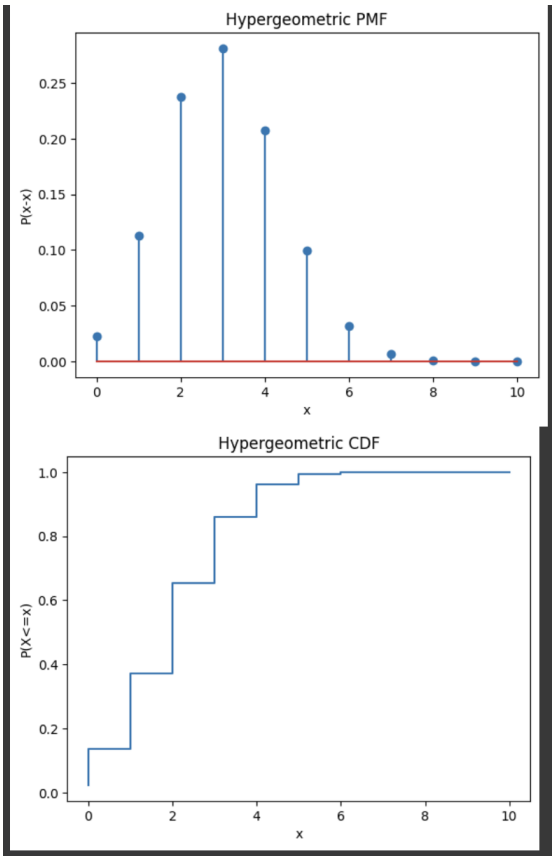
1. BINOMIAL DISTRIBUTION:

Code:

```
import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np

# Define the parameters of the hypergeometric distribution
N = 100 # population size
K = 30 #number of successes in the population
n = 10 # sample size

# Create a hypergeometric distribution object
hypergeom_dist = stats.hypergeom(N, K, n)

# Generate the probability mass function (PMF)
x = np.arange(max(0, n-(N-K)), min(n, K)+1) # possible values of the random
variable
pmf= hypergeom_dist.pmf(x) # probability mass function

# Plot the PMF
```

```
plt.stem(x, pmf)
plt.title('Hypergeometric PMF')
plt.xlabel('x')
plt.ylabel('P(x-x)')
plt.show()

# Generate the cumulative distribution function (CDF)
cdf = hypergeom_dist.cdf(x) # cumulative distribution function

# Plot the CDF
plt.step(x, cdf)
plt.title('Hypergeometric CDF')
plt.xlabel('x')
plt.ylabel('P(X<=x)')
plt.show()
```

Output:



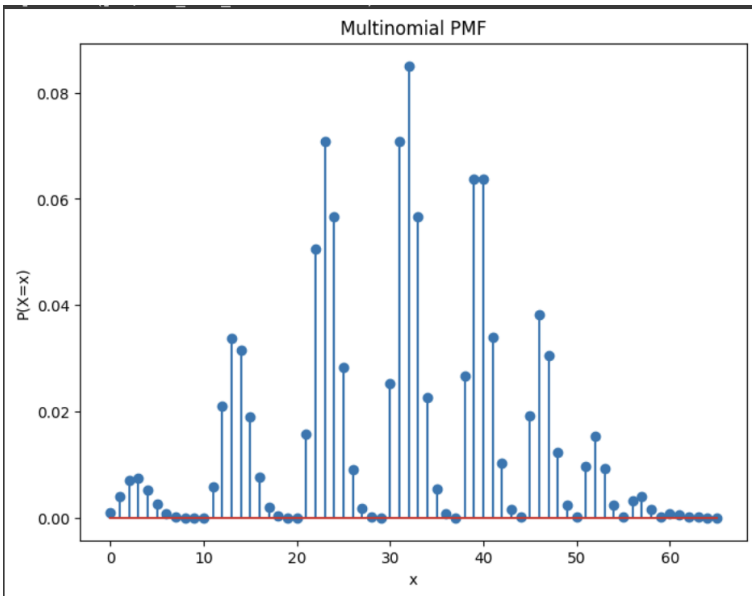
2. MULTINOMIAL DISTRIBUTION:

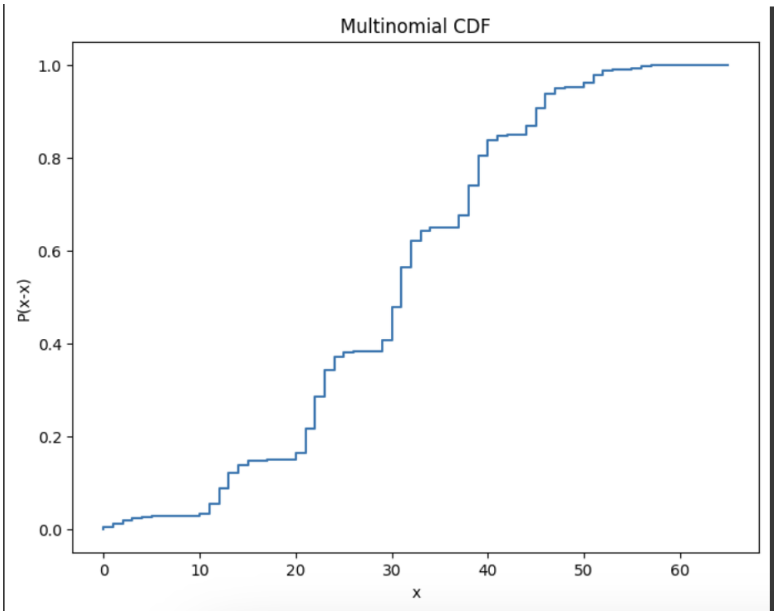
Code:

```
import numpy as np
from scipy.stats import multinomial
import matplotlib.pyplot as plt
```

```
# Define the parameters of the multinomial distribution
n = 10 # number of trials
pvals = [0.3, 0.2, 0.5] # probabilities of the outcomes

# Create a multinomial distribution object
multinom_dist = multinomial(n, pvals)

# Generate the probability mass function (PMF)
x = np.array([[i, j, n-i-j] for i in range(n+1) for j in range(n-i+1)]) #
possible values of the random variable
pmf = multinom_dist.pmf(x) # probability mass function

# Plot the PMF
plt.figure(figsize=(8, 6))
plt.stem(pmf, use_line_collection=True)
plt.title('Multinomial PMF')
plt.xlabel('x')
plt.ylabel('P(X=x)')
plt.show()

# Generate the cumulative distribution function (CDF)
cdf = np.cumsum(pmf) # cumulative distribution function

# Plot the CDF
plt.figure(figsize=(8, 6))
plt.step(np.arange(len(cdf)), cdf)
plt.title('Multinomial CDF')
plt.xlabel('x')
plt.ylabel('P(x-x)')
plt.show()
```

Output:





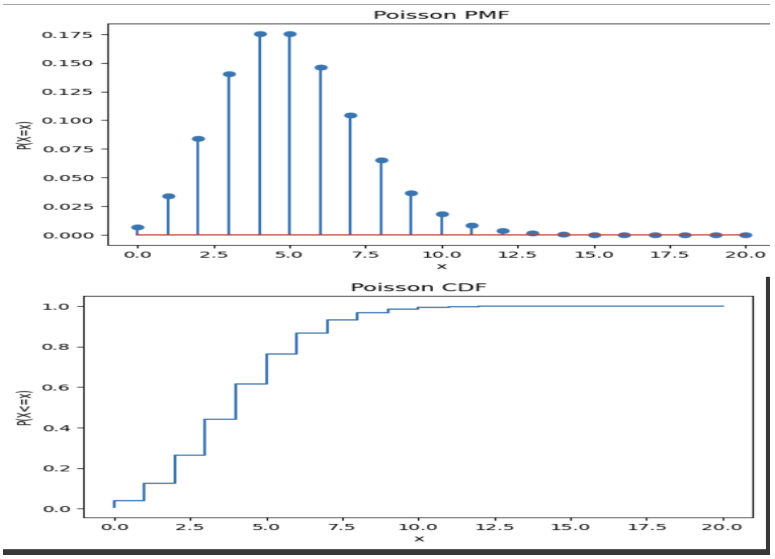
3. POISSON DISTRIBUTION:

Code:

```
import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np

# Define the parameter Lambda of the Poisson distribution
lam = 5

# Create a Poisson distribution object
poisson_dist = stats.poisson(mu=lam)

# Generate the probability mass function (PMF)
```

```
x = np.arange(0, 21) # possible values of the random variable
pmf = poisson_dist.pmf(x) # probability mass function

# Plot the PMF
plt.stem(x, pmf)
plt.title('Poisson PMF')
plt.xlabel('x')
plt.ylabel('P(X=x)')
plt.show()

# Generate the cumulative distribution function (CDF)
cdf = poisson_dist.cdf(x) # cumulative distribution function

# Plot the CDF
plt.step(x, cdf)
plt.title('Poisson CDF')
plt.xlabel('x')
plt.ylabel('P(X<=x)')
plt.show()
```

Output:



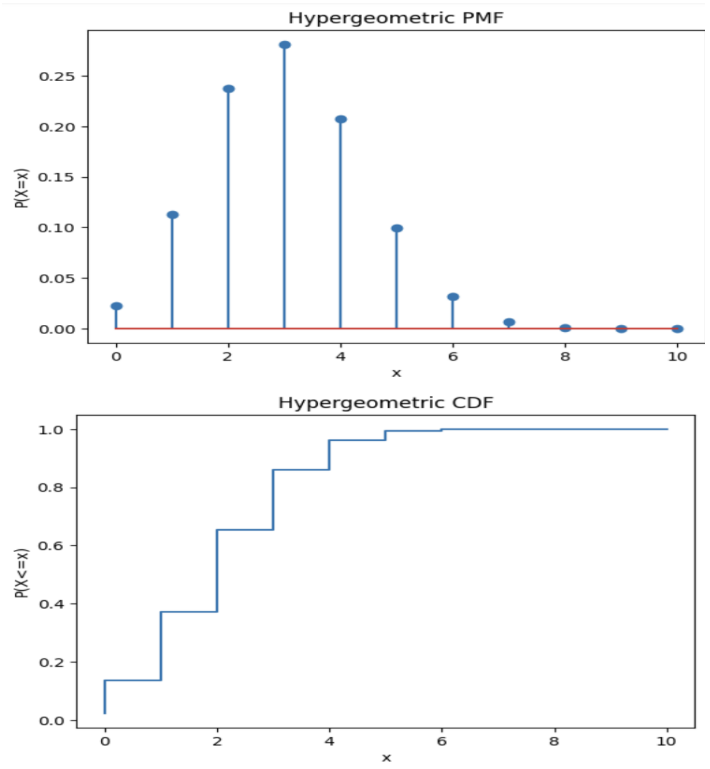
4. HYPERGEOMETRIC DISTRIBUTION:

Code:

```
import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np

# Define the parameters of the hypergeometric distribution
N = 100 # population size
K = 30 # number of successes in the population
```

```
n = 10 # sample size

# Create a hypergeometric distribution object
hypergeom_dist = stats.hypergeom(N, K, n)

# Generate the probability mass function (PMF)
x = np.arange(max(0, n-(N-K)), min(n, K)+1) # possible values of the random
variable
pmf = hypergeom_dist.pmf(x) # probability mass function

# Plot the PMF
plt.stem(x, pmf)
plt.title('Hypergeometric PMF')
plt.xlabel('x')
plt.ylabel('P(X=x)')
plt.show()

# Generate the cumulative distribution function (CDF)
cdf = hypergeom_dist.cdf(x) # cumulative distribution function

# Plot the CDF
plt.step(x, cdf)
plt.title('Hypergeometric CDF')
plt.xlabel('x')
plt.ylabel('P(X<=x)')
plt.show()
```

Output:



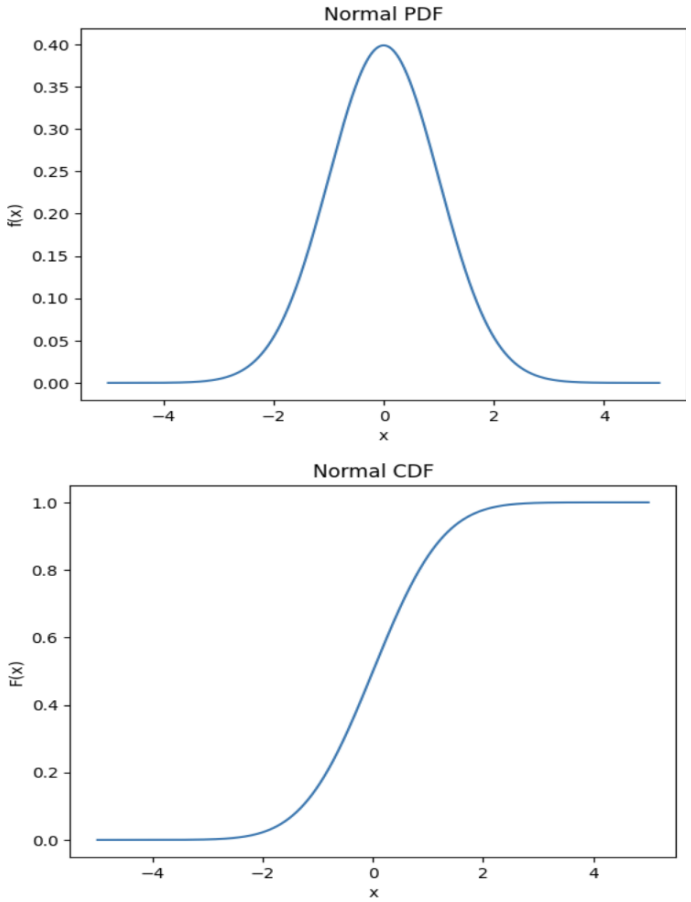
5. NORMAL DISTRIBUTION:

Code:

```
import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np

# Define the parameters of the normal distribution
mu = 0 # mean
sigma = 1 # standard deviation

# Create a normal distribution object
norm_dist = stats.norm(mu, sigma)

# Generate the probability density function (PDF)
x = np.linspace(-5, 5, num=1000) # possible values of the random variable
pdf = norm_dist.pdf(x) # probability density function

# Plot the PDF
plt.plot(x, pdf)
plt.title('Normal PDF')
plt.xlabel('x')
plt.ylabel('f(x)')
plt.show()
```

```
# Generate the cumulative distribution function (CDF)
cdf = norm_dist.cdf(x) # cumulative distribution function

# Plot the CDF
plt.plot(x, cdf)
plt.title('Normal CDF')
plt.xlabel('x')
plt.ylabel('F(x)')
plt.show()
```

Output:



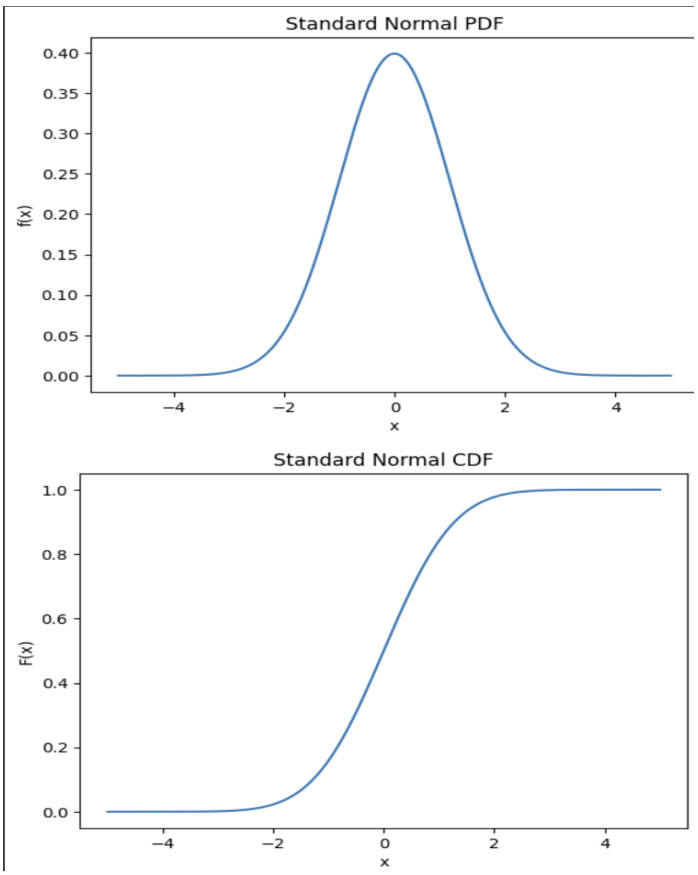
6. STANDARD NORMAL DISTRIBUTION:

Code:

```
import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np

# Create a standard normal distribution object
std_norm_dist = stats.norm()
```

```
# Generate the probability density function (PDF)
x = np.linspace(-5, 5, num=1000) # possible values of the random variable
pdf = std_norm_dist.pdf(x) # probability density function

# Plot the PDF
plt.plot(x, pdf)
plt.title('Standard Normal PDF')
plt.xlabel('x')
plt.ylabel('f(x)')
plt.show()

# Generate the cumulative distribution function (CDF)
cdf = std_norm_dist.cdf(x) # cumulative distribution function

# Plot the CDF
plt.plot(x, cdf)
plt.title('Standard Normal CDF')
plt.xlabel('x')
plt.ylabel('F(x)')
plt.show()
```

Output:

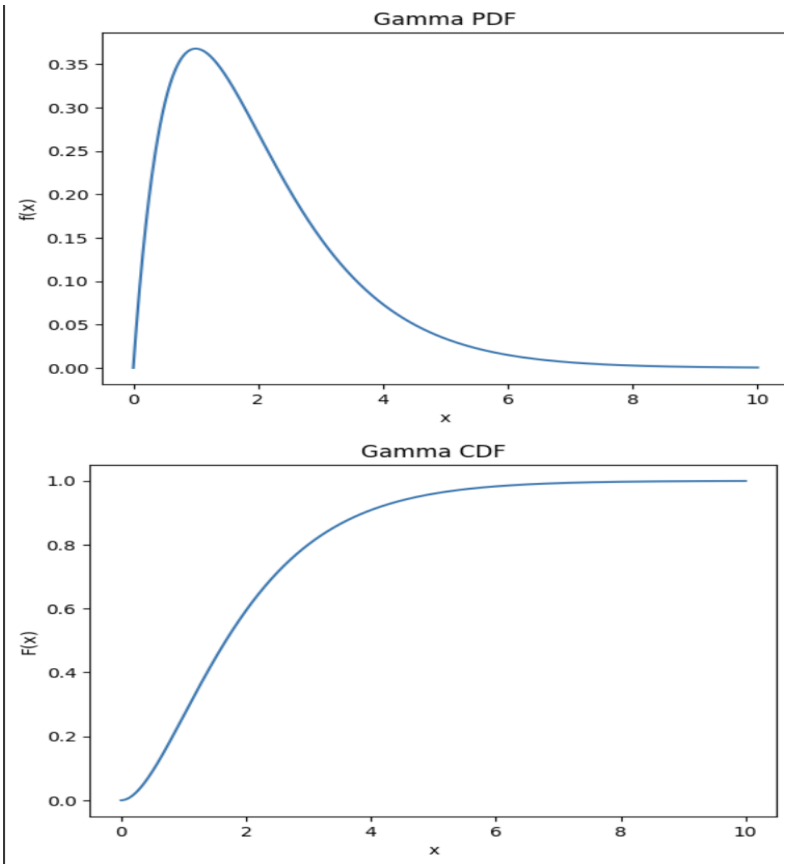


7. GAMMA DISTRIBUTION:

Code:

```
import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np

##Define the parameters of the gamma distribution
shape = 2 # shape parameter (alpha)
scale = 1 # scale parameter (beta)

#Create a gamma distribution object
gamma_dist = stats.gamma(a=shape, scale=scale)

#Generate the probability density function (PDF)
x = np.linspace(0, 10, num=1000) # possible values of the random variable
pdf = gamma_dist.pdf(x) # probability density function

#Plot the PDF
plt.plot(x, pdf)
plt.title('Gamma PDF')
plt.xlabel('x')
plt.ylabel('f(x)')
plt.show()

#Generate the cumulative distribution function (CDF)
cdf = gamma_dist.cdf(x) # cumulative distribution function

# Plot the CDF
plt.plot(x, cdf)
plt.title('Gamma CDF')
plt.ylabel("F(x)")
plt.xlabel('x')
plt.show()
```

Output:



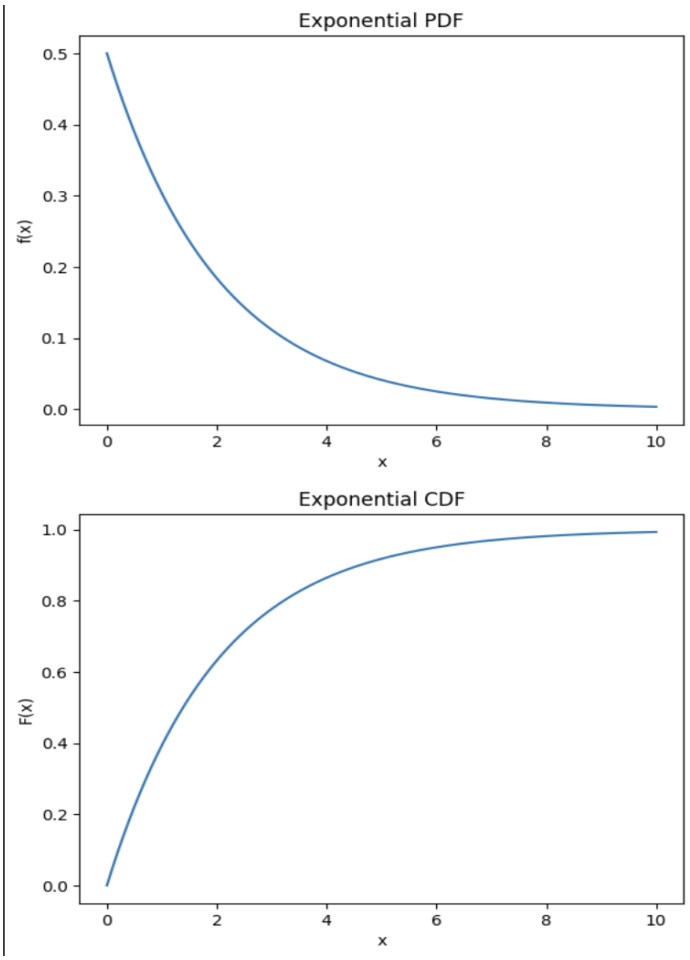
8. EXPONENTIAL DISTRIBUTION:

Code:

```
import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np

# Define the rate parameter of the exponential distribution
rate = 0.5

# Create an exponential distribution object
exp_dist = stats.expon(scale=1/rate)

# Generate the probability density function (PDF)
x = np.linspace(0, 10, num=1000) # possible values of the random variable
pdf = exp_dist.pdf(x) # probability density function

# Plot the PDF
plt.plot(x, pdf)
plt.title('Exponential PDF')
plt.xlabel('x')
plt.ylabel('f(x)')
plt.show()
```

```
# Generate the cumulative distribution function (CDF)
cdf = exp_dist.cdf(x) # cumulative distribution function

# Plot the CDF
plt.plot(x, cdf)
plt.title('Exponential CDF')
plt.xlabel("x")
plt.ylabel("F(x)")
plt.show()
```

Output:



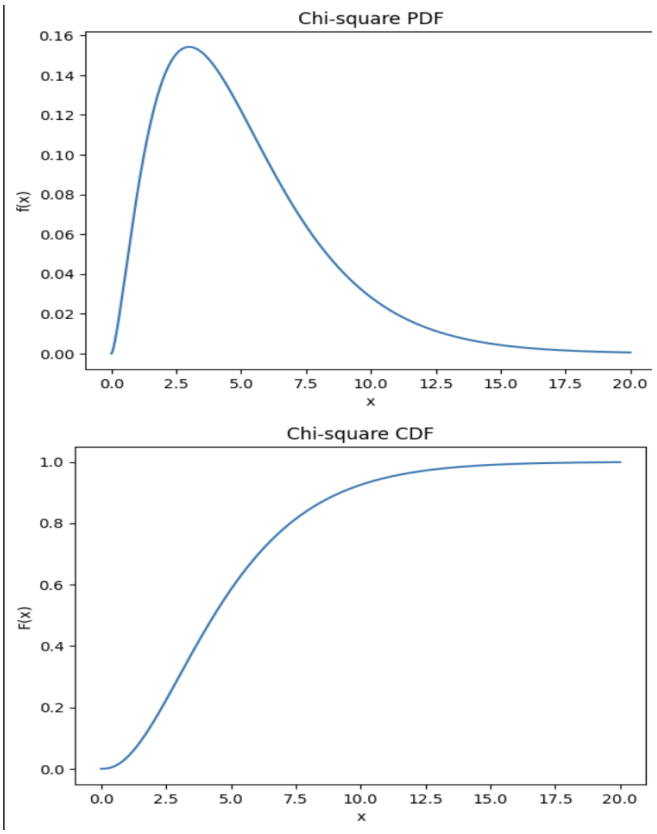
9. CHI-SQUARE DISTRIBUTION:

Code:

```
import scipy.stats as stats
```

```
import matplotlib.pyplot as plt
import numpy as np

# Define the degrees of freedom parameter of the chi-square distribution
dof = 5

# Create a chi-square distribution object
chi2_dist = stats.chi2(df=dof)

# Generate the probability density function (PDF)
x = np.linspace(0, 20, num=1000) # possible values of the random variable
pdf = chi2_dist.pdf(x) # probability density function

# Plot the PDF
plt.plot(x, pdf)
plt.title('Chi-square PDF')
plt.xlabel('x')
plt.ylabel('f(x)')
plt.show()

# Generate the cumulative distribution function (CDF)
cdf = chi2_dist.cdf(x) # cumulative distribution function

# Plot the CDF
plt.plot(x, cdf)
plt.title('Chi-square CDF')
plt.xlabel('x')
plt.ylabel("F(x)")
plt.show()
```

Output:



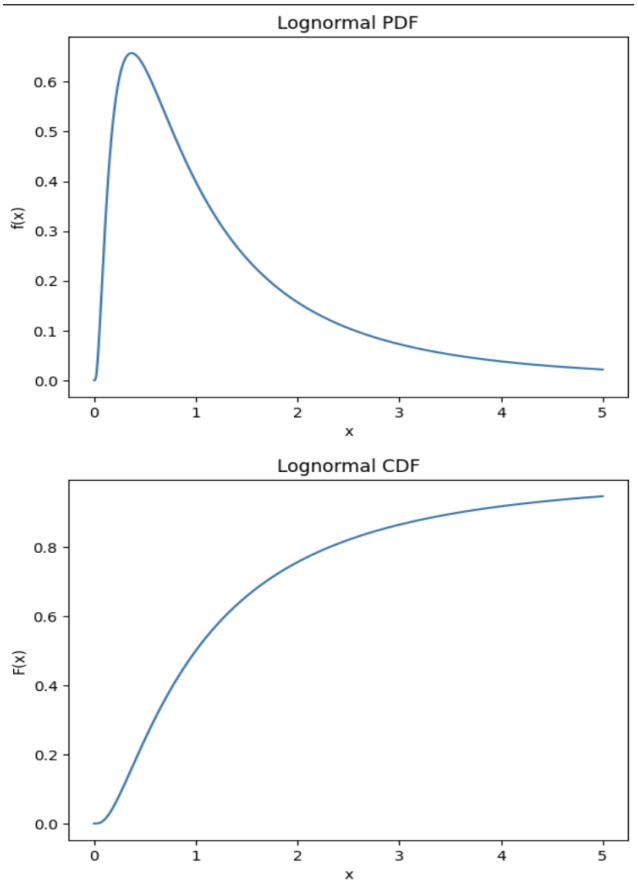
10. LOGNORMAL DISTRIBUTION:

Code:

```
import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np

# Define the parameters of the lognormal distribution
sigma = 1
mu = 0

# Create a lognormal distribution object
lognorm_dist = stats.lognorm(s=sigma, scale=np.exp(mu))

# Generate the probability density function (PDF)
x = np.linspace(0, 5, num=1000) # possible values of the random variable
pdf = lognorm_dist.pdf(x) # probability density function

# Plot the PDF
plt.plot(x, pdf)
plt.title('Lognormal PDF')
plt.xlabel('x')
plt.ylabel('f(x)')
plt.show()
```

```
# Generate the cumulative distribution function (CDF)
cdf = lognorm_dist.cdf(x) # cumulative distribution function

# Plot the CDF
plt.plot(x, cdf)
plt.ylabel('F(x)')
plt.title('Lognormal CDF')
plt.xlabel('x')
plt.show()
```

Output:



11. WEIBULL DISTRIBUTION:

Code:

```
import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np

# Define the shape and scale parameters of the Weibull distribution
shape = 2
scale = 3

# Create a Weibull distribution object
```

```
weibull_dist = stats.weibull_min(c=shape, scale=scale)

# Generate the probability density function (PDF)
x = np.linspace(0, 10, num=1000) # possible values of the random variable
pdf = weibull_dist.pdf(x) # probability density function

# Plot the PDF
plt.plot(x, pdf)
plt.title('Weibull PDF')
plt.xlabel('x')
plt.ylabel('f(x)')
plt.show()

# Generate the cumulative distribution function (CDF)
cdf = weibull_dist.cdf(x) # cumulative distribution function

# Plot the CDF
plt.title('Weibull CDF')
plt.plot(x, cdf)
plt.xlabel('x')
plt.ylabel('F(x)')
plt.show()
```

Output: