Day 4

10 August 2024 17:55

Feedback

- Reading from PDFs only
- Either the topic was taught in school OR you will learn this in future class!!!
- I am new to AI/Python - need more detailed sessions
- Daily bit of 10 min assignments will help, covering
- Some topics are rushed
-

Topic

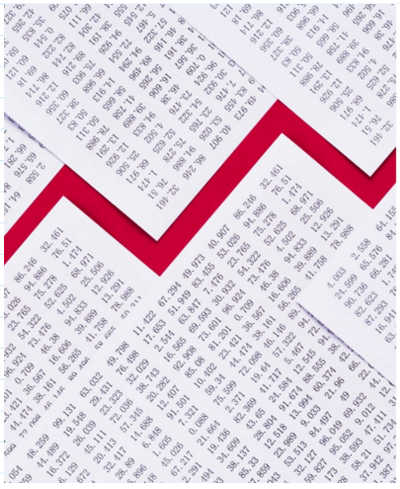
- What it is
- Need for it
- Example code
- Do's and Don'ts
- Use cases
  - Some use cases will be generic
  - Some may be related to ML/DL ** (wait)

Random variables

A variable whose value is unknown or a function that assigns values to each of an experiment's outcomes

Example: rolling a fair six-sided die. We can define a random variable XX to represent the outcome of the roll. The possible outcomes are the numbers 1 through 6, and the value of X will depend on which number is rolled.

For instance, let's designate the variable X to denote the total sum obtained from three dice rolls. In this scenario, X could potentially take on values such as 3 (if each die shows a 1), 18 (if each die shows a 6), or any integer between 3 and 18, as each die's outcome can range from 1 to 6.



Examples

- In the corporate world, random variables can be assigned to properties such as:
  - The average price of an asset over a given time period,
  - The return on investment after a specified number of years,
  - The estimated turnover rate at a company within the following six months, etc.
- Risk analysts
  - the probability of an adverse event occurring.

More examples

...

1. Number of Defective Items
2. temperature in a particular city at a given time
3. In financial markets, the price of a stock at a given point in time.
4. The amount of time a customer waits in line at a grocery store checkout.
5. Blood pressure readings for individuals in a population.
6. The grades obtained by students in a class.

Types of random variables

Discrete Random Variables

- variables take on a countable number of distinct values.
- Examples include the number of students in a class, the number of heads when flipping coins, or the number of defects in a batch of products.
- probability distribution of a discrete random variable is described by a probability mass function (PMF).

Continuous Random Variables

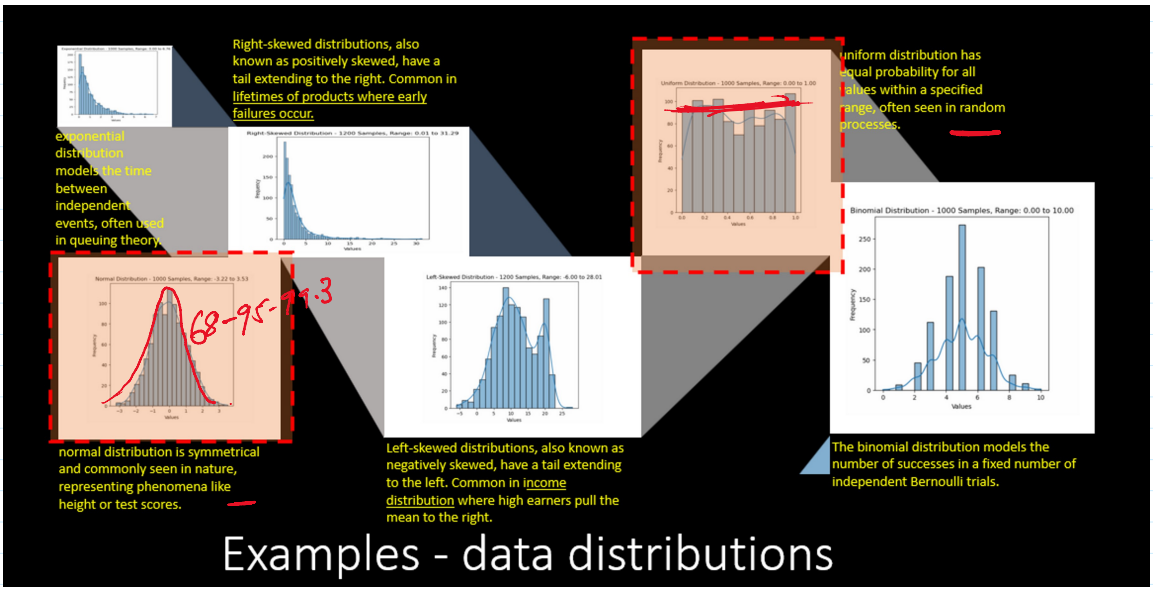
- variables can take on any value within a certain range.
- Examples include temperature, height, or time taken to complete a task.
- probability distribution of a continuous random variable is described by a probability density function (PDF).

What is data distribution?

In statistics, a data distribution refers to the way in which values or outcomes are spread or distributed across a dataset.

provides information about the possible values a variable can take and how frequently those values occur.

Understanding the distribution of data is fundamental for statistical analysis as it influences the choice of appropriate statistical methods and helps draw meaningful conclusions.



[1 2 11 100]

How is data distribution different from dispersion methods

Aspect	Data Distribution	Dispersion Methods
Definition	way values are spread or distributed across a dataset.	quantify the spread or variability of values within a dataset.
Focus	Concentrates on the frequency	Specifically addresses the extent to which data points deviate from the central tendency.
Key Aspects	Probability densities, Skewness, symmetry, kurtosis are essential characteristics.	Range, variance, standard deviation, interquartile range (IQR), etc. are key measures.
Example	A dataset could have a normal distribution, skewed to the right or left, or exhibit other patterns.	High standard deviation indicates greater variability in the dataset.
Scope	Concerned with the overall shape	Focuses on the variability of values.
Measurement	Often described qualitatively	Provides concrete numerical values for analysis.

Probability mass function PMF

A probability mass function (PMF) is a function that describes the probability distribution of a discrete random variable.

discrete random variables can only take on a countable number of distinct values.

The PMF assigns a probability to each possible value that the discrete random variable can take.

The sum of probabilities assigned by the PMF to all possible values must equal 1.

fair six-sided die

the PMF assigns a probability of 1/6 to each outcome, since each face has an equal chance of occurring when rolling a fair six-sided die.

$$P(X = 1) = \frac{1}{6}$$
$$P(X = 2) = \frac{1}{6}$$
$$P(X = 3) = \frac{1}{6}$$
$$P(X = 4) = \frac{1}{6}$$
$$P(X = 5) = \frac{1}{6}$$
$$P(X = 6) = \frac{1}{6}$$

It means ...

- Let X be a discrete random variable on a sample space S. Then the *probability mass function* $f(x)$ is defined as

$$f(x) = \mathrm{P}[X = x].$$

- Each probability mass function satisfies the following two conditions:

$$\begin{array}{ll} \text{(i)} & f(x) \geq 0 \text{ for all } x \in S, \\ \text{(ii)} & \sum_{x \in S} f(x) = 1. \end{array}$$

The PMF can be used to calculate the expected value (or mean) of X:

$E(X) = \sum_i x_i \cdot P(X = x_i)$

where:

- x_i are the possible values of the random variable,
- $P(X = x_i)$ is the probability mass function (PMF) evaluated at x_i,
- and the summation is taken over all possible values of X.

expected value

Example

1. The expected number of complaints per day $E[X]$ can be calculated as:

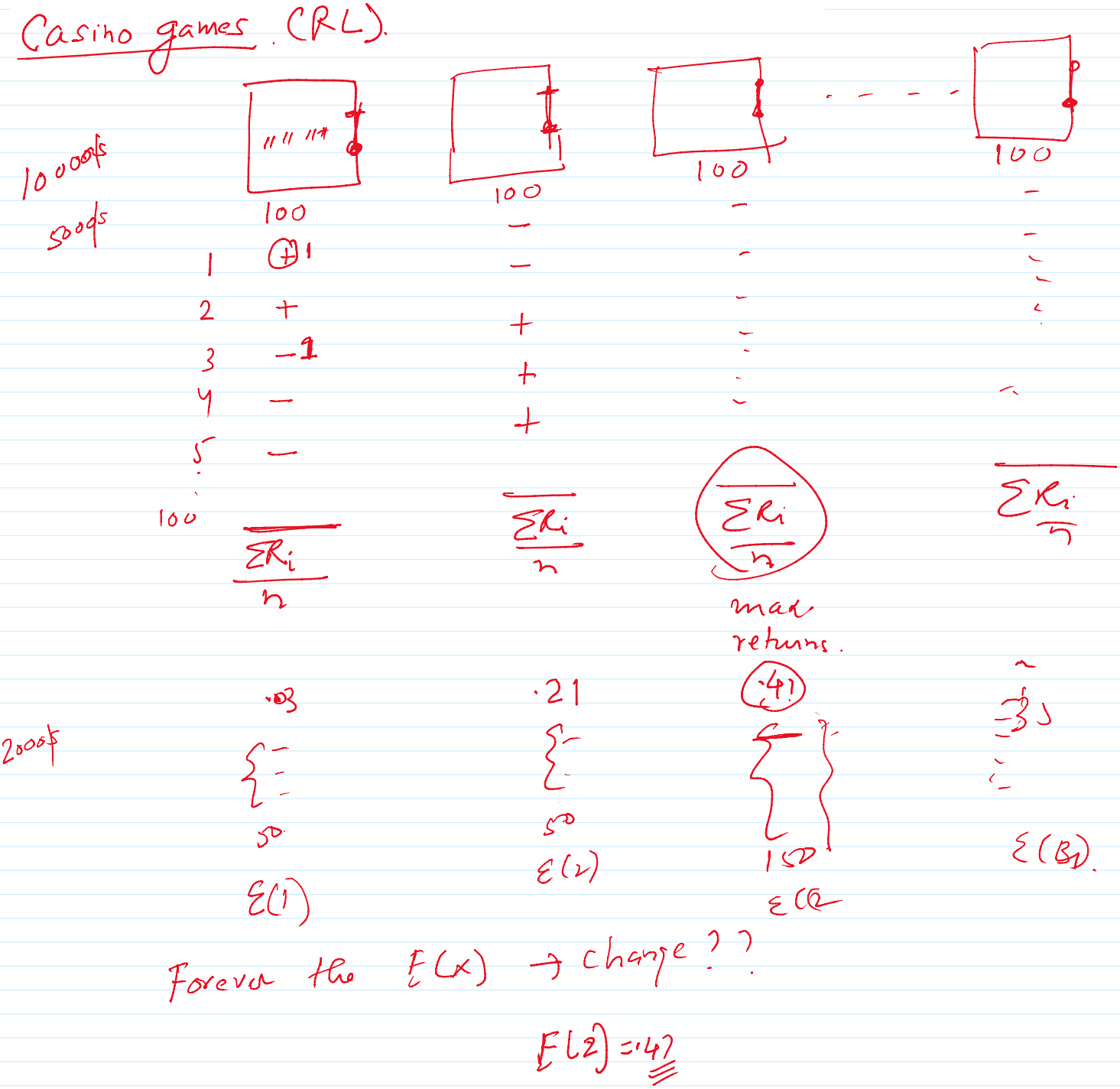
$E[X]$

= 0×0.05+1×0.1+2×0.15+3×0.16+4×0.2+5×0.13+6×0.1+7×0.07+8×0.04
= 0+0.1+0.3+0.48+0.8+0.65+0.6+0.49+0.32
= 3.74

2. What is the probability that the number of complaints will exceed the expected number?

= 0.2 + 0.13 + 0.1 + .07 + .04
= 0.54

Complaints	Probability
0	0.05
1	0.1
2	0.15
3	0.16
4	0.2
5	0.13
6	0.1
7	0.07
8	0.04



Types of discrete probability distributions

Bernoulli Distribution

- **Description**: represents the probability distribution of a random variable that can take on only two values, typically labeled as success (usually denoted by 1) and failure (usually denoted by 0), with a fixed probability of success p and failure $1-p$.
- **Usage**: used to model binary outcomes, such as success/failure, yes/no, heads/tails in a coin flip, etc.

Bernoulli distribution

- pertains to scenarios featuring a single trial with two potential outcomes
- experiments posing a binary question
  - whether a coin will land on heads,
  - if a die roll will result in a 6,
  - if an ace will be drawn from a deck of cards, or
  - if voter X will opt for "yes" in a referendum.
  - a team will win a championship or not
- Essentially, Bernoulli trials encompass situations where the two potential results can be framed as "success" or "failure," though these terms aren't strictly literal.
- In this context, "success" simply denotes achieving a "yes" outcome (e.g., rolling a six, drawing an ace, etc.).

Basic Properties of Bernoulli distribution

- The **expected value** is

  $E(X)$
  $= 0 \times (1-p) + 1 \times p$
  $= p$
- The **variance** is

  $Var(X)$
  $= E(X^2) - E(X)^2$
  $= 1^2 \times p + 0^2 \times (1-p) - p^2$
  $= p - p^2$
  $= p(1-p)$
- The **mode**, the value with the highest probability of occurring, is 1 if $p>0.5$ and 0 if $p<0.5$.
  - If $p=0.5$, success and failure are equally likely and both 0 and 1 are modes.

key assumptions

Binary Outcome: the outcome of interest can be categorized into only two possible outcomes, typically denoted as success and failure, yes and no, or 1 and 0.

Consistent Probability: The probability of success (p) must remain constant for each trial. This assumption implies that the underlying process generating the outcomes does not change over time or across trials.

Independence: The outcome of one trial must be independent of the outcome of another trial. … the occurrence or non-occurrence of success in one trial should not influence the probability of success in subsequent trials.

Types of discrete probability distributions

Binomial Distribution:

- **Description**: The binomial distribution describes the probability distribution of the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success p.
- **Usage**: It is used to model situations involving a fixed number of independent trials with two possible outcomes, such as the number of successful surgeries in a series of operations, the number of heads obtained in a series of coin flips, etc.

What is binomial distribution?

01

essentially a series of independent Bernoulli trials.

02

outcome of each trial is either success or failure (p or 1-p),

03

Binomial distribution summarizes the total number of successes across these trials.

Binomial vs Bernoulli

binomial distribution, extends the concept of the Bernoulli distribution to scenarios where the event is repeated multiple times

Each trial follows a Bernoulli distribution (with two possible outcomes: success and failure)

trials are independent of each other

probability of success (p) remains constant across all trials

probability mass function (PMF) of the binomial distribution

binomial distribution allows us to calculate the probability of getting a specific number of successes (or failures) out of a fixed number of independent trials.

provides a probability distribution for the number of successes.

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Where:

- n is the number of trials,
- k is the number of successes,
- p is the probability of success in each trial.

Mean variance mode of Binomial distribution

1. **Mode of a Binomial Distribution:**

$$\text{Mode} = \lfloor (n+1) \times p \rfloor$$

where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x.

2. **Mean of a Binomial Distribution:**

$$\mu = np$$

3. **Variance of a Binomial Distribution:**

$$\sigma^2 = np(1-p)$$

1. $\lfloor 3.7 \rfloor = 3$ because the greatest integer less than or equal to 3.7 is 3.
2. $\lfloor 2 \rfloor = 2$ because 2 is already an integer.
3. $\lfloor -1.5 \rfloor = -2$ because the greatest integer less than or equal to -1.5 is -2.
4. $\lfloor 4 \rfloor = 4$ because 4 is already an integer.

parameters (μ, σ)

parametric (μ, σ)

non-parametric

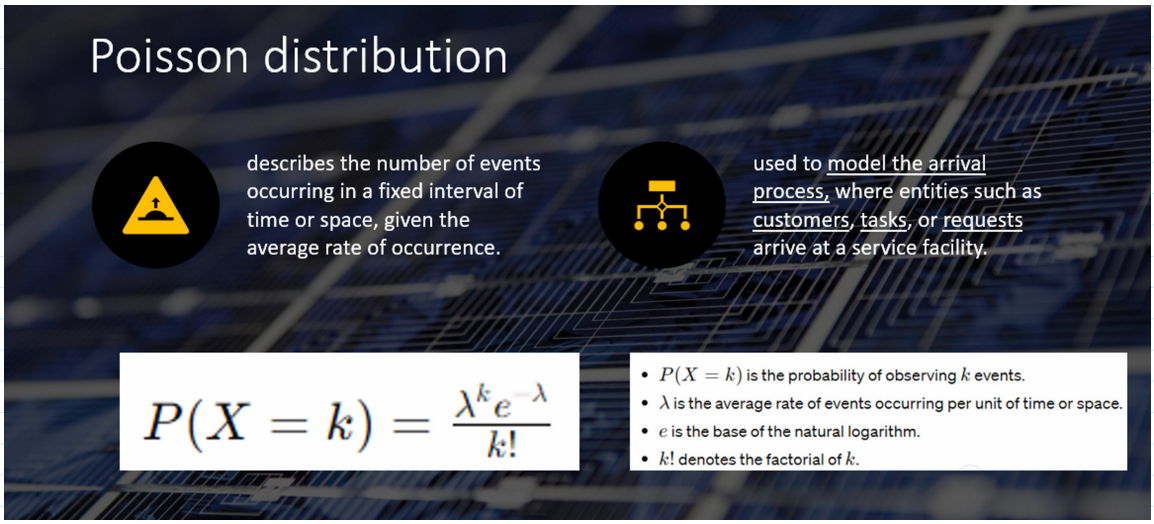
Types of discrete probability distributions

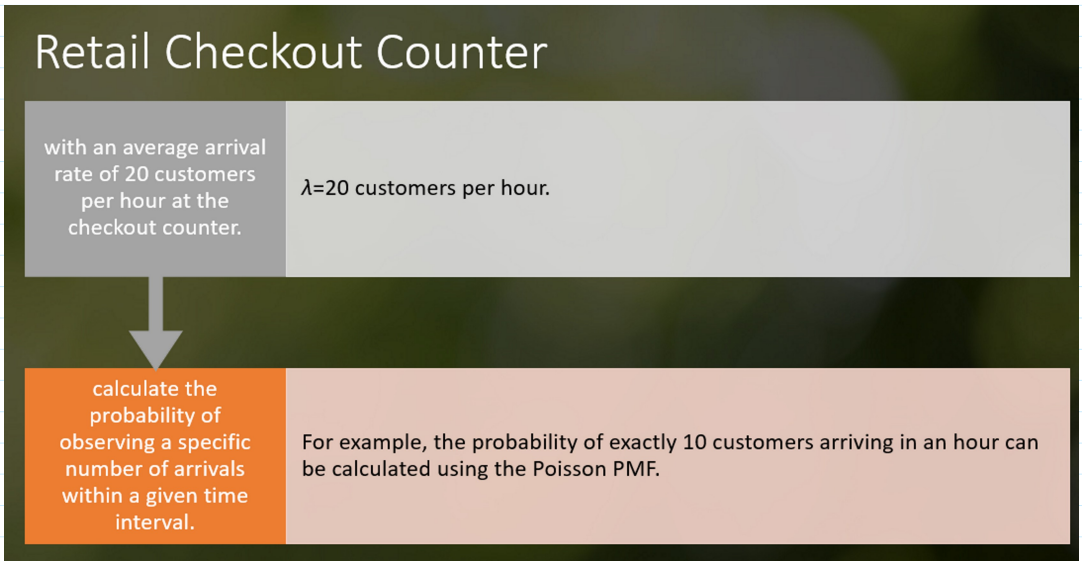
Types of discrete probability distributions

Poisson Distribution:

Description: represents the probability distribution of the number of events occurring in a fixed interval of time or space, given the average rate of occurrence λ.

Usage: commonly used to model rare events or occurrences that happen independently at a constant average rate, such as the number of phone calls received by a call center in a given hour, the number of accidents at a specific intersection in a day, etc.





Similar examples

Call Center Operations

- calls from customers arrive at a rate of 5 calls per minute. λ=5 calls per minute.
- model the arrival process of calls, allowing call center managers to analyze call volumes and plan staffing levels accordingly.

Website Traffic

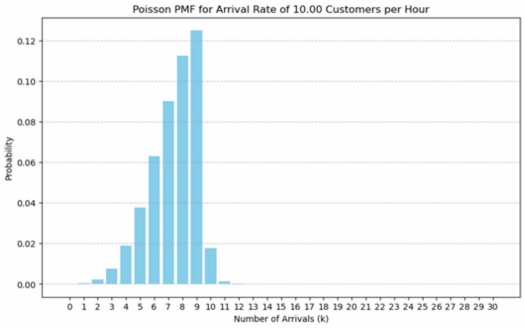
- the number of user requests or visits per minute can be modeled using the Poisson distribution. If the website receives an average of 100 requests per minute, λ=100 requests per minute.

Hospital Emergency Room

- Hospital administrators can use the Poisson distribution to analyze patient arrival patterns and optimize staffing levels and resource allocation in the emergency room.

Network Packet Arrival

- Network administrators can use the Poisson distribution to analyze network traffic patterns and plan network capacity and bandwidth requirements.




```
lambda_val = 10
k          = 10

poisson_pmf(lambda_val, k)
```

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Example

Example

```
lambda_val = 10
k          = 10

poisson_pmf(lambda_val, k)
```

pmf

Cumulative probability.

- 0.1251100357211333
- result 0.12 represents the probability associated with observing exactly 10 events (customer arrivals) in the specified scenario.
- It indicates that there is approximately a 12% chance of exactly 10 customers arriving at the retail checkout counter within the given time interval, based on the Poisson distribution with an average arrival rate of 10 customers per hour.

10-08-2024

CDF of Poisson distribution

$$F(X \leq k) = \sum_{i=0}^{k} \frac{\lambda^i e^{-\lambda}}{i!}$$

For the Poisson distribution, the CDF is the cumulative sum of the PMF up to a given value.

It gives the probability that the number of events is less than or equal to a certain value.

why Poisson distribution is used in sampling arrival data

Memoryless Property

- exhibits the memoryless property, which means that the probability of an arrival occurring in a given time interval is independent of the number of arrivals that have occurred before.
- In queuing systems, this property aligns well with the assumption of random and independent arrivals.

Models Rare Events

- Poisson distribution is suitable for modeling rare events, where the probability of multiple arrivals within a short time period is low.
- This is often the case in queuing systems where arrivals are infrequent or occur at a relatively low rate compared to the duration of the time intervals being considered.