Probability distribution and data

By

Dr Shaik A Qadeer

Professor MJCET

Content

- Random variable
- Probability distribution
  - Discrete distribution(Bio-nomial, Poisson and Geometric distribution)
  - Continuous distribution(Uniform, Exponential and Normal)

Random variable

- A **random variable** is a **function** that maps every **outcome** in the **sample space** to a **real number**. It can both be **discrete** and **continuous**
- **Discrete random variable** – If the random variable X can assume only a finite or countably infinite set of values, then it is called a discrete random variable. Examples:
    1. Credit rating (low, medium, and high credit rating)
    2. Customer churn (churn and do not churn)
    3. Fraud (fraudulent transaction and genuine transaction)
- They are described using probability mass function (PMF) and cumulative distribution function (CDF)

Random variable..

- **Continuous random variable** – A random variable X which can take a value from an infinite set of values is called a continuous random variable.
- Examples:
    1. Market share of a company (any value between 0 and 100%).
    2. Percentage of attrition of employees of an organization.
    3. Time-to-failure of an engineering system.
- They are described using probability density function (PDF) and cumulative distribution function (CDF)

Discrete Probability functions

- Bio-nomial distribution ,
- Poisson distribution and
- Geometric distribution

Bionomial distribution function

- It is a discrete probability distribution function
- A random variable X is said to follow a binomial distribution if:
  1. Random variable can have only two outcomes – success and failure
  2. Objective is to find the probability of getting x successes out of n trials
  3. Probability of success is p and probability of failure is (1-p)
  4. Probability p is constant and does not change between trials

Calculation of bionomial distribution

1) By probability mass function (PMF): This is used for exactly equal case and

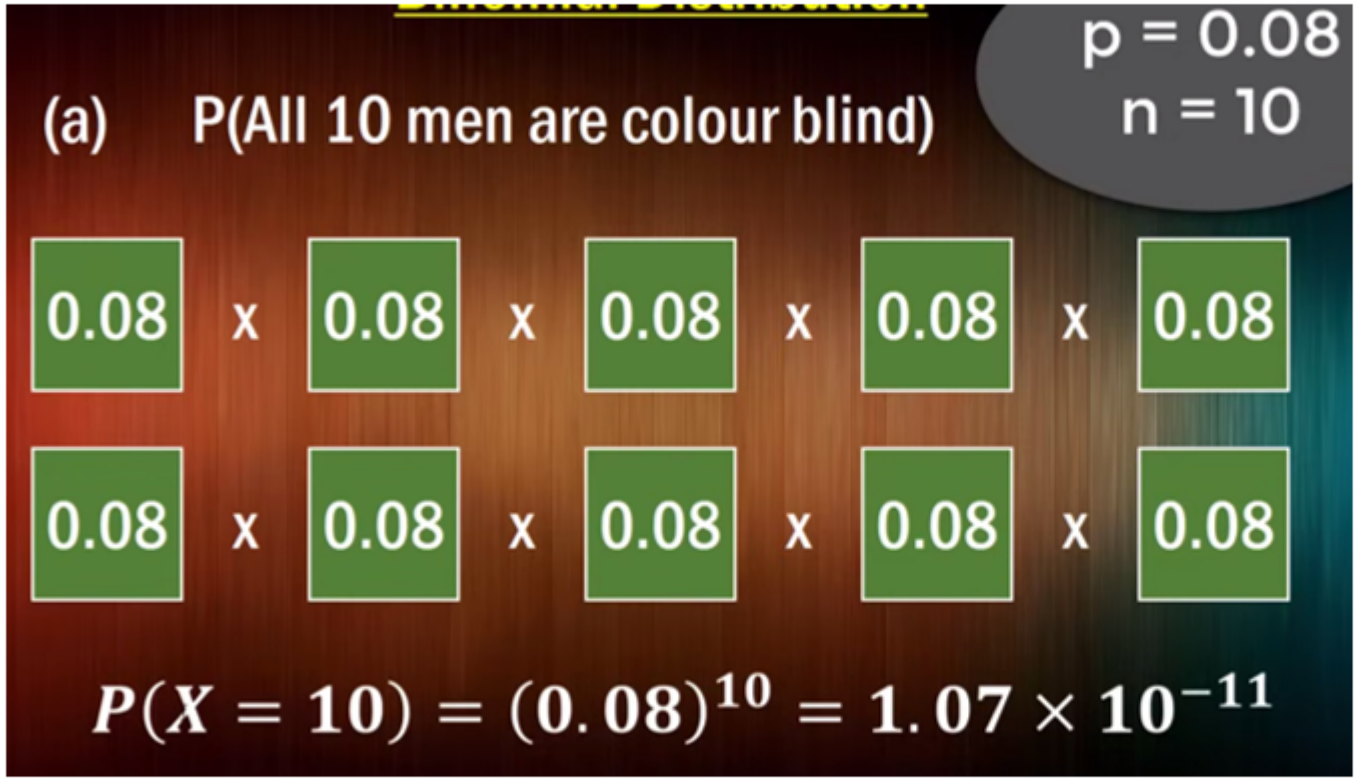
$$P(x) = {}^{n}C_{x} p^{x} q^{n-x} = \frac{n!}{(n-x)!x!} p^{x} q^{n-x}$$

2) Cumulative distribution function (PDF): This is use for less than or equal to (or maximum)cases

$$F(r) = \sum_{x=0}^{r} \binom{n}{x} p^{x} \, q^{(n-x)}$$

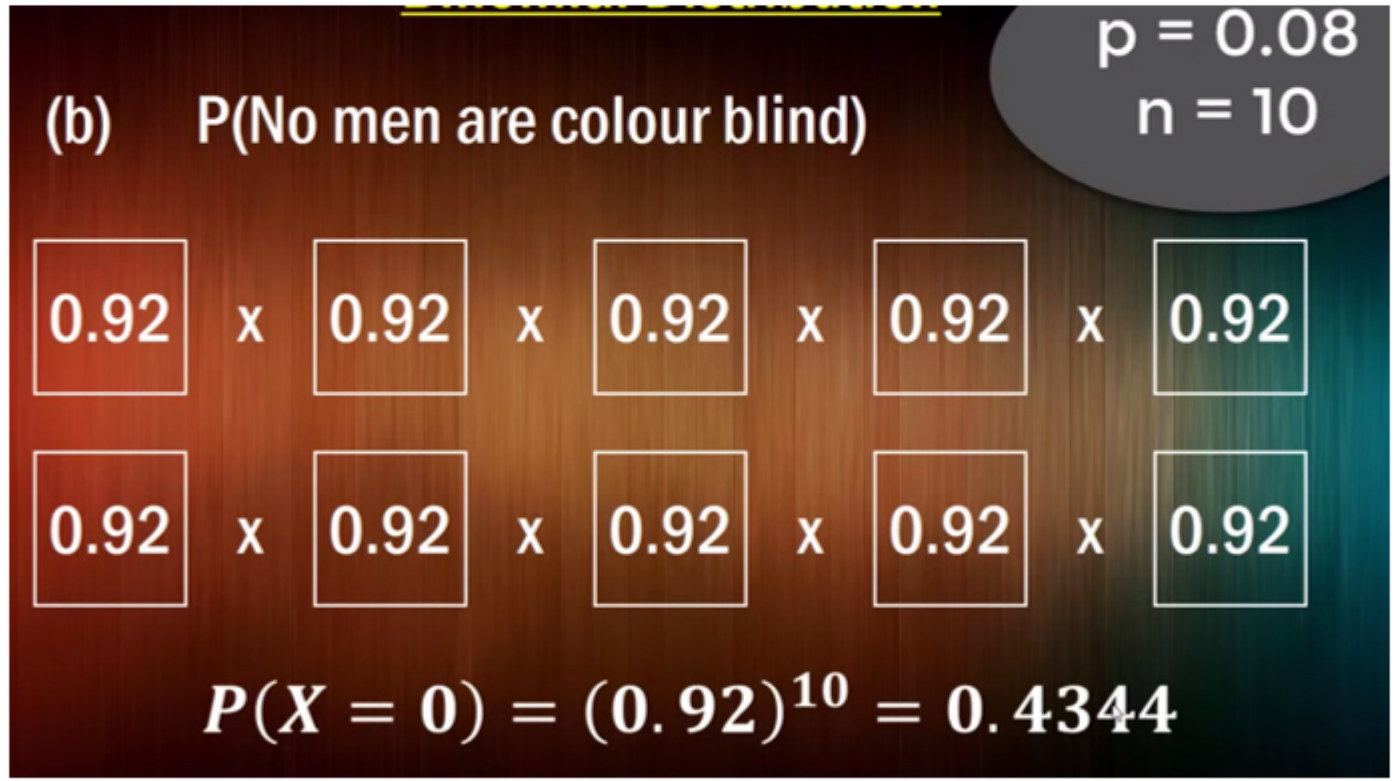
Case study of Probability calculation using PMF

Case study of Probability calculation using PMF. “All 10 mens are blind”



$P(X=10) = {}^{10}C_{10} \times 1.07 \times 10^{-11} = 1.07 \times 10^{-11}$

Case study of Probability calculation using PMF. “No mens are blind”

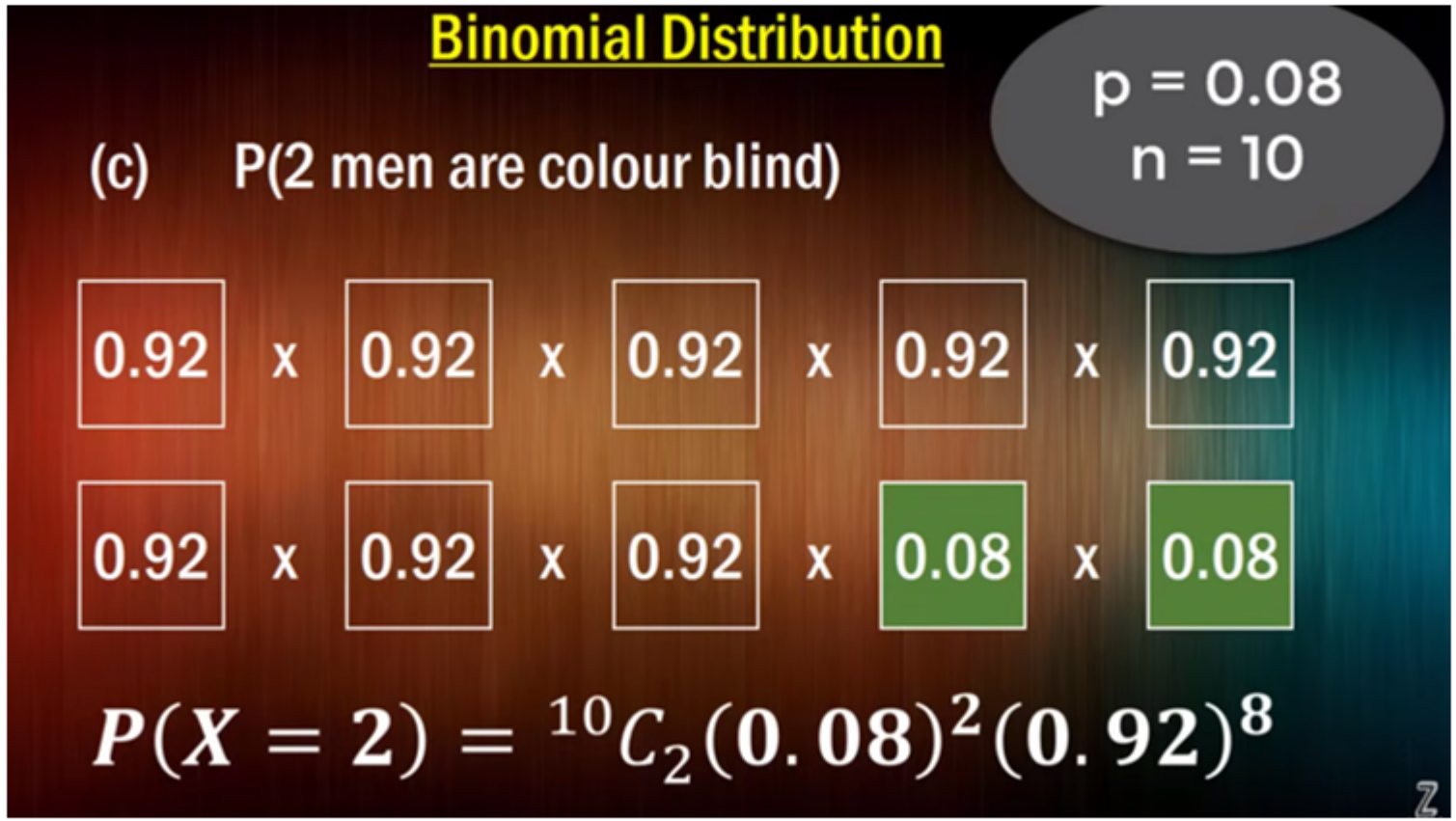


$P(X=0) \quad {}^{10}$

$= {}^{10}C_0 \times (0.92)^{10}$

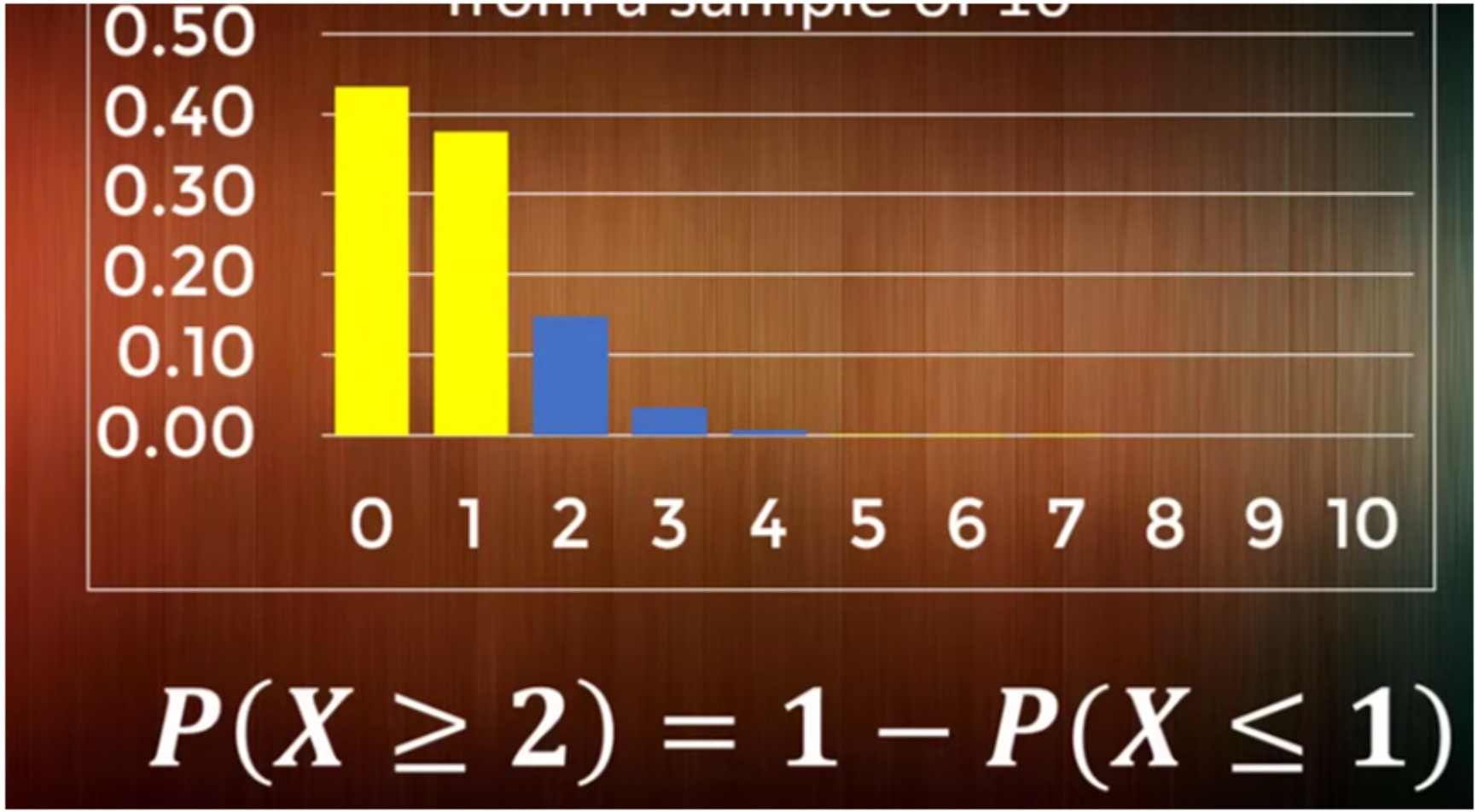
$= 0.4344$

Case study of Probability calculation using PMF. Exactly 2 mens are blind



=0.148

Case study of Probability calculation using PMF. Alteast 2 mens are blind



$= 1 -$
$P(x=0)$
$-$
$P(x=1)$
$= 1 - 0.378 -$
0434

Poisson distribution

- Consider the following business problems:

    1. Number of cancellation of orders by customers at an e-commerce portal

    2. Number of customer complaints

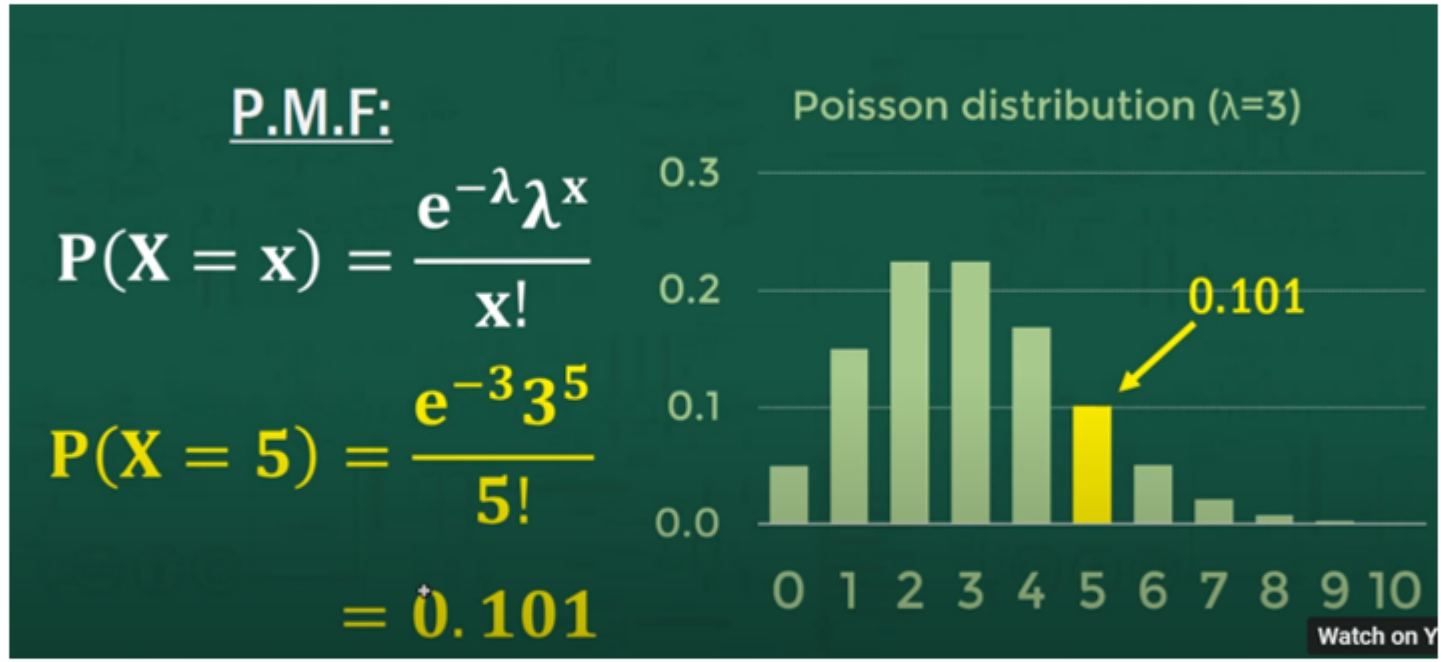
    3. Number of cash withdrawals at an ATM

- All these problems are can be describe by the number of events occurring in a fixed intervals of time
- This can be done with poisons distribution

Poisson distribution..

- It’s a discrete distribution
- Its describe the number of events occurring in a fixed intervals of time
- It requires only one parameter(lamda=time interval)

PMF of poisons distribution



Probability of getting 5th event in time interval equals to 3(lambda)

Question on poisson dis

- Q:The number of calls arriving at a call center follows a Poisson distribution at 10 calls per hour.

1. Calculate the probability that the number of calls will be maximum 5.

2. Calculate the probability that the number of calls over a 3-hour period will exceed 30 calls.

1)Calculate the probability that the number of calls will be maximum 5.

- We can use for this cdf function, as it is maximum number finding probability
- stats.poisson.cdf(event, lambda)=stats.poisson.cdf(5, 10)

=0.067

2) Calculate the probability that the number of calls over a 3-hour period will exceed 30 calls

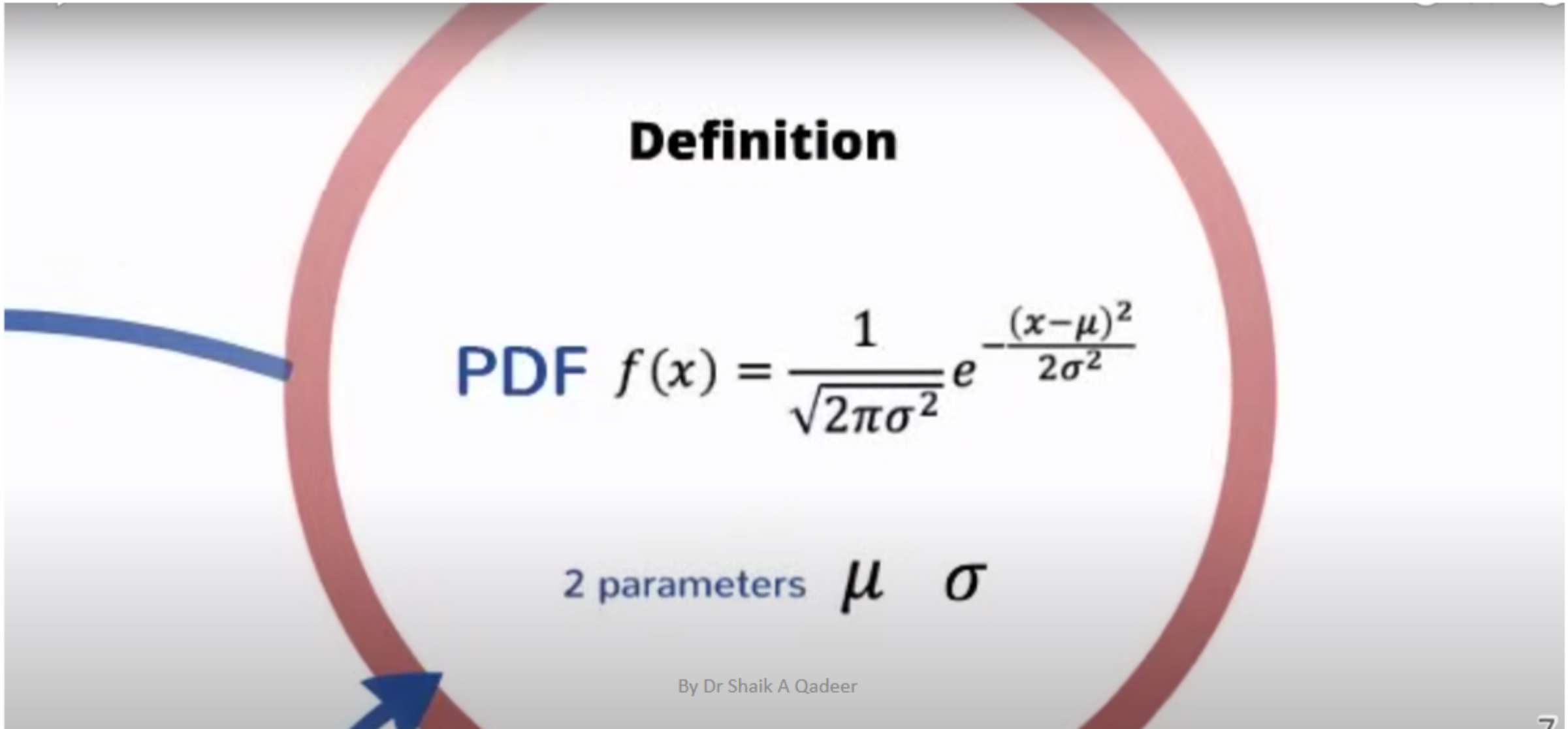
- One period is 10 calls per hour , so 3 hour period=30
- It is a question of knowing probability after this period so
- The python code is “1 - stats.poisson.cdf(30, 30)”

  =0.45

Normal distribution: Intro

- Also known as Gaussian distribution
- A continuous distribution
- Normal distribution is observed across many naturally occurring measures like: age, salary, sale volume, birth weight, height, etc.
- Popularly known as bell curve

Normal distribution: Intro.. PDF of it is



Normal distribution: Let us dive into normal distribution with a case study

- Imagine a scenario where an investor wants to understand the risks and returns associated with various stocks before investing in them.
- We will evaluate two stocks: BEML and GLAXO.
- The daily trading data for each stock is taken for the period starting from 2010 to 2016 from BSE site.
- Reference: (www.bseindia.com)