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

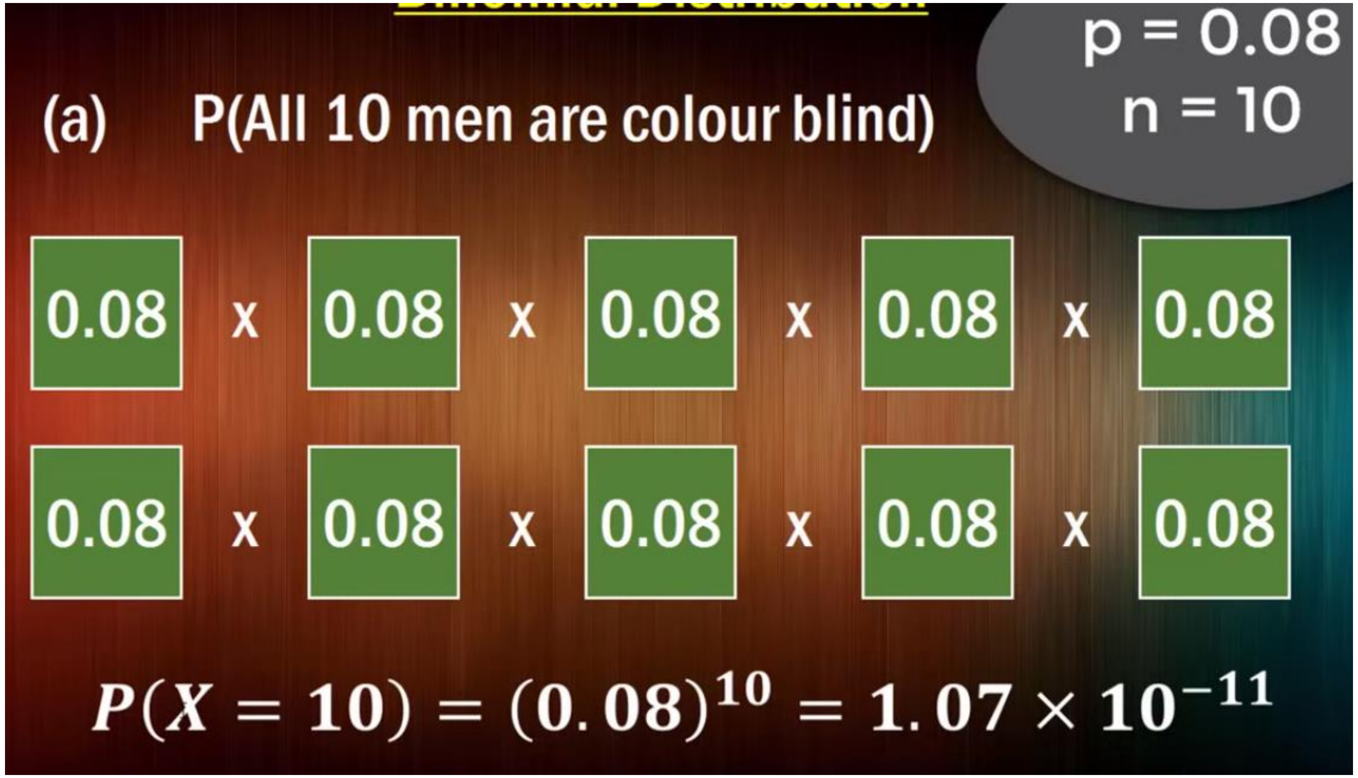
Studies show colour blindness affects about 8% of men.
A random sample of 10 men is taken.

$p = 0.08$
$n = 10$

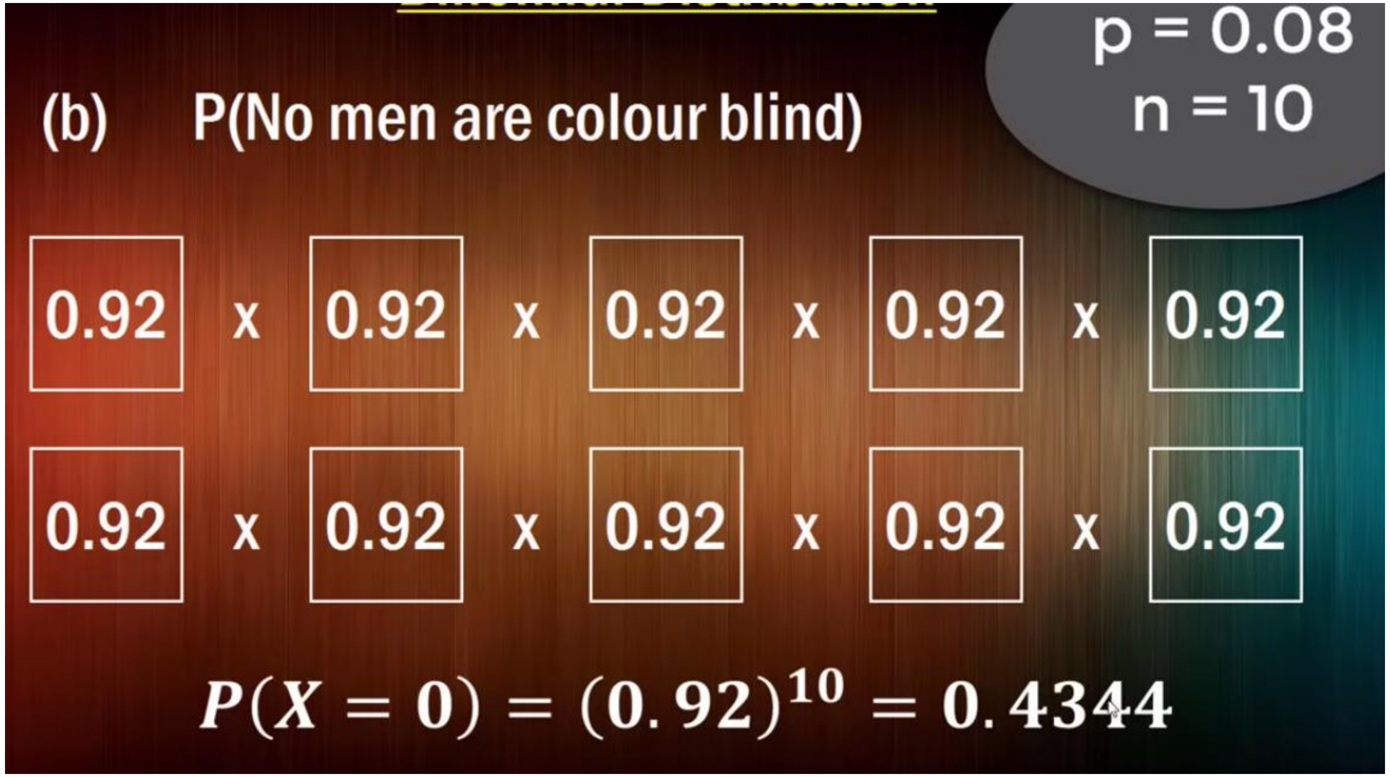
Find the probability that:

(a) All 10 men are colour blind
(b) No men are colour blind
(c) Exactly 2 men are colour blind
(d) At least 2 men are colour blind

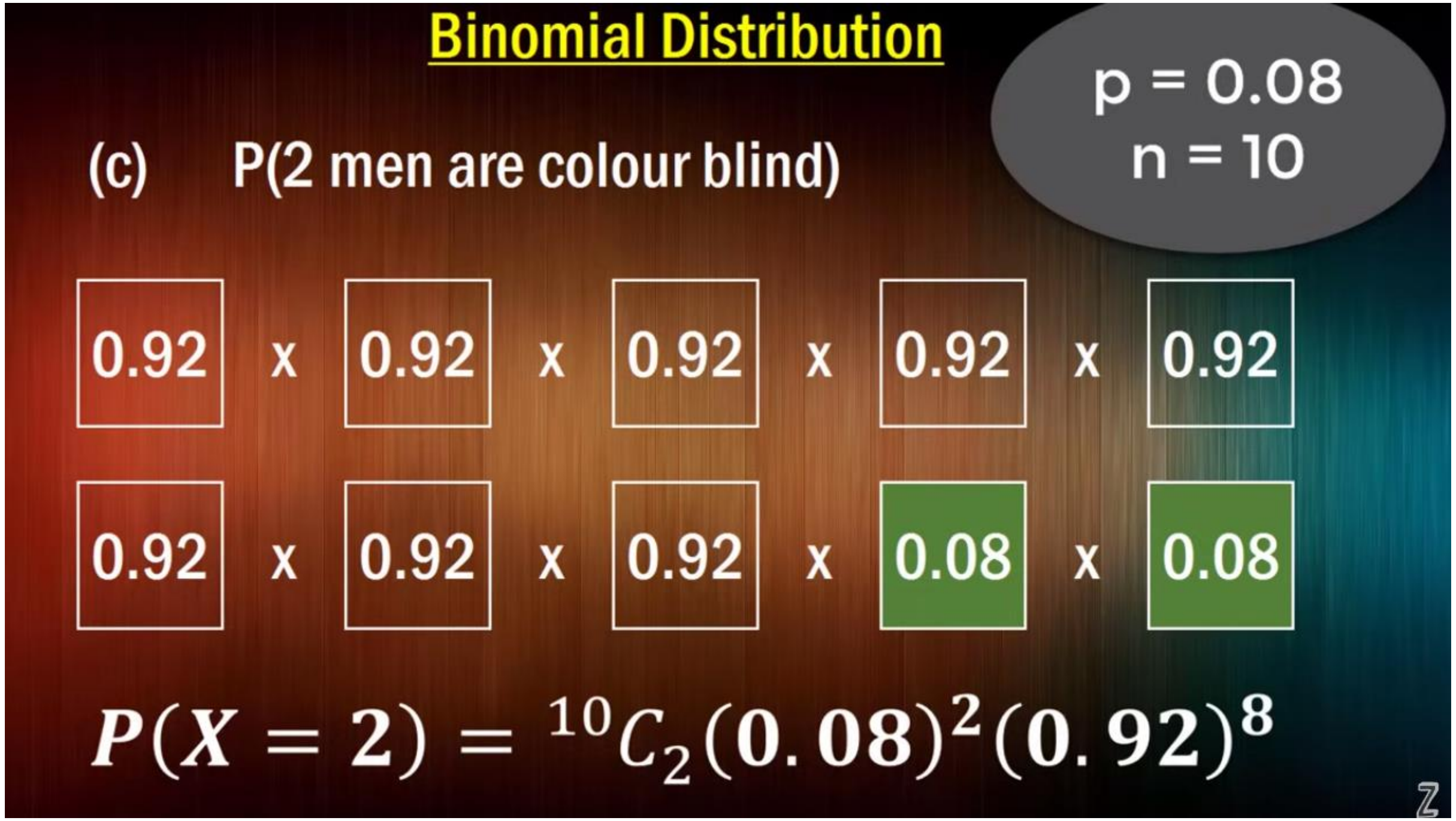
Case study of Probability calculation using PMF. "All 10 mens are blind"



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

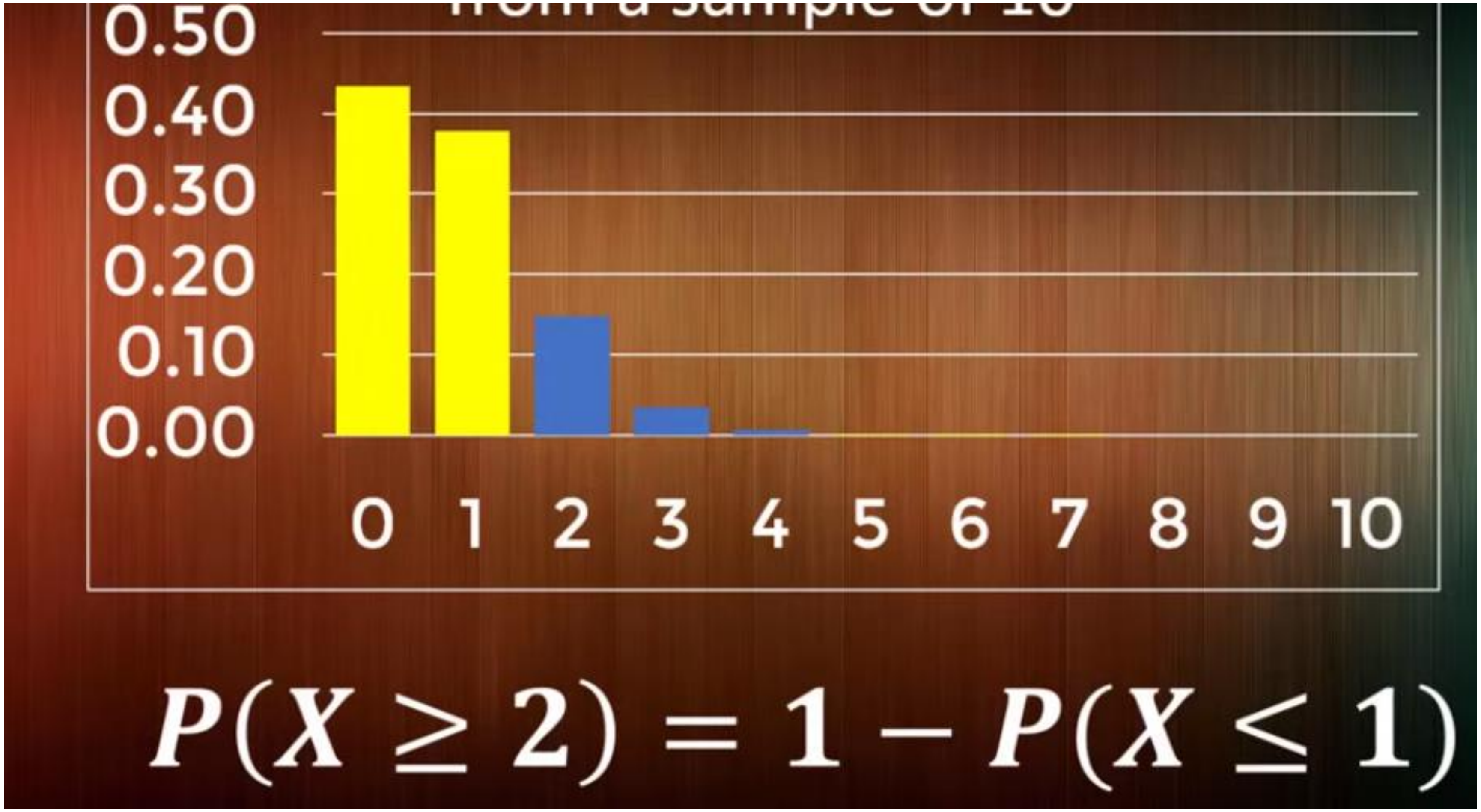


Case study of Probability calculation using PMF. Exactly 2 mens are blind



=0.148

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



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

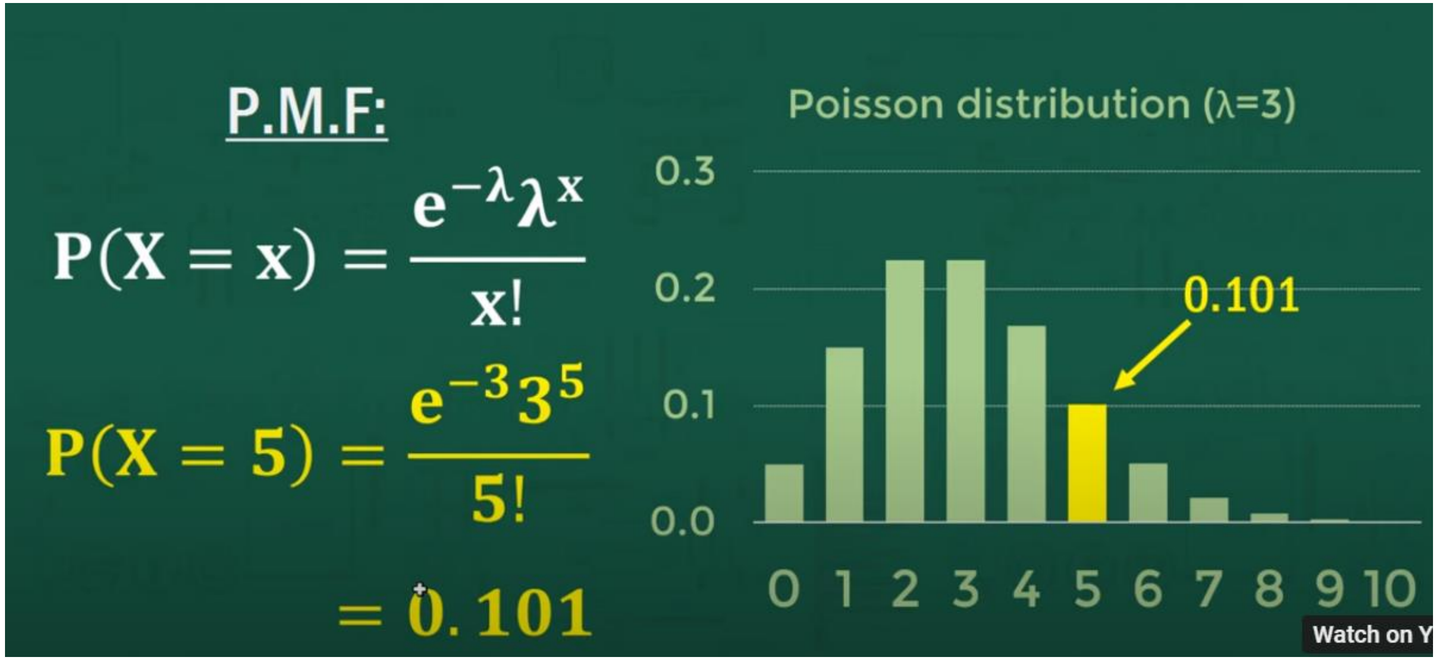
  3. Number of cash withdrawals at an ATM

- 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

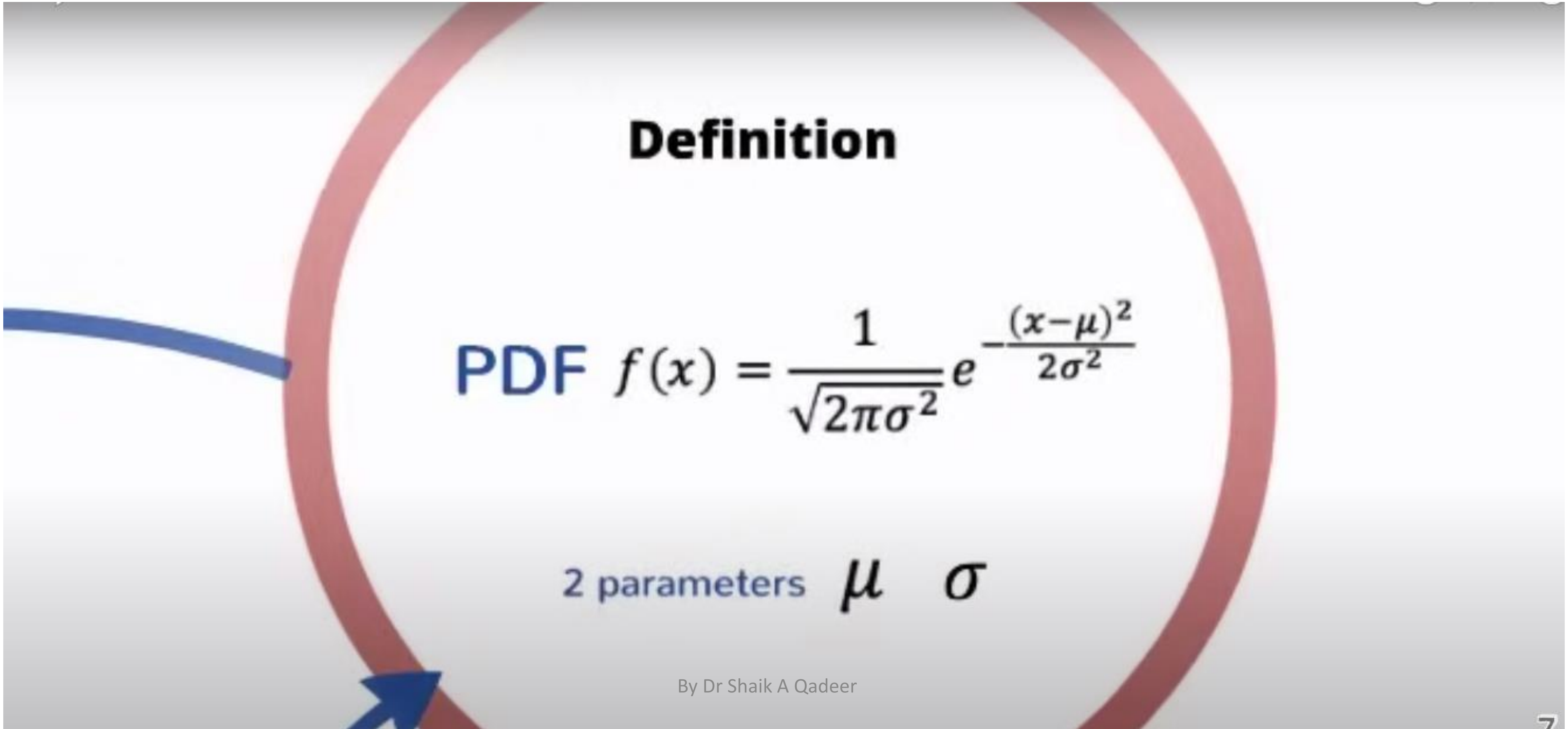


Probability of getting 5th event in time interval equals to 3(lambda)

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)

Normal distribution..
Solution: loading the data(BEML)

```
import pandas as pd
import numpy as np
import warnings

beml_df = pd.read_csv('BEML.csv')
beml_df[0:5]
```

	Date	Open	High	Low	Last	Close	Total Trade Quantity	Turnover (Lacs)
0	2010-01-04	1121.0	1151.00	1121.00	1134.0	1135.60	101651.0	1157.18
1	2010-01-05	1146.8	1149.00	1128.75	1135.0	1134.60	59504.0	676.47
2	2010-01-06	1140.0	1164.25	1130.05	1137.0	1139.60	128908.0	1482.84
3	2010-01-07	1142.0	1159.40	1119.20	1141.0	1144.15	117871.0	1352.98
4	2010-01-08	1156.0	1172.00	1140.00	1141.2	1144.05	170063.0	1971.42

Normal distribution..
Solution: loading the data(GLAXO)..

```
glaxo_df = pd.read_csv('GLAXO.csv')
glaxo_df[0:5]
```

	Date	Open	High	Low	Last	Close	Total Trade Quantity	Turnover (Lacs)
0	2010-01-04	1613.00	1629.10	1602.00	1629.0	1625.65	9365.0	151.74
1	2010-01-05	1639.95	1639.95	1611.05	1620.0	1616.80	38148.0	622.58
2	2010-01-06	1618.00	1644.00	1617.00	1639.0	1638.50	36519.0	595.09
3	2010-01-07	1645.00	1654.00	1636.00	1648.0	1648.70	12809.0	211.00
4	2010-01-08	1650.00	1650.00	1626.55	1640.0	1639.80	28035.0	459.11

Normal distribution.. Solution:..

- Selecting Date and Close columns from the DataFrames, since the analysis will involve only daily prices.

```
beml_df = beml_df[['Date', 'Close']]
glaxo_df = glaxo_df[['Date', 'Close']]
```

- Setting the Datetime Index

```
glaxo_df = glaxo_df.set_index(pd.DatetimeIndex(glaxo_df['Date']))
beml_df = beml_df.set_index(pd.DatetimeIndex(beml_df['Date']))
```

Normal distribution.. Solution:..

- Plotting the trend of close prices of GLAXO stock.

```
import matplotlib.pyplot as plt
import seaborn as sn
%matplotlib inline

plt.plot(glaxo_df.Close);
plt.xlabel('Time');
plt.ylabel('Close Price');
```

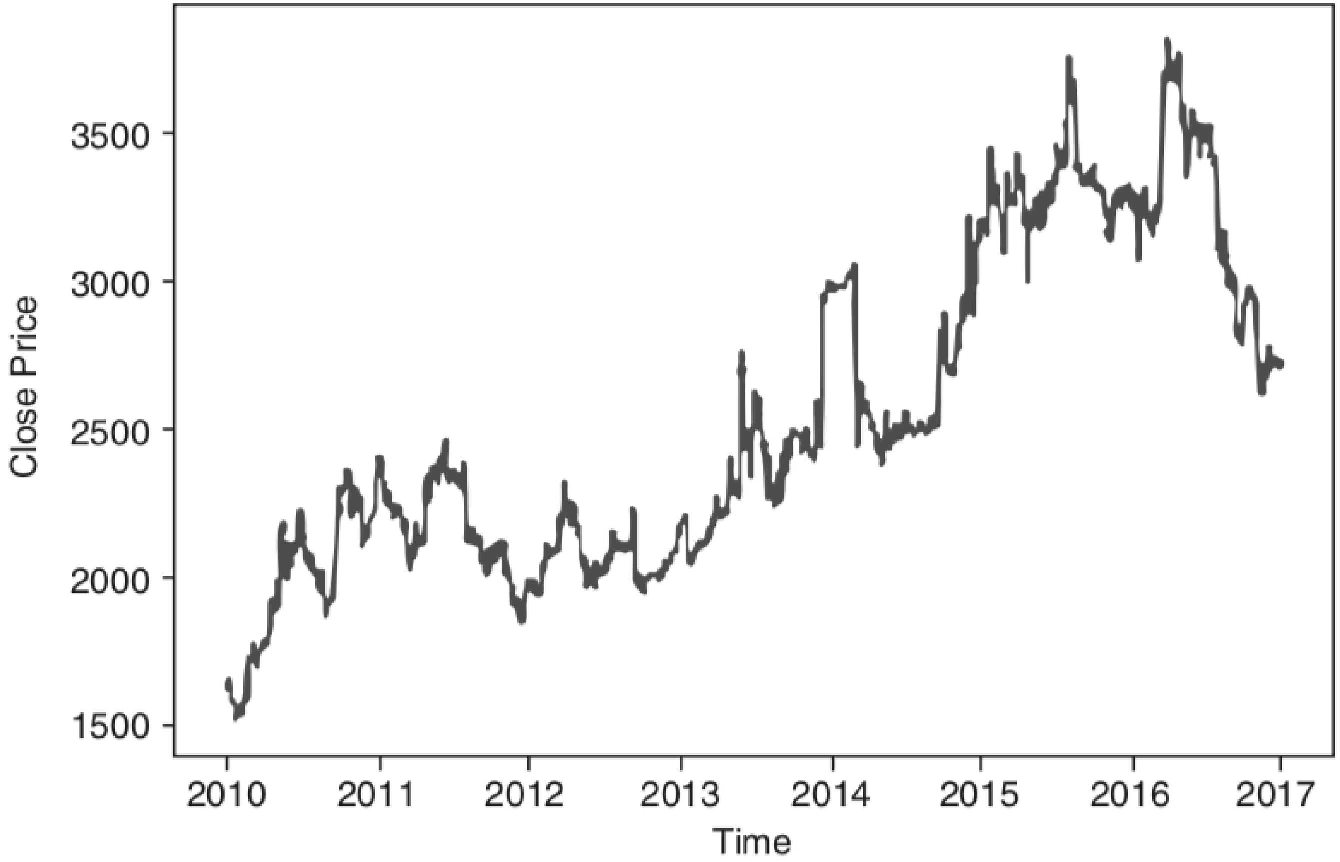


FIGURE 3.4 Close price trends of GLAXO stock.

Normal distribution.. Solution:..

- Plotting the trend of close prices of BEML stock.

```
plt.plot(beml_df.Close);
plt.xlabel('Time');
plt.ylabel('Close');
```

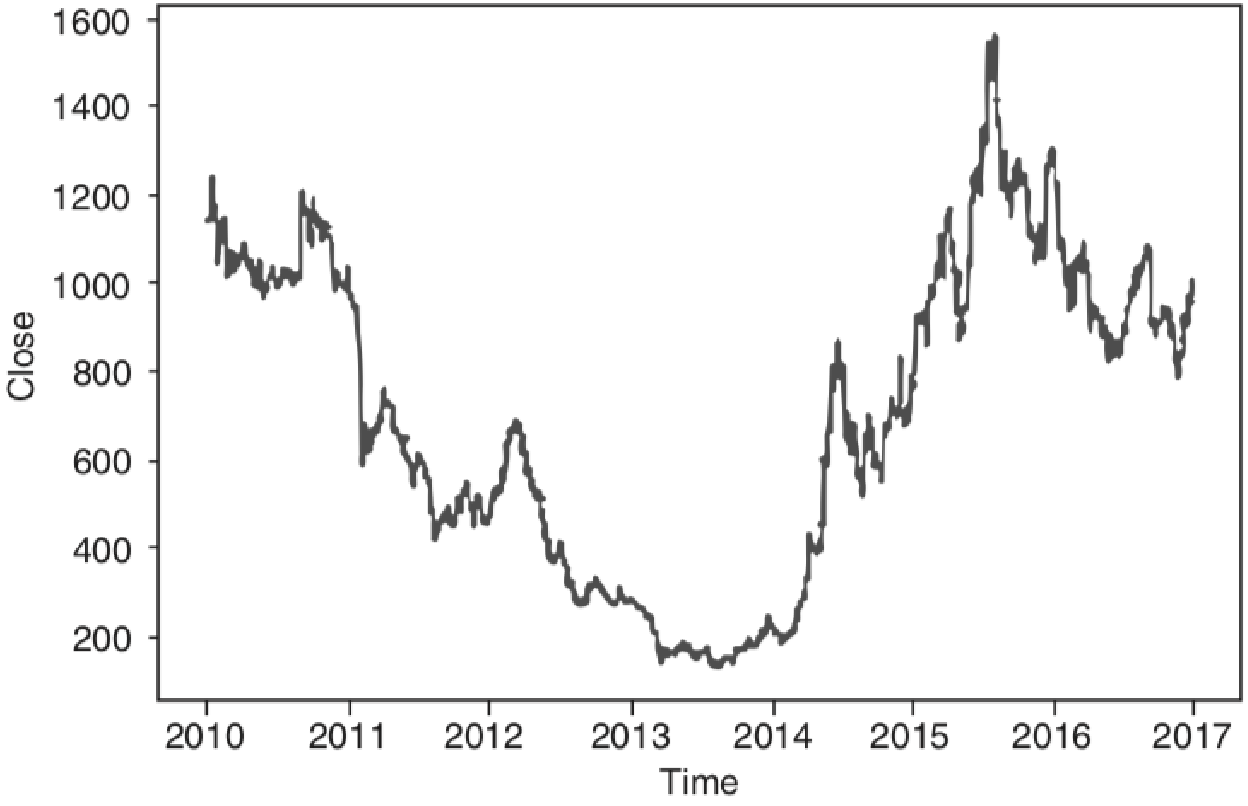


FIGURE 3.5 Close price trends of BEML stock.

ND Solution:..

- The behavior of daily returns on the stocks is called Gain.

$$gain = \frac{ClosePrice_t - ClosePrice_{t-1}}{ClosePrice_{t-1}}$$

- In Pandas it can be calculated as

```
glaxo_df['gain'] = glaxo_df.Close.pct_change(periods = 1)
beml_df['gain'] = beml_df.Close.pct_change(periods = 1)
glaxo_df.head(5)
```

	Date	Close	Gain
Date			
2010-01-04	2010-01-04	1625.65	NaN
2010-01-05	2010-01-05	1616.80	–0.005444
2010-01-06	2010-01-06	1638.50	0.013422
2010-01-07	2010-01-07	1648.70	0.006225
2010-01-08	2010-01-08	1639.80	–0.005398

ND Solution:..

- Plotting gain against time

```
plt.figure(figsize = (8, 6));
plt.plot(glaxo_df.index, glaxo_df.gain);
plt.xlabel('Time');
plt.ylabel('gain');
```

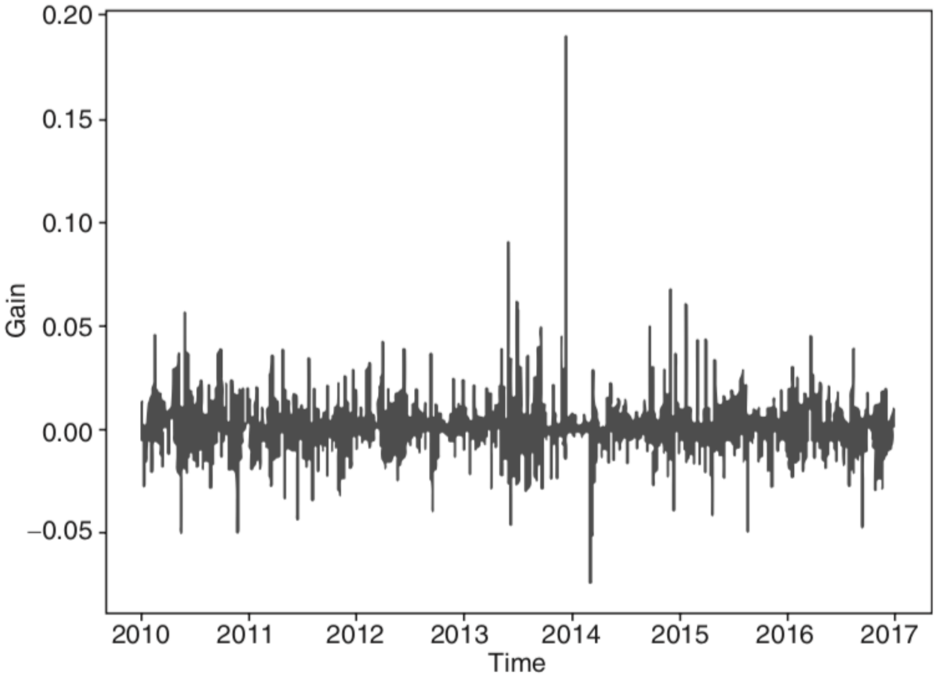


Figure: Daily gain of Glaxo stock

ND Solution:..

- Distribution plot of gain for both BEML and GLAXO stocks

```
sn.distplot(glaxo_df.gain, label = 'Glaxo');
sn.distplot(beml_df.gain, label = 'BEML');
plt.xlabel('gain');
plt.ylabel('Density');
plt.legend();
```

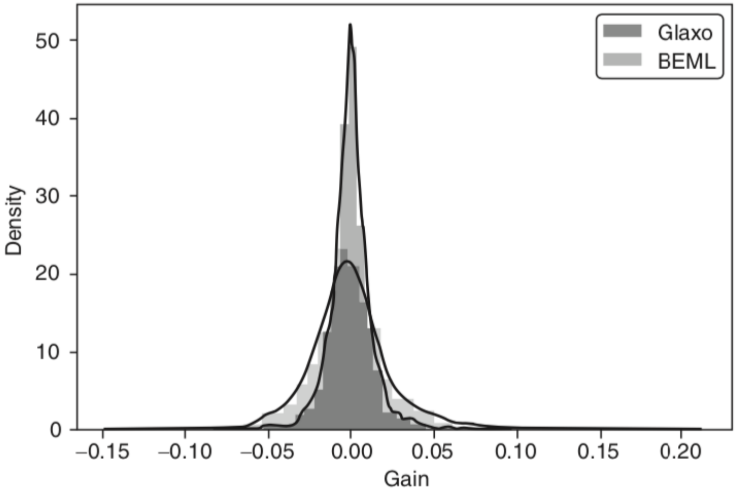


FIGURE 3.7 Distribution plot of daily gain of BEML and Glaxo stocks.

- Gain seems to be normally distributed for both the stocks with a mean around 0.00
- BEML seems to have a higher variance than GLAXO

ND Solution:..

- The sample mean of a normal distribution is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Variance is given by

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

ND Solution:..

- In Pandas, the sample mean and standard deviation for daily returns

for GLAXO and BEML are

```
print("Daily gain of Glaxo")
print("--------------------")
print("Mean: ", round(glaxo_df.gain.mean(), 4))
print("Standard Deviation: ", round(glaxo_df.gain.std(), 4))
```

```
Daily gain of Glaxo
---------------------
Mean:                  0.0004
Standard Deviation:    0.0134
```

```
print("Daily gain of BEML")
print("--------------------")
print("Mean: ", round(beml_df.gain.mean(), 4))
print("Standard Deviation: ", round(beml_df.gain.std(), 4))
```

```
Daily gain of BEML
---------------------
Mean:                  0.0003
Standard Deviation:    0.0264
```

ND Solution:..

- The describe() method of DataFrame returns the detailed statistical summary of a variable

```
beml_df.gain.describe()
```

```
count      1738.000000
mean          0.000271
std           0.026431
min          -0.133940
25%          -0.013736
50%          -0.001541
75%           0.011985
max           0.198329
Name: gain, dtype: float64
```

- BEML stock has higher risk as standard deviation of BEML is 2.64% whereas the standard deviation for GLAXO is 1.33%