## Probability Distributions

- Discrete Distributions: Random variable takes discrete values
  - Bernoulli Distribution
  - Binomial Distribution
  - Poisson Distributon
- Continous Distributions: Random variable takes continuous values in a range
  - Normal Distribution
  - Uniform Distribution
  - Exponential Distribution etc.

## Binomial distribution

- The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N.
- the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes(x) in a sequence of n independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: success (with probability p) or failure (with probability q = 1 − p).
- A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment. That is, when n = 1, the binomial distribution is a Bernoulli distribution.
- The probability mass function is given by -

$$P_x = \binom{n}{x} p^x q^{n-x}$$

- The mean of Binomial distribution is $np$ and the variance is $npq$

Examples:

- Gender of babies delivered in a hospital
- Fatal side effect deaths for a Schedule H drug
- Getting a head when a coin is tossed several times

[1]:
```
#Binomial distribution
from scipy.stats import binom
```

Example:

Consider a random experiment of tossing a coin 6 times. Find the following:

- What is the probability of getting exactly 0 heads ?
- What is the proability of getting 2 or less number of heads?

***Solution***: Here, $n = 6$, $p = q = \frac{1}{2}$ and $x = 0$

- We need to find $P(X = 0)$ and $P(X \leq 2)$

[2]:
```
# P(X=0)
# binom.pmf has a syntax as: pmf(x,n,p)
binom.pmf(0,6,0.5)
```

[2]:
```
0.015625
```

[2]:
```
0.015625
```

[3]:
```
#Find the probability of seeing less or equal to 2 heads.
# P(x<=2)= P(x=0)+ P(x=1) + P(x=2)
binom.pmf(0,6,0.5) + binom.pmf(1,6,0.5) + binom.pmf(2,6,0.5)
```

[3]:
```
0.34375000000000006
```

[3]:
```
#Alternatively, we can use Cumulative probability
binom.cdf(2,6,0.5)
```

[3]:
```
0.34375
```

## Tasks:

- What is the proability of getting exactly 6 heads?
- What is the probability of getting 5 or more heads?

Example:

Assume that there is a road junction and several vehicles passes by everyday. Some drivers will ask for the direction at that junction. The probability of a driver seeking for direction is 0.45. On a given day, you observe 200 vehicles passing through the junction. Then, what is the probability that

- exactly 100 drivers will stop and ask for direction?
- atleast 50 drivers will ask for direction?
- at the most 20 drivers will ask for direction?

Markdown

1. Exactly 100 drivers will stop and ask for direction: n = 200, p =0.45, q = 0.55, and x=100 P(X=x) = ?

[4]:
```
#exactly 100 drivers will stop and ask for direction
# P(X=100)
binom.pmf(100,200,0.45)
```

[4]: 0.020625365698943285

[5]:
```
# atleast 50 stop and ask direction
# P(X>=50) = 1 - p(X<=49)
1- binom.cdf(49,200,0.45)
```

[5]: 0.99999999844398

[6]:
```
#at the most 20 drivers stop and ask direction
# p(X<=20)
binom.cdf(20,200,0.45)
```

[6]: 3.978917632494344e-27

## Let's work on some real data

[1]:
```
import pandas as pd
```

[2]:
```
camp=pd.read_excel('ProbDist_Data.xlsx',sheet_name='binom')
camp.head()
```

[2]:

| | Card_ID | Campaign_Responce | Registration_Date | Gender | Birth_Date |
|---|---|---|---|---|---|
| 0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |

Markdown

[2]:

| | Card_ID | Campaign_Responce | Registration_Date | Gender | Birth_Date |
|---|---|---|---|---|---|
| 0 | 100005950 | False | 1998-11-18 | M | 1984-02-06 |
| 1 | 100022191 | True | 1999-09-15 | F | 1959-09-11 |
| 2 | 100025442 | False | 1998-05-12 | M | 1970-08-25 |
| 3 | 100026513 | False | 1999-02-12 | M | 1951-03-12 |
| 4 | 100039145 | False | 2000-08-12 | M | 1949-06-08 |

[ ]:

[9]:
```
camp.shape
```

[9]:
```
(297, 5)
```

[10]:
```
camp['Campaign_Responce'].value_counts()
```

[10]:
```
False    248
True      49
Name: Campaign_Responce, dtype: int64
```

[13]:
```
# Check the probability of 'True' and 'False' by dividing respective counts by total number of records.
# That is, Prob(True) =49/297
# and Prob(False) = 248/297
49/297
```

[13]:
```
0.16498316498316498
```

Search

15]:
```
# This can be achieved by adding a parameter to value_counts() function:
camp['Campaign_Responce'].value_counts(normalize=True)
```

15]:
```
False    0.835017
True     0.164983
Name: Campaign_Responce, dtype: float64
```

What is the probability that upto 15 customers will respond to campaign out of 150 randomly selected customers?

n=150, p= 0.165, P(X<=15)=?

12]:
```
binom.cdf(15,150,0.165)
```

12]:
```
0.016547176458314 22
```

What is the probability that between 15 and 20 customers will respond to campaign out of 150 randomly selected customers?

16]:
```
# P(15 <= X <=20) = P(0<= X <=20) - P(0<= X <=14)
binom.cdf(20,150,0.165) - binom.cdf(14, 150, 0.165)
```

16]:
```
0.16699220469937 01
```

# Poisson distribution

- It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.
- When the number of successes (p) are very small compared to the number of trials (n), then we go for Poisson distribution.
- Examples:
  - A call center receives an average of 180 calls per hour, 24 hours a day. The calls are independent; receiving one does not change the probability of when the next one will arrive. The number of calls received during any minute has a Poisson probability distribution: the most likely numbers are 2 and 3 but 1 and 4 are also likely.
  - The number of decay events that occur from a radioactive source during a defined observation period.
  - The number of road accidents in a given day at a particular road junction. Imagine number of vehicles passing by the junction per day. It is quite a huge number (like 1000, 10000 etc). But, the number of accidents observed per day will be 1-2 or at the max 10. So, this number is very small compared to actual number of vehicles that have passed the junction. Thus, this is a poisson distribution.
- If X follows poisson distribution with parameter $\lambda$ then probability mass function is given by

$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for all $x = 0, 1, 2, 3, 4, \ldots$

- Mean=variance=$\lambda$

```
from scipy.stats import poisson
import pandas as pd
```

Example:

- Find the Probability of exactly 4 accidents in a month, given that average is 5 accidents/ month.
- Solution: Here, $\lambda = 5$. We need to compute $P(X = 4)$

**Example:**

- Find the Probability of exactly 4 accidents in a month, given that average is 5 accidents/ month
- Solution: Here, $\lambda = 5$. We need to compute $P(X = 4)$

8]:
```
#The syntax of poisson.pmf is pmf(x, lambda)
poisson.pmf(4,5)
```

8]:
```
0.17546736976785063
```

- What is the probability of 3 or lesser accidents in a month?

6]:
```
#P(x<=3)=?
# use cumulative probability
poisson.cdf(3,5)
```

6]:
```
0.2650259152973616
```

6]:
```
#Alternatively,
poisson.pmf(0,5)+poisson.pmf(1,5)+poisson.pmf(2,5)+poisson.pmf(3,5)
```

6]:
```
0.26502591529736164
```

- What is the probability that 5 or more accidents in a month?

```
# Pr(X>=5) =?

1- poisson.cdf(4,5)
```

```
0.5595067149347874
```

Let's work on some real time dataset

```
cart=pd.read_excel('ProbDist_Data.xlsx',sheet_name='poisson', usecols = [0])
cart.head()
```

| | Cart Addition |
|---|---|
| 0 | 1 |
| 1 | 1 |
| 2 | 9 |
| 3 | 1 |
| 4 | 1 |

```
cart.shape
```

```
(49, 1)
```

```
cart.mean()
```

```
Cart Addition    1.44898
```

- What is the probability of seeing atleast 2 items in the cart?
- Solution: Compute $P(X >= 2)$

```
# P(x>=2)= 1-P(x<2)
# P(x>=2) = 1- {P(X=0)+P(X=1)}

1- poisson.cdf(1,1.44898)
```

```
0.42495580982638104
```

- What is the probability of seeing 6 to 9 items in the cart?
- Solution: Compute $P(6 \leq X \leq 9)$
- If we use cumulative distribution function, then $P(6 \leq X \leq 9) = P(X \leq 9) - P(X \leq 6)$

```
poisson.cdf(9,1.45) - poisson.cdf(5,1.45)
```

```
0.0037871655781491764
```

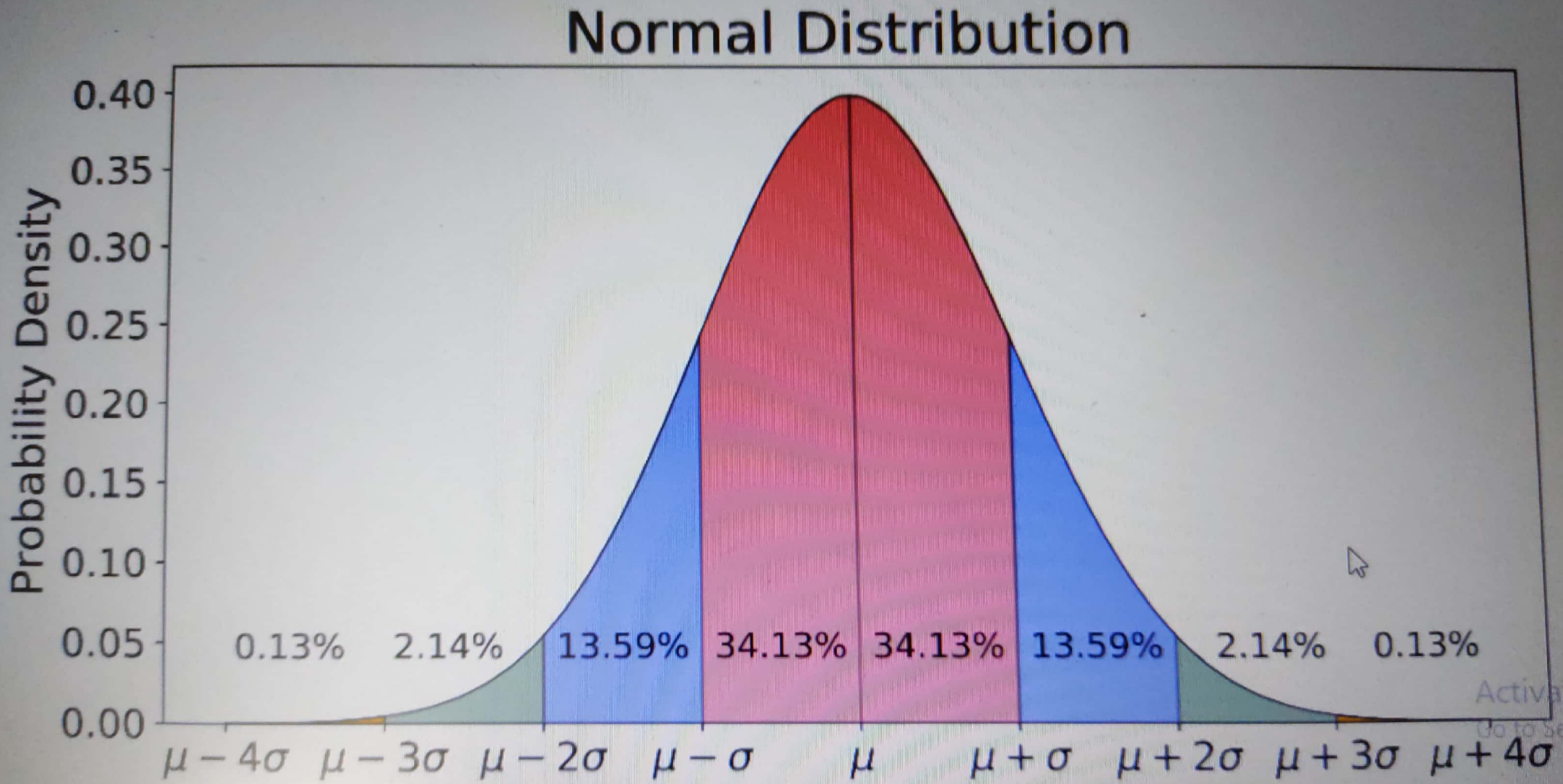
## Normal distribution:

- It is a continuous probability distribution for a real-valued random variable.
- The probability density function of Normal distribution with parameters $\mu$ and $\sigma$ is given by -

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Examples:

- Marks scored by students in an exam
- Sales of car in an year



Edit View Run Kernel Settings Help

Markdown

JupyterLab

**Empirical Rule**

- Approximately 68% of all observations fall within one standard deviation of the mean.
- Approximately 95% of all observations fall within two standard deviations of the mean.
- Approximately 99.7% of all observations fall within three standard deviations of the mean.

**Example:**

- Let the average sales of a particular product is 10000 and standard deviation is 2400. Then what is the probability of getting more than 12000 sales?

- Solution: Given that $\mu = 10000$ and $\sigma = 2400$. We need to compute $P(X >= 12000)$

11]:
```
from scipy.stats import norm
1 - norm.cdf(12000,10000,2400)
```

11]: 0.20232838096364314

**Business Inference:**

- When we know that average sales is 10000 with a standard deviation of 2400, then the chances of having more than 12000 sales is around 20.23%

- What is the probability of getting fewer than 9200 sales?

[2]:
```
# P(X<=9200)
norm.cdf(9200,10000,2400)
```

[2]: 0.36944134018176367

Search

- What is the probability of getting between 9000 to 12000 sales?

```
#P(x<=12000)-P(x<=9000)
norm.cdf(12000,10000,2400) - norm.cdf(9000,10000,2400)
```

```
0.4592104995256672
```

**Risk measurement using Normal distribution**

- Consider a stock with returns mean=10,std=5. What are the chances that the returns will be < 0?

```
norm.cdf(0,10,5)
```

```
0.022750131948179195
```

**Business Inference:**

- Given that, as per the past performance, the returns on the stock is 10% with the standard of deviation of 5% What is the chance of you loosing the money (returns is less than 0%).
- The chance of you loosing your money is 0.023. That is, it is around 2.3%

- What are the chances of losing money when std=10%?

```
norm.cdf(0,10,10)
```

```
0.15865525393145707
```

As standard deviation of the distribution increases the risk of losing money increases.

**Example:**

A monthly balance in the bank account of credit card holders is assumed to be normally distributed with mean 500USD and variance 100USD.

- What is the probability that the balance can be more than 513.5USD?
- If there are 1000 customers,
  - how many customers will have account balance more than 513.5 USD?
  - How many people will have less than 520 USD as their account balance?

3]:
```
mu= 500
var= 100
sd= 10
```

4]:
```
## p(x>=513.5) =?
1 - norm.cdf(513.5,mu,sd)    # 8.85% chance
```

4]:
```
0.08850799143740196
```

2]:
```
N= 1000
round(N* (1-norm.cdf(513.5,mu,sd)),0)
```

2]:
```
89.0
```

[13]:
```
mu= 500
var= 100
sd= 10
```

[14]:
```
## p(x>=513.5) =?
1 - norm.cdf(513.5,mu,sd)    # 8.85% chance
```

[14]:
```
0.08850799143740196
```

[12]:
```
N= 1000
round(N* (1-norm.cdf(513.5,mu,sd)),0)
```

[12]:
```
89.0
```

[15]:
```
round(N*norm.cdf(520,mu,sd),0)
```

[15]:
```
977.0
```

[ ]: