

Disclaimer: Please note that any topics that are not covered in today's lecture will be covered in the next lecture.

Content

- Introduction
- Log Normal Distribution.
 - Real life dataset
 - Key Characteristics of a Log-Normal Distribution
- Poisson Distribution
 - Application
 - Rules of Poisson Distribution
 - Poisson approximation to Binomial

✓ Log Normal Distribution

Imagine that you are a data scientist at Amazon/Swiggy/Zomato
You've collected a bunch of data on delivery times.

Generally how much time delivery takes?

- Let's assume around 30 mins, maybe sometimes less than 30 maybe more

Now, if we take thousands of these delivery time data points and plot a histogram,

- It may be a bit skewed to the right. Sometimes deliveries are quicker than 30 minutes, and sometimes they take a bit longer.

The lognormal distribution is a continuous probability distribution that models this type of right-skewed data.

Suppose X is the actual data

- Now the beauty of log normal is when you take **the logarithm (log) of the actual delivery time data** and plot a new histogram,
- The new histogram tends to be more symmetrical, like a bell curve.

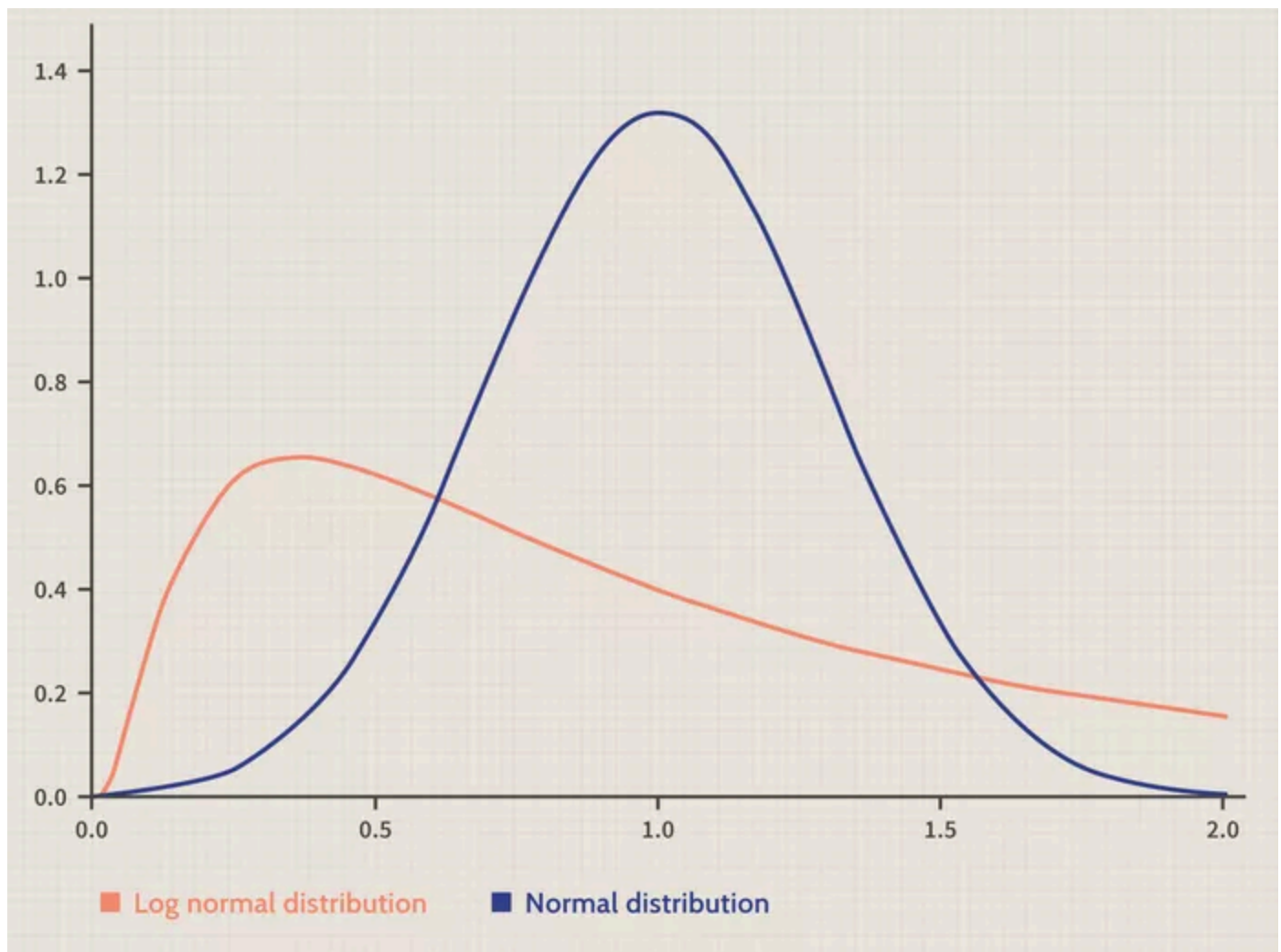
In simple terms, **the Log-Normal Distribution takes the original data, does some math (logarithm), and makes it look more like a normal, symmetrical distribution.**

So, in the language of distributions, we say the "**original delivery time data (X) is log-normal**"

- It means if X follows a log-normal distribution, $\log(X)$ follows a normal (bell-shaped) distribution.

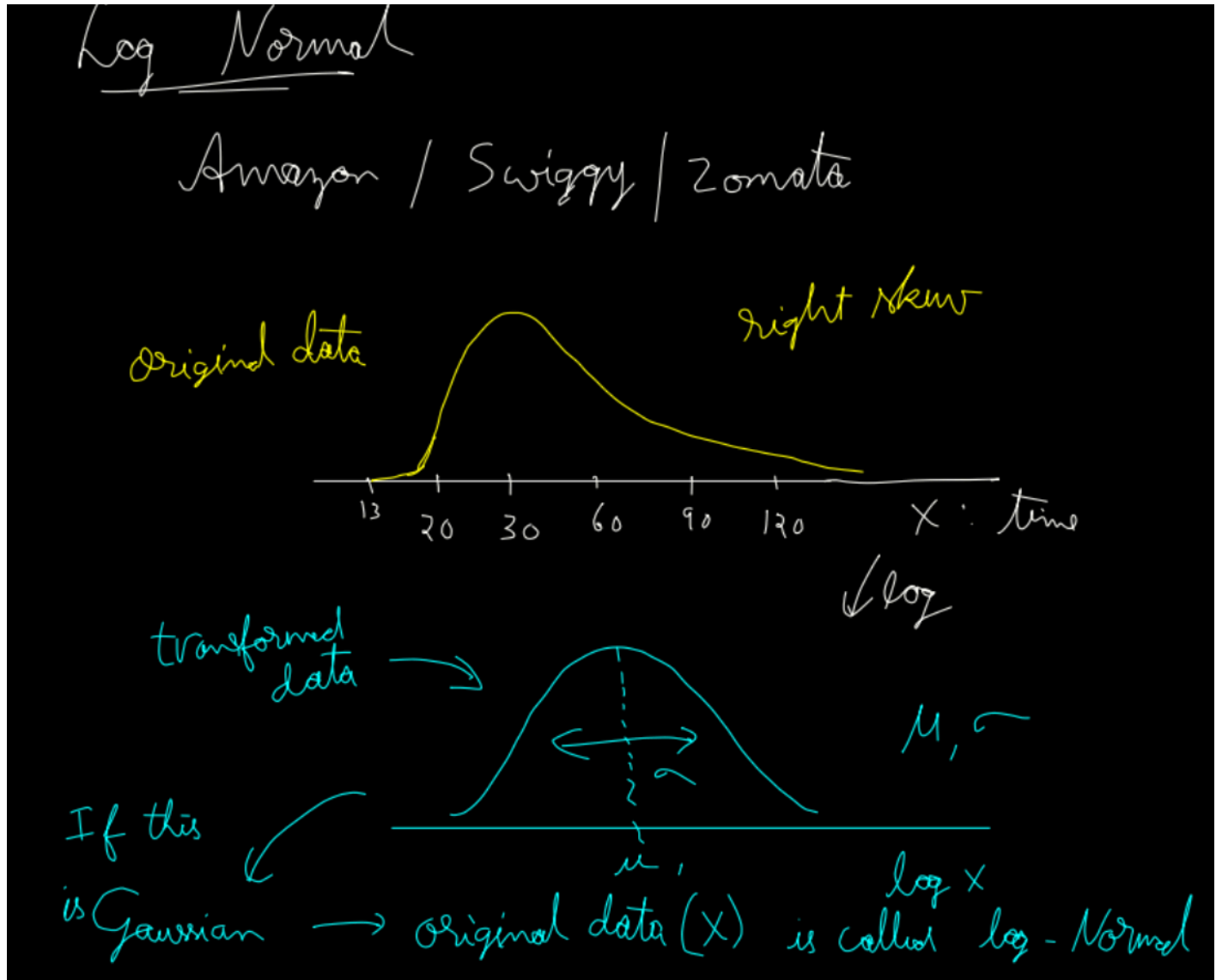
You can **exponentiate a normal distribution ($\exp(X)$) to obtain the lognormal distribution.**

In this manner, you can transform back and forth between pairs of related log normal and normal distribution



We can see in this image that the original data follows log normal distribution and if we take log of this distribution, it'll look more symmetrical like a bell shaped curve.

We will implement this on a real life dataset



✓ Real Life dataset

Let's have a look into the dataset which has waiting time records

```
!wget --no-check-certificate https://drive.google.com/uc?id=1SIZC
```

```
--2024-01-18 10:35:32-- https://drive.google.com/uc?id=1SIZC1FZvZAhVzRvnZ7IFWBl
Resolving drive.google.com (drive.google.com)... 74.125.31.113, 74.125.31.102, 7
Connecting to drive.google.com (drive.google.com)|74.125.31.113|:443... connecte
HTTP request sent, awaiting response... 303 See Other
Location: https://drive.usercontent.google.com/download?id=1SIZC1FZvZAhVzRvnZ7IFWBl
--2024-01-18 10:35:32-- https://drive.usercontent.google.com/download?id=1SIZC1FZvZAhVzRvnZ7IFWBl
Resolving drive.usercontent.google.com (drive.usercontent.google.com)... 173.194.100.100
Connecting to drive.usercontent.google.com (drive.usercontent.google.com)|173.194.100.100|:443...
HTTP request sent, awaiting response... 200 OK
Length: 1656272 (1.6M) [application/octet-stream]
Saving to: 'waiting_time.csv'
```

```
waiting_time.csv 100%[=====>] 1.58M --.-KB/s in 0.01s
```

```
2024-01-18 10:35:32 (140 MB/s) - 'waiting_time.csv' saved [1656272/1656272]
```



Importing Libraries

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
from scipy.stats import poisson, binom
```

```
data = pd.read_csv("/content/waiting_time.csv")
data.head()
```

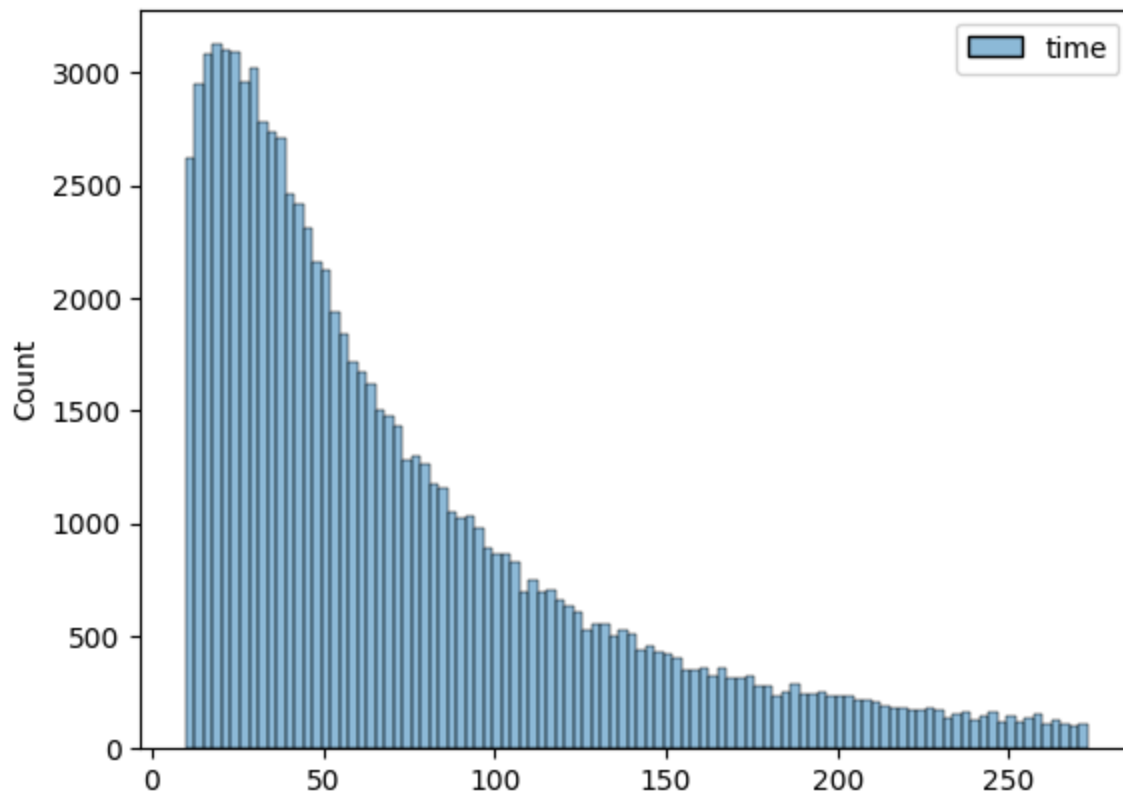
```
↔
```

	time
0	184.003075
1	36.721521
2	29.970417
3	75.640285
4	61.489439

Let's plot this data

```
sns.histplot(data, bins=100)
```

↔ <Axes: ylabel='Count'>



Observation

We can observe that it is right skewed.

Now,

Q1. How can we answer questions related to the data which is distributed in this way?

We can transform this data using a **log** and let's see the distribution of transformed data.

✓ Log Normal Distribution Parameters

As we know, the random variable for the original data is X and after transforming it using log, the random variable of transformed data is $\ln(x)$.

If you have the **mean (μ) and standard deviation (σ) of the natural logarithm** of a random variable X and you want to **find the mean and standard deviation of the original random variable X** (which follows a log-normal distribution), you can use the following relationships:

- **Mean of original** $X = \exp\left(\mu + \frac{\sigma^2}{2}\right)$

$$m = e^{\mu + \frac{\sigma^2}{2}}$$

- **Variance of original** $X = [\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2).$

$$\text{Var}[X] = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$

Let's transform our original data:

```
data_log = np.log(data)
```

```
data_log
```

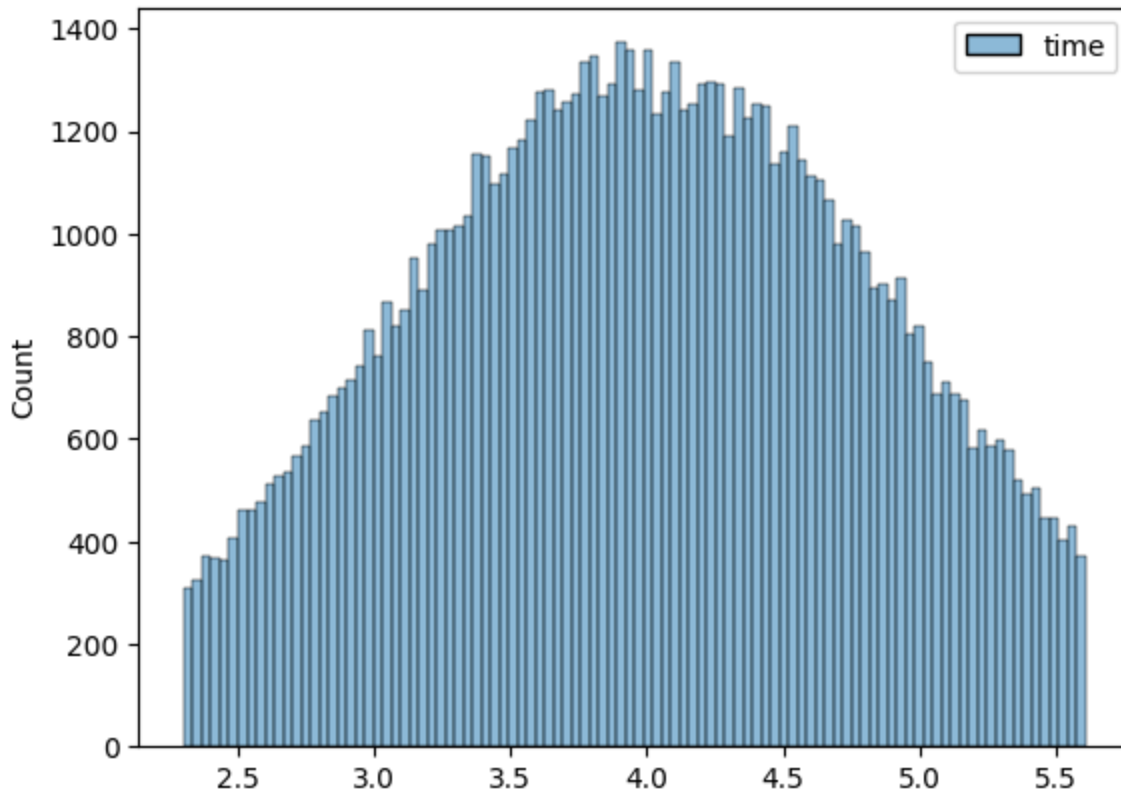


	time
0	5.214952
1	3.603363
2	3.400211
3	4.325989
4	4.118865
...	...
90041	4.911816
90042	2.722871
90043	5.336766
90044	4.945125
90045	3.926311

90046 rows × 1 columns

```
sns.histplot(data_log, bins=100)
```

```
<Axes: ylabel='Count'>
```



Observation

- We can observe that after applying logarithm to the right skewed data, we get the distribution which is approximately normal.
- We converted our data in such a format such that we are able to utilize the properties of gaussian distribution

This is known as **log normal transformation**

Q1. Why did we specifically choose log?

1. Symmetry:

- Our original data is right-skewed, with a long tail on the right side indicating occasional very long delivery times.
- The **logarithmic transformation** compresses larger values more than smaller values.
 - The extreme right tail is pulled in, making the distribution more symmetric.

2. Stabilizing Variance:

- In the original delivery time data, you might observe that the variance (spread) of delivery times increases as the mean delivery time increases.
- Taking the logarithm can **stabilize the variance**.

We can observe that the spread of delivery times after the transformation is more consistent across different independent variables.

In summary,

Applying a logarithmic transformation to the right skewed data can make the distribution more symmetric and stabilize the variance, making it potentially more useful for certain statistical analyses.

Let's understand this with the help of an example:

Suppose we have values like,

X: 1, 10, 100, 1000, 10000

Now let's take a log of all these values, we will get:

- $\ln(1) = 0$
- $\ln(10) = 2.30$
- $\ln(100) = 4.60$
- $\ln(1000) = 6.90$
- $\ln(10000) = 9.21$

Observation:

- We can clearly observe that after taking log of all the values it compresses larger values more than smaller values.
 - 10,000 got transformed into 9.21 and we can clearly see how much compressed the values got.
 - This can bring symmetry to the distribution.
- We can also observe that after applying the log, variance also got stabilized.

Example on stabilizing variance:

We can also observe that after applying the log, variance also got stabilized.

- Let's consider a simple example to illustrate this:

Suppose you have a set of positive numbers with increasing variance: Original Data:

1, 2, 4, 8, 16, 32, . . .

- If you observe the differences between consecutive values, you'll see that the differences increase:

Differences: 1, 2, 4, 8, 16, . . .

Now, if you take the logarithm of the original data:

- Log-Transformed Data: $\ln(1), \ln(2), \ln(4), \ln(8), \ln(16), \dots$

The differences between the log-transformed values are now constant around 0.693

Differences:

$\ln(2) - \ln(1), \ln(4) - \ln(2), \ln(8) - \ln(4), \ln(16) - \ln(8), \dots$

This constant difference suggests a stabilized variance, which can be beneficial in statistical analyses.

Key Characteristics of a Log-Normal Distribution

Let's understand the key characteristics of a log-normal distribution.

1. Positivity:

- All values in a log-normal distribution are **positive because the logarithm of any positive number is always real**.
- Eg. Let's say the original value X is -1.5 then the log normal distribution value will be e^X which is $e^{1.5} = 0.223$, this comes out to be positive.

2. Skewedness:

- If the original data is right-skewed, the log-normal transformation can make it more symmetric and bell-shaped.

3. Multiplicative Processes:

- Log-normal distributions are suitable for modelling scenarios where the final outcome is influenced by the product of independent factors.
- In our dataset, we are aware that **delivery times may get affected by various independent factors like traffic, order processing time, etc.**

In summary, a log-normal distribution is a good fit for positively skewed, ensuring **positivity, and aligning with multiplicative processes** often seen in real-world scenarios.

Now, let's see what is poisson distribution

✓ Poisson Distribution

Scenario: Traffic at a Toll Booth

Imagine you're at a toll booth on a highway, observing the number of vehicles passing through the toll booth in a given time period.

The Poisson distribution comes into play when we want to **model the number of events that occur in a fixed interval of time or space**.

- In this case, **vehicles passing through the toll booth** are our event.



Explanation:

The Poisson distribution is a **discrete probability distribution** particularly useful when dealing with events that occur randomly and independently, but with a known average rate.

In our toll booth scenario, we can make a few key observations:

1. **Fixed Interval:**

- Let's say we want to study the number of vehicles passing through the toll booth in a specific time period, say 1 hour.

2. Average Rate:

- We have an average rate of vehicles passing through the toll booth, let's say 30 vehicles per hour.
- It is denoted as λ (lambda), which represents the average rate of occurrence of the event within a given interval.
- Here, λ is 30 vehicles per hour.

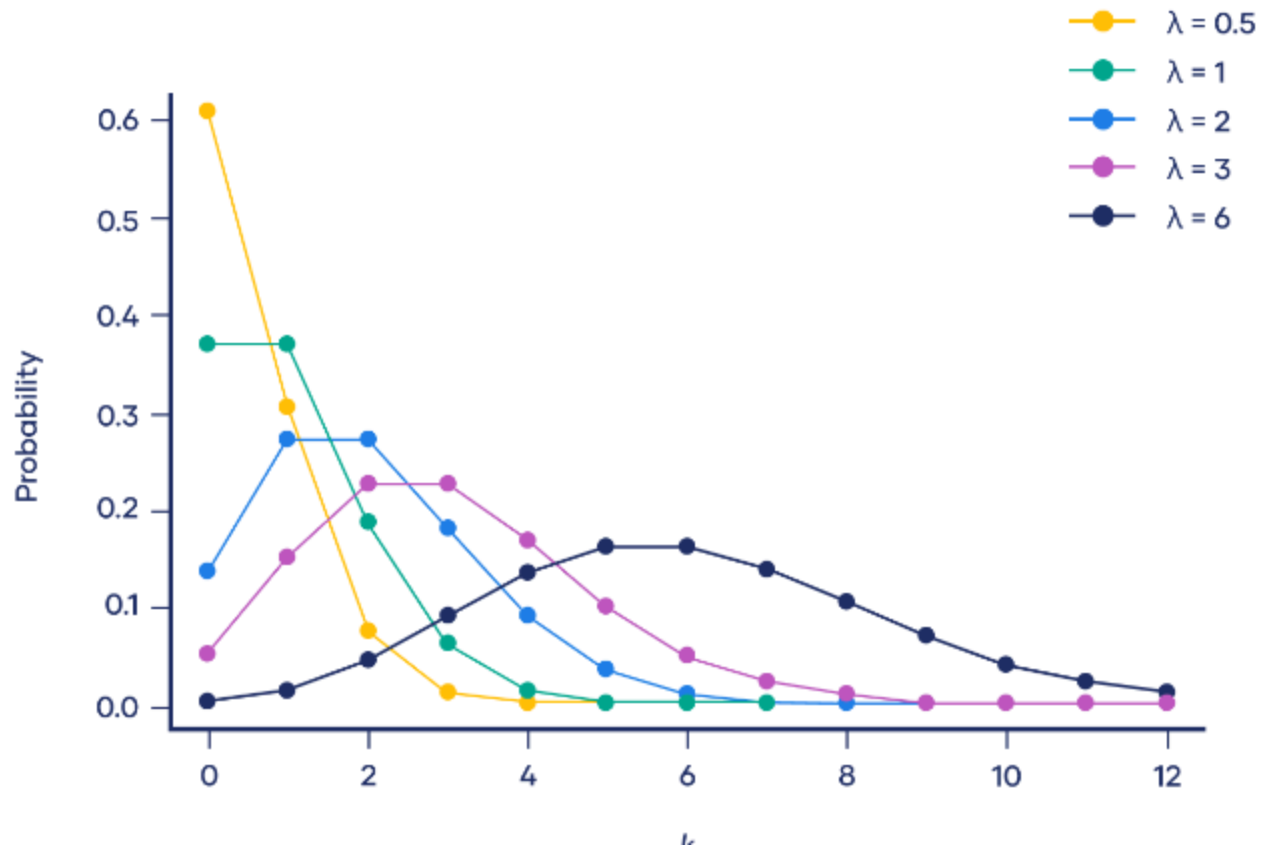
Now, the poisson distribution **helps us answer some questions** like:

Q1. What is the probability of exactly 25 vehicles passing through the toll booth in the next hour?

Q2. What is the probability of more than 40 vehicles passing through the toll booth in the next hour?

This toll booth scenario is just one example of how the Poisson distribution is applied in various fields.

The graph below shows examples of Poisson distributions with different values of λ .



- When λ is low, the distribution is much longer on the right side of its peak than on its left.
- As λ increases, the spread of distribution also increases
 - If you keep increasing, the distribution looks more and more similar to a normal distribution.

Poisson Distribution Formula

If a random variable X follows a Poisson distribution, then the probability that $X = k$ successes can be found by the following formula:

$$P[X = k] = \frac{\lambda^k * e^{-\lambda}}{k!}$$

where:

- λ : rate or mean number of successes that occur during a specific interval
- k : number of successes
- e : a constant equal to approximately 2.71828

This is also known as **Probability Mass Function (PMF)** of poisson distribution as using this formula we can calculate the probability of exact events.

✓ Example 1:

suppose a particular hospital experiences an average of 2 births per hour.

We can use the formula above to determine

Calculate the probability of experiencing 0, 1, 2, 3 births, etc. in a given hour:

Here, rate (λ) = 2

e = constant= 2.71828

$$P[X = 0] = \frac{2^0 * e^{-2}}{0!} = 0.1353$$

$$P[X = 1] = \frac{2^1 * e^{-2}}{1!} = 0.2707$$

$$P[X = 2] = \frac{2^2 * e^{-2}}{2!} = 0.2707$$

$$P[X = 3] = \frac{2^3 * e^{-2}}{3!} = 0.1804$$

Here, we can also find the probability using the function in python i.e. **poisson.pmf()** as it is asking for the probability of an exact value

- We have to pass 2 parameters in this function, **k**: number of events and **mu**: average or rate

P[x=0]

```
poisson.pmf(k=0, mu=2)
```

```
↩ 0.1353352832366127
```

P[x=1]

```
poisson.pmf(k=1, mu=2)
```

```
↩ 0.2706705664732254
```

```
# P[x=2]
```

```
poisson.pmf(k=2, mu=2)
```

```
↔ 0.2706705664732254
```

```
# P[x=3]
```

```
poisson.pmf(k=3, mu=2)
```

```
↔ 0.18044704431548356
```

Let's look into more examples.

Applications:

1) Football Match Goals

Imagine we have collected data for all the football matches ever happened, now we want to analyze the distribution of goals.

- We observe that the average goal per 90 mins match is 2.5
 - So the rate will be 2.5 goals per match ($\lambda = 2.5$).

Q. If I want to know the probability of getting 1 goal in last 30 mins?

This is where poisson distribution comes into play.

Here, the rate is 2.5 goal/ 90 mins (per match) which mean average number of goals in 90 mins

- What will be the average number of goals in 45 mins?

2.5 goals -> 90 mins

Average goals for half of the time will be half of the total average rate

Rate: $2.5/2 = 1.25/45$ mins (per 45 mins)

Similarly, we can define a range for 30 mins,

1.25 goals -> 45 mins

x goals? -> 30 mins

$$x = (30 * 1.25)/45$$

$$\text{Rate} = 0.833 \text{ goals}/30 \text{ mins}$$

So,

Q1. How long should you stay to witness a goal on average?

- On average, **staying at least 45 minutes increases the probability of witnessing a goal** during a football match.
- Because staying at least 45 minutes aligns with the average goal rate of 1.25 goals per 45 minutes.
- This duration maximizes the likelihood of experiencing a goal during a football match based on the observed rate of scoring.

Next example,

2) Support Phone Calls

Think about a support centre that receives 100 calls per hour.

- So the average call received per minute will be,

$$100 \text{ calls} \rightarrow 60 \text{ mins (1 hr)}$$

$$x \text{ calls} \rightarrow 1 \text{ min}$$

$$\text{Rate: } 100/60 = 1.666 \text{ calls/min}$$

This allows us to analyze the probability of receiving a certain number of calls within a specific time frame.

- The call center management can use this rate to determine the optimal number of customer service representatives to have on duty during different time periods.
- For instance, during peak times, when the call rate is high, more staff may be required to handle the increased volume.

One more example

3) Hospital OPD Patients

Consider a hospital's Outpatient Department (OPD) where, on average, 200 patients visit in a day ($\lambda = 200$).

- The average hourly rate of patient arrivals can be calculated by dividing the daily rate by the number of working hours.
- For example, if the facility operates for 8 hours, the hourly rate would be $\frac{200}{8} = 25$ patients per hour.
- The facility can use this information for resource planning, such as determining the optimal number of staff, doctors, and examination rooms needed to handle the expected patient load efficiently.

These are some real life examples where poisson distribution can help us understand the likelihood of an event occurring in a specific time interval or space

✓ Rules of Poisson Distribution

Key rules that govern the Poisson distribution:

1. Counting:

- The Poisson distribution is tailored for **counting the number of discrete events happening within a fixed interval**
- The events can take on values like 0, 1, 2, 3, and so on.

2. Independence:

- The occurrence of one event should not affect the occurrence of another event.
- Events are considered to be independent if the probability of one event happening doesn't change based on whether another event has occurred.

For example,

- **if an accident occurs in Delhi at 4 PM, it will have no impact on the occurrence of an accident in Mumbai at the same time.**
- Each event is independent of the other, and the outcome in one location does not influence or affect the outcome in the other location.

3. Rate (λ or μ):

- The distribution is defined by a single parameter often denoted as λ (lambda) or μ (mu), which represents the average rate of occurrence of the event within the given interval.
- This rate remains constant throughout the interval and doesn't change based on the occurrences.

4. No Simultaneous Events:

- The Poisson distribution assumes that there cannot be more than one occurrence of the event at exactly the same time or within an infinitesimally small interval of time or space.
- For instance, if a family of five people enters a store, it's counted as a single event, not five separate events.
- Another example, two goals can't be scored at a same time

Let's look at the some examples using Poisson distribution

✓ Example 2:

A city sees 3 accidents per day on average.

Find the probability that there will be 5 accidents tomorrow.

Solution:

Given,

The rate is given as 3 accidents per day on average,

- $\lambda = 3$

Let " X " denote the number of accidents tomorrow.

- We say " X " is Poisson distributed with rate (λ) = 3

So, the probability that there will be 5 accidents tomorrow is $P[X = 5]$

By using the formula,

- $$P[X = 5] = \frac{\lambda^5 * e^{-\lambda}}{5!} = \frac{3^5 * e^{-3}}{5!}.$$

Using python,

```
# P[X=5]
poisson.pmf(k=5, mu=3)

↔ 0.10081881344492458
```

There is a 10% chance that there will be 5 accidents tomorrow.

Next question

Q1. Find the probability that there will be 5 or fewer accidents tomorrow?

Here we want to calculate $P[X \leq 5]$,

We will use **poisson.cdf()** here as we want to calculate cumulative probability.

$$P[X \leq 5] = P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3] + P[X = 4] + P[X = 5]$$

We can directly find it using **poisson.cdf()**

```
# P[X ≤ 5]
poisson.cdf(k=5, mu=3)

↔ 0.9160820579686966
```

✓ Example 3:

Let “X” be the number of typos in a page in a printed book, with mean of 3 typos per page. What is the probability that a randomly selected page has atmost 1 typo?

Here, rate (λ) = 3

we want to find for atmost 1 type, so we need to find

$$P[X \leq 1] \text{ which will be } P[X = 0] + P[X = 1].$$

We can directly use **poisson.cdf()** here

```
# P[x≤1]
poisson.cdf(k=1, mu=3)

↔ 0.1991482734714558
```

```
prob = poisson.pmf(k=0, mu=3) + poisson.pmf(k=1, mu=3)
prob
```

⇒ 0.1991482734714558

There is a 19% chance that a randomly selected page has atmost 1 typo

✓ Poisson approximation to Binomial

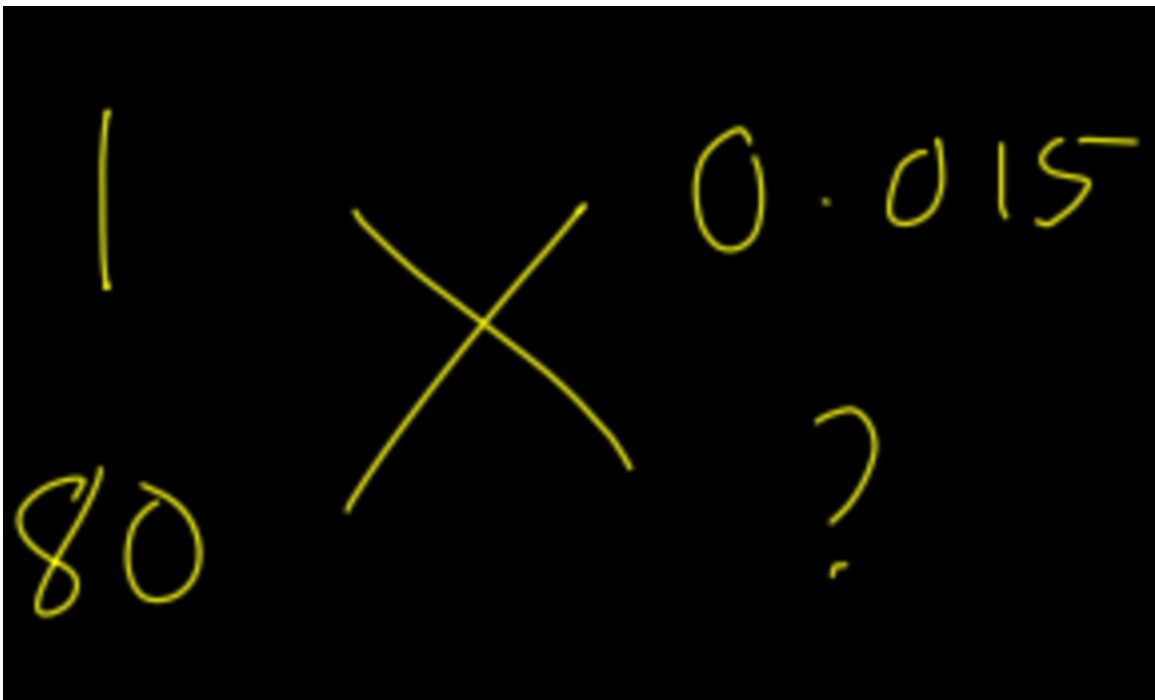
There are 80 students in a kinder garden class.

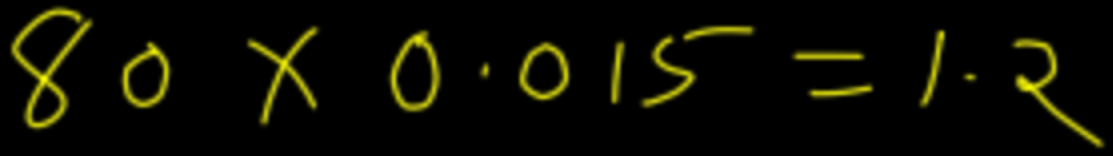
Each one of them has 0.015 probability of forgetting their lunch on any given day.

- (a) What is the average or expected number of students who forgot lunch in the class?
- (b) What is the probability that exactly 3 of them will forget their lunch today?

Solution:

First question,





$$80 \times 0.015 = 1.2$$

```
rate = 80*0.015 # total rate
```

```
rate
```

```
⇒ 1.2
```

Conclusion:

This implies that, on average, there are 1.2 students who forget their lunch in a given period.

(b) What is the probability that exactly 3 of them will forget their lunch today?

here, $k = 3$ and $\lambda = 1.2$

We can directly use `poisson.pmf()`

```
poisson.pmf(k=3, mu=1.2)
```

```
⇒ 0.08674393303071422
```

There is 8.67% chance that exactly 3 of them will forget their lunch today

Q. Can I model this question into binomial distribution?

We have 80 students, we can define two probabilities here

- probability of success $P(s) = \text{student forgot the lunch} = 0.015$
- probability of failure $P(f) = 1 - P(s) = 1 - 0.015$

We want $P[X = 3]$,

- we can represent it as **out of 80 trials, I want 3 success**

Using binomial formula, it will be

- ${}^{80}C_3(0.015)^3(1 - 0.015)^{77}$

we just make this question of binomial

```
binom.pmf(k=3, n=80, p=0.015) # Large n, small p, np=mu
```

```
⇒ 0.08660120920447557
```

We got the similar answers using both **poisson and binomial**.

In binomial

- We are counting the number of successes in n trials where $P(s) = p$

In poisson

- Counting number of occurrences in a given time interval.
- Now, for 1 success we have probability p so for n success, the probability will be

1 success -> p

n success -> ?

for n success -> np

Here, the $P(s)$ for 1 student is 0.015. So $P(s)$ for 80 students will be $80 * 0.015 = 1.2$

From this, we can observe that $\lambda = np$

This approximation is known as the **Poisson approximation to the binomial distribution**

✓ Conditions for a reasonable approximation:

- The binomial distribution converges towards the Poisson distribution as the number of trials (n) goes to infinity while the product np converges to a finite limit.
- Therefore, the Poisson distribution with parameter $\lambda = np$ can be used as an approximation to $B(n, p)$ of the binomial distribution **if n is sufficiently large and p is sufficiently small**

- For a reasonable approximation:
 - This approximation is good if $n \geq 20$ and $p \leq 0.05$ such that $np \leq 1$,
 - or if $n > 50$ and $p < 0.1$ such that $np < 5$,
 - or if $n \geq 100$ and $np \leq 10$.

The concept of "large enough" for the number of trials (n) is not fixed