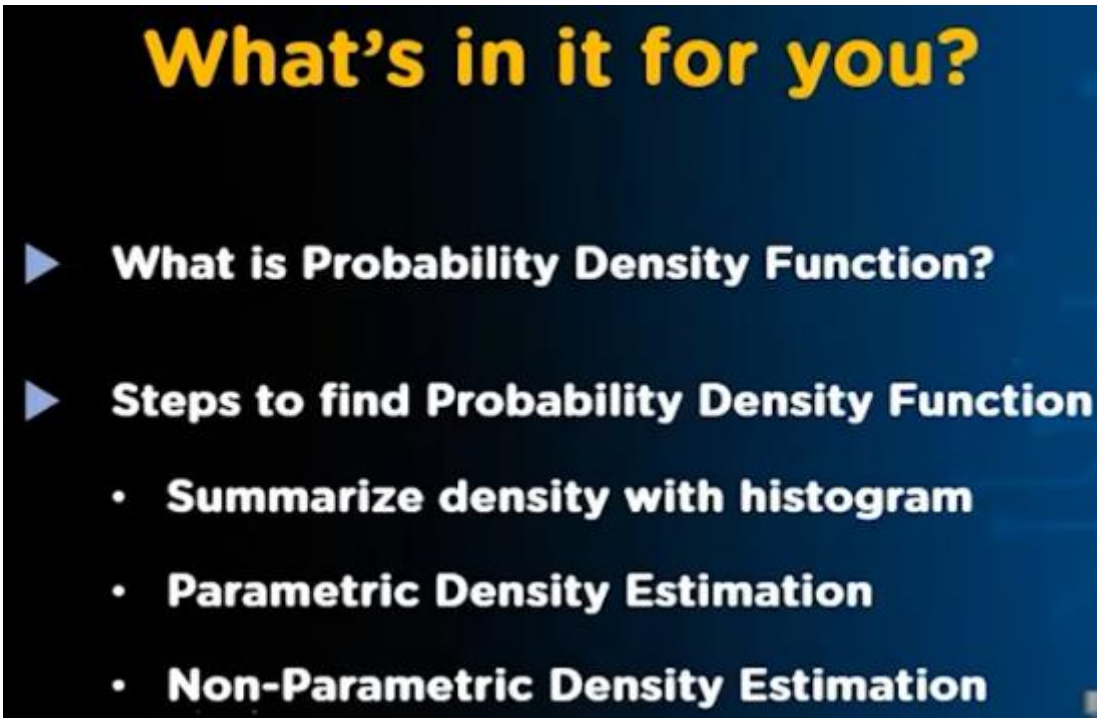
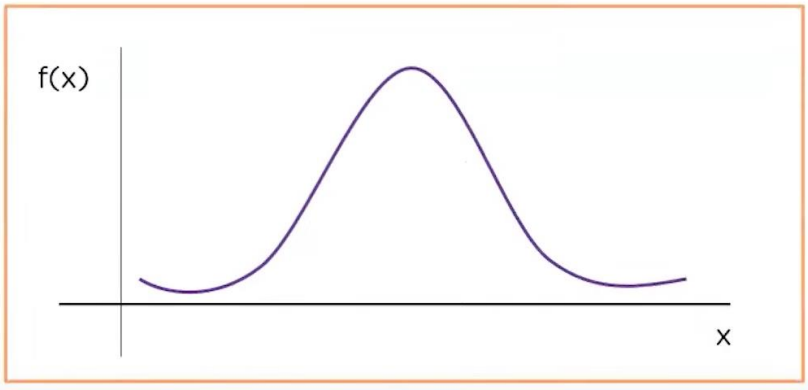
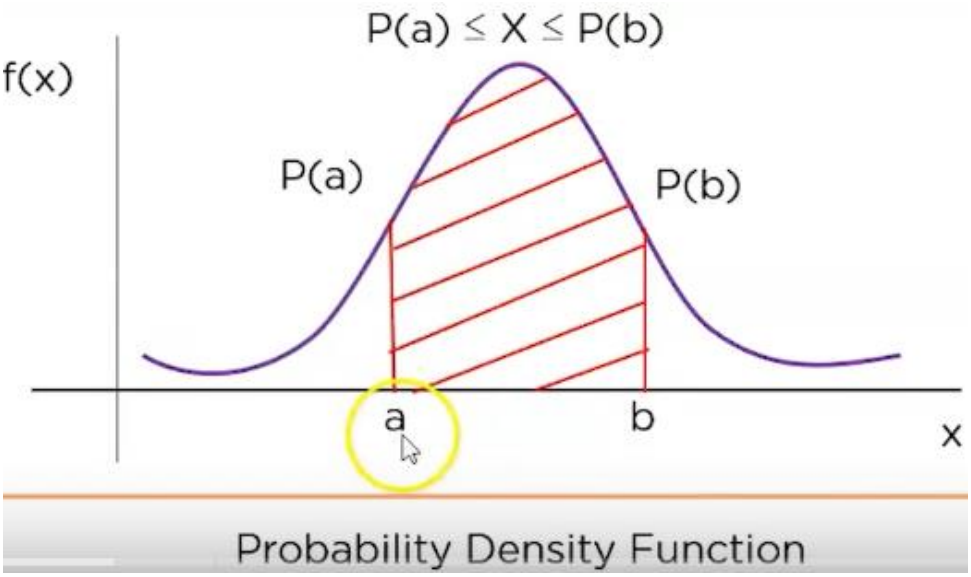


A function which defines the relationship between a random variable and its probability, such that you can find the probability of the variable using the function is called a Probability Density Function (PDF). Consider a variable x with a pdf of f(x)



Probability Density Function

Movements of the Distribution: **Mean and Variance**.



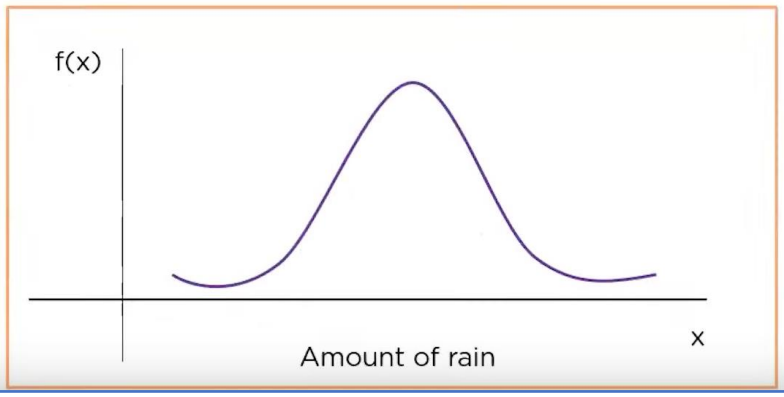
Discrete Random Variable

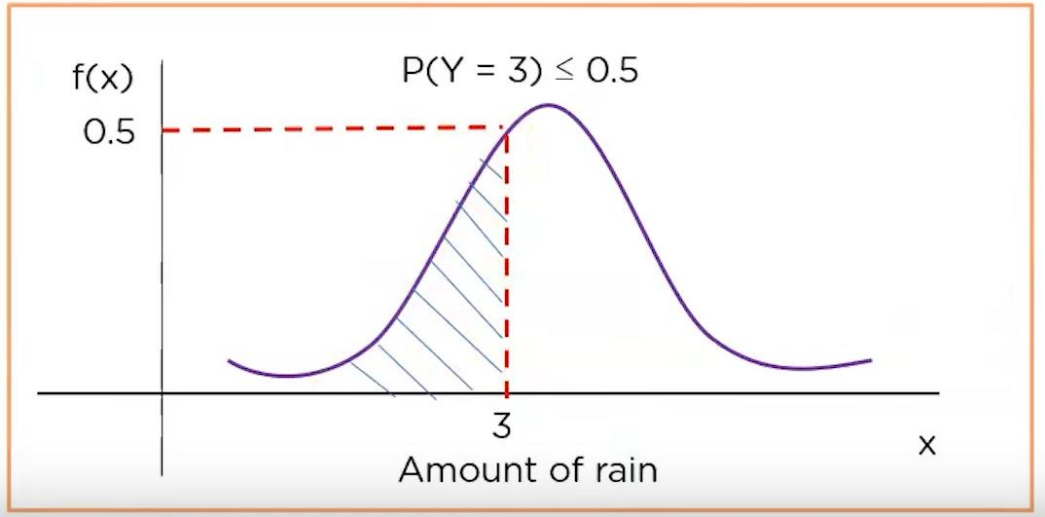
A variable which can only take in a value within a certain range. The value is usually within a certain distance of another finite value, eg : Sum of two dice

Continuous Random Variable

A continuous random variable is one which can take on infinite different values within a range of values, eg : Height of a student

Consider the graph below which shows the rainfall distribution in a year in a city. The x-axis has the rainfall in inches and the y-axis has the probability density function. The probability of some amount of rainfall is obtained by finding area of curve on the left of it



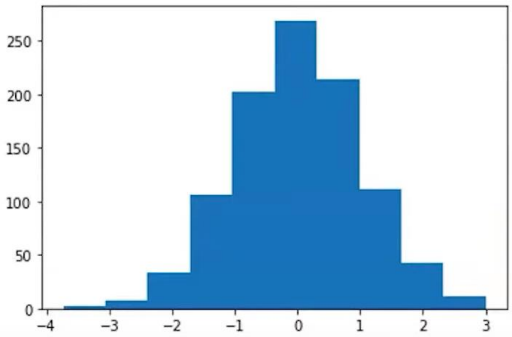


The exact probability can be found from the area of the curve which falls to the left of the tree.

Steps to find the PDF:

1. Summarize Density with Histogram

We first convert the data into discrete form by plotting it as a histogram. A histogram is a graph which has categorical values on x-axis and bins of different heights which give us a count of the values in that category



Histogram

```
In [1]:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

```
In [2]:
df = sns.load_dataset('tips')
df.head(3)
```

Out[2]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3

```
In [4]:
n = len(df)
prob_smoker = (df['smoker'] == 'Yes').sum() / n
prob_male = (df['sex'] == 'Male').sum() / n
```

```
In [5]:
prob_smoker
```

```
Out[5]: 0.38114754098360654
```

```
In [7]:
df['smoker']
```

```
Out[7]: 0    No
        1    No
        2    No
        3    No
        4    No
        5    No
        6    No
```

```
In [8]:
df['smoker'] == 'Yes'
```

```
Out[8]: 0    False
        1    False
        2    False
        3    False
        4    False
        5    False
        6    False
```