**Discrete Distributions:**

These are the type of distributions which has discrete values. this value can be whole numbers. And its not necessary to be related with other values.

Eg-

Roll of dice: here the die has 6 faces, so the probability of each face is 1/6.

**Uniform continues Distribution:**

Here the values are related to each other in form of range the probability of continues dist is related to are of distribution under the curve.

Eg-

Time interval of a bus stop

**Binomial Distribution:**

Here the outcome can be of just 2 types 1 or 0. Here the values should be independent.

Eg-

Tossing of coin

**Normal Distribution:**

Here the data is distributed such that the curve forming will be symmetrical and the mean is the line passing through it.

Mean

Here we can see that our data displays an symmetrical curve and mean is in middle of the curve.

In normal distribution if the standard deviation is high then the height of the curve decreases and vise versa.

Cumulative distribution frequency(cdf) will take the area of curve and will find probability according to it. For example we can use it to find the probability of men’s height is lesser than 180 cm from given distribution code- norm.cdf(180,150,5) 150-> mean and 7-> standard deviation

A percentage point function(ppf) takes percentage and provides the accurate point on the graph at which a certain percentage matches. For example we if we want to get height of a person who is more than 90% of the sample data then we can find if by using norm.ppf(0.9,150,7) 150-> mean and 7-> standard deviation. Here the 90% is cdf don’t think 90% of data.

We can also generate random values so that we can compare it with real data and come to a conclusion. We can do it using norm.rvs(150,7, size=10) 150-> mean and 7-> standard deviation, and size gives no of samples .

**Poisson Distribution:**

This Distribution can be used to predict the how many times a event can occur over a period of time.

For Example- let avg no of customers arriving on Saturday be around 400. Then Poisson distribution can predict what is the chance of getting more than or equal to 600 customers on that day.

We can use poisson.pmf to find probability of an even occurring

Poisson.pmf(600,400) 600->value to find probability of , and 400-> mean no of customers.

Same we can use cdf to find the probability of certain value less than or equal to.

Poisson.cdf(600,400) 600->value to find probability of , and 400-> mean no of customers.

For greater than we can subtract 1-lesser than value

We can also sample data by ourself using rvs

Poisson.rvs(400,size=10) 400->mean , and size specifies no of samples.

**Difference between poisson and normal distribution**

Poisson; this takes only discrete values as its mean for example if u roll a dice the outcomes are in whole numbers which are in between 1-6, but if u perform this test several times u may get different outcomes so to select a single outcome u take mean of it and calculate its probability.

Normal; this used on continues set of data u are working with sales data for a year then the data contain profit or loss is continues (as they are not any randomly picked no.s) so this data can be in float values and we use mean and standard deviation to find its probability.

**Exponential Distribution:**

This algo is used on continues data to find the time taken by a particular thing to occur

Eg:

It is used to find time taking to radioactive decay of a molecule given its mean decay and time.

Lets assume a phone booth is getting 4hrs of calls per day so we have to predict how many calls it will receive after 2 days.

First set scale=1/4 as we need to calculate it per hr

Than we can use pdf function expon.pdf(2,scale=0.25) where 2-> day to get avg calls, scale is avg calls as calculated above.

We can also get probability for how many calls happen in last 2 days.

We use cdf function expon.cdf(2,scale=0.25)

**Difference between exponential and normal distribution**

Clearly both use continues variables to predict the values but exponential is more often used to predict time related data while others can be predicted using normal distribution.

**Correlation**

It can be used to find the strength of relation between 2 data ,a data can be strongly correlated if the data tends towards +1 or -1 but week when it is towards 0. But the downside of correlation is it can only relate to linear data. It works poorly on non-linear data. And it also don’t identify confounding columns ,for eg, Lets coffee and lung cancer has strong correlation of 0.88 but we cant conclude that coffee is cause of lung cancer. There may be other columns which lead to lung cancer, it may be people who drink coffee tends to smoke quite often, so here the smoking column is an example of confounding column.