Mean,Mode and Median:

Mean,Mode and Median are different measures of center in a numerical data set. They each try to summarize a dataset with a single number to represent a "typical" data point from the dataset.

Mean:

1. Mean: The "average" number found by adding all data points divided by the number of datapoints.

Calculating the Mean: The arithmetic mean is sum of all data points divided by the number of data points

$$\text{mean} = \frac{\text{sum of data}}{\#\text{ of data points}}$$

Here's the same formula written more formally:

$$\text{mean} = \frac{\sum x_i}{n}$$

Mathematically solved:

Question: Find the mean of this data: 1,5,4,5,8,1,2,3,4 Answer: First we have to add all the numbers: 1+5+4+5+8+1+2+3+4=33 Mean=33/9=3.666 The mean of the data points is 3.666

Implementation in python:

In [1]:

```
import numpy as np
import pandas as pd

data=np.array([1,5,4,5,8,1,2,3,4])

print("The mean of the data point is ",data.mean())
```

```
The mean of the data point is  3.6666666666666665
```

Note: Mean is the most common measures of central tendency but it has a huge downside because it is easily affected by outliers - which value is significantly greater than other values in the dataset.

Let us see an example:

Mathematically solved:

Question: Find the mean of this data: 1,5,4,5,8,1,2,3,4,50 Answer: First we have to add all the numbers: 1+5+4+5+8+1+2+3+4+50=83 Mean=83/10=8.3 The mean of the data points is 8.3

Implementation in python:

In [2]:

```
import numpy as np
import pandas as pd

data=np.array([1,5,4,5,8,1,2,3,4,50])

print("The mean of the data point is ",data.mean())
```

```
The mean of the data point is  8.3
```

Explanation: As you can see because of one outlier value 50 your mean get corrupted by a high margin from 3.66 to 8.3.

Mode:

Mode: The most frequent number that is, the number that occurs the highest number of times.

Finding the mode: The mode is the most commonly occuring data points in a dataset. The mode is useful when there are lots of repeated values in a dataset. There can be no mode, one mode or multiple modes.

Mathematically solved:

Question: Find the mode of this data: 1,5,4,8,1,2,3,9 Answer: As we can see the 1 is occuring the highest times. So, the mode is 1

Implementation in python:

In [3]:

```
import numpy as np
import statistics

data=np.array([1,5,4,8,1,2,3,9])

print("The mode of the data point is ",statistics.mode(data))
```

```
The mode of the data point is  1
```

Median:

Median: The middle number; found by ordering all the data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

Finding the median: The median is the middle point in a dataset - half of the datapoints are smaller than the median and half of the data points are larger.

To find the median:

1. Arrange the data points from smallest to largest.
2. If the number of data points is odd, the median is the middle data points in the list.
3. If the number of data point is even , then the median is the average of the two middle data points in the list

Mathematically solved:(Odd number of points)

Question: Find the median of this data: 1,5,4,8,1,2,3,9,7 Answer: Step1: Put the data in order first: 1,1,2,3,4,5,7,8,9 Step2: There is an odd number in data points. so, the median is 4 the middle data point.

Implementation in python:

In [4]:

```
import numpy as np
import statistics

data=np.array([1,5,4,8,1,2,3,9,7])

print("The mode of the data point is ",statistics.median(data))
```

```
The mode of the data point is  4
```

Mathematically solved:(Even number of points)

Question: Find the median of this data: 1,5,4,8,1,2,3,9 Answer: Step1: Put the data in order first: 1,1,2,3,4,5,8,9 Step2: There is an even number in data points. so, the median is mean of 3 and 4 which is 3.5.

Implementation in python:

In [5]:

```
import numpy as np
import statistics

data=np.array([1,5,4,8,1,2,3,9])

print("The mode of the data point is ",statistics.median(data))
```

```
The mode of the data point is  3.5
```

Variability:

Variability (also called spread or dispersion) refers to how spread out a set of data is. Variability gives you a way to describe how much data sets vary and allows you to use statistics to compare your data to other sets of data.

Measures of variability:

Following are some of the measures of variability that offers to differentiate between datasets:

- Variance
- Standard Deviation
- Range
- Interquartile range

1. Variance

The variance is a numerical measure of how the data values is dispersed around the mean..
Mathematically, it is defined as the average of squared differences from the mean value.

To calculate the variance follows these steps:

1. Find out the mean
2. Then for each number subtract the mean and square the result
3. The worked out the average of those squared difference

Formulae:

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

Whree

- σ² specifies variance of the data set
- xi specifies i^th value in data set
- μ specifies the mean of data set
- n specifies total number of observations

Mathematically solved:

Example: Find the variance of the numbers 3, 8, 6, 10, 12, 9, 11, 10, 12, 7.

Solution:

Given,

3, 8, 6, 10, 12, 9, 11, 10, 12, 7

Step 1: Compute the mean of the 10 values given.

Mean = (3+8+6+10+12+9+11+10+12+7) / 10 = 88 / 10 = 8.8

Step 2: Make a table with three columns, one for the X values, the second for the deviations and the third for squared deviations. As the data is not given as sample data so we use the formula for population variance. Thus, the mean is denoted by μ.

Value X	$X - \mu$	$(X - \mu)^2$
3	-5.8	33.64
8	-0.8	0.64
6	-2.8	7.84
10	1.2	1.44
12	3.2	10.24
9	0.2	0.04
11	2.2	4.84
10	1.2	1.44
12	3.2	10.24
7	-1.8	3.24
Total	0	73.6

Step 3:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

$= 73.6 / 10$

$= 7.36$

Implementation in python:

In [6]:

```
import numpy as np
import statistics

X=np.array([3,8,6,10,12,9,11,10,12,7])

print("The variance is: ",statistics.variance(X))
```

```
The variance is:  8
```

2. Standard Deviation

- Standard Deviation is a measure of spread in statistics. It is used to quantify the measure of spread. Variation of a set of data values. It is very much similar to variance, gives the measure of deviation whereas variation provide the squared value.
- Standard Deviation measures the spreadness of data values with respect to mean and mathematically, is calculated as square root of variance.

Formulae:

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

Where

- σ specifies standard deviation of the data set
- xi specifies i^th value in data set
- μ specifies the mean of data set
- n specifies total number of observations

Mathematically solved:

From the previous question of variance part just underroot the variance √variance √7.366 2.714

Implementation in python:

In [7]:

```
import numpy as np
import statistics

X=np.array([3,8,6,10,12,9,11,10,12,7])

print("The variance is: ",statistics.stdev(X))
```

```
The variance is:  2.8284271247461903
```

3. Range:

The range is the difference between the lowest and highest values.
Intuition: The range will tell us how spread our data is.

$$\text{Range}(X) = \text{Max}(X) - \text{Min}(X)$$

Mathematically solved:

Question: Find the range of this data: 4,6,9,3,7 Answer: The lowest value in the dataset is 3 The highest value in the dataset is 9 Range=Max(x) - Min(x) =9-3 so, the range is 6

Implementation in python:

In [8]:

```
import numpy as np

x=np.array([4,6,9,3,7])

print("The range is: ",max(x)-min(x))
```

```
The range is:  6
```

Mid range:

The mid range of a set of statistical data values is the arithmetic mean of the maximum and minimum value in dataset

$$\text{Midrange} = \frac{\text{Max}(x) + \text{Min}(x)}{2}$$

Mathematically solved:

Question: Find the midrange of this data: 4,6,9,3,7 Answer: The lowest value in the dataset is 3 The highest value in the dataset is 9 Range=(Max(x) + Min(x))/2 =(9+3)/2 so, the midrange is 6

Implementation in python:

In [9]:

```
import numpy as np

x=np.array([4,6,9,3,7])

print("The midrange is: ",(max(x)+min(x))/2)
```

```
The midrange is:  6.0
```

4. Interquartile Range(IQR):

- Before going to the IQR we have to know the concept of **Percentile** and **Quartile**:

Percentile:

Percentile is a measure used in statistics indicating the values which a given percentage of observation in a group of observation falls

How to calculate percentile:

Let us suppose you have 100 datapoints in a random variable:

X={x1,x2,x3,x4,x5,x6,.....,x95,x96,x97,x98,x99,x100}

Step1: Rank the values in the dataset in order from smallest to largest.
X={x1n,x2n,x3n,x4n,x5n,x6n,.....,x95n,x96n,x97n,x98n,x99n,x100n}

Step2: Multiply K(percent) by n(total number of values in the dataset). This is the index . You will refer to this in the next steps as the position of values in your dataset.

```
0.01*100=1st percentile
0.02*100=2nd percentile
0.25*100=25th percentile
0.50*100=50th percentile
0.75*100=75th percentile
0.1*100=100th percentile
```

Step3: If the index is not a round number, round it up(or down, if it is closer to the lower number)to the nearest whole number.

Implementation in python:

In [10]:

```
import numpy as np

data=range(0,101)

print("1st percentile of data is: ",np.percentile(data,1))
print("5th percentile of data is: ",np.percentile(data,5))
print("25th percentile of data is: ",np.percentile(data,25))
print("50th percentile of data is: ",np.percentile(data,50))
print("75th percentile of data is: ",np.percentile(data,75))
print("100th percentile of data is: ",np.percentile(data,100))
```

```
1st percentile of data is:  1.0
5th percentile of data is:  5.0
25th percentile of data is:  25.0
50th percentile of data is:  50.0
75th percentile of data is:  75.0
100th percentile of data is:  100.0
```

In [11]:

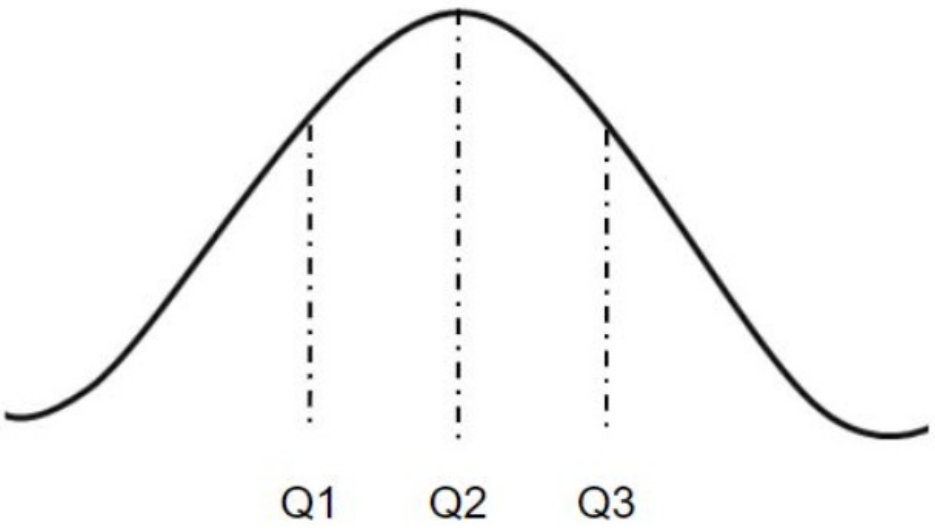
```
import numpy as np

data=range(0,501)

print("1st percentile of data is: ",np.percentile(data,1))
print("5th percentile of data is: ",np.percentile(data,5))
print("25th percentile of data is: ",np.percentile(data,25))
print("50th percentile of data is: ",np.percentile(data,50))
print("75th percentile of data is: ",np.percentile(data,75))
print("100th percentile of data is: ",np.percentile(data,100))
```

```
1st percentile of data is:  5.0
5th percentile of data is:  25.0
25th percentile of data is:  125.0
50th percentile of data is:  250.0
75th percentile of data is:  375.0
100th percentile of data is:  500.0
```

Quartile:

A quartile is a type of quantile which divides the number of data points into four more or less equal parts or quarters.

1. **1. Q1(First quratile/Lower quartile/25th percentile): Splits off the lowest 25% of data from the highest 75%.**
2. **2. Q2(Second quartile/Median/50th percentile): Cuts the dataset into half.**
3. **3. Q3(Third quartile/Upper quartile/75th percentile): Splits off the highest 25% of data from the lowest 75%.**



Implementation in python:

In [12]:

```
import numpy as np

data=range(0,101)

print("Q1 of data is: ",np.quantile(data,0.25))
print("Q2 of data is: ",np.quantile(data,0.50))
print("Q3 of data is: ",np.quantile(data,0.75))
print("Q4 of data is: ",np.quantile(data,0.1))
```

```
Q1 of data is:  25.0
Q2 of data is:  50.0
Q3 of data is:  75.0
Q4 of data is:  10.0
```

In [13]:

```
import numpy as np

data=range(0,501)

print("Q1 of data is: ",np.quantile(data,0.25))
print("Q2 of data is: ",np.quantile(data,0.50))
print("Q3 of data is: ",np.quantile(data,0.75))
print("Q4 of data is: ",np.quantile(data,0.1))
```

```
Q1 of data is:  125.0
Q2 of data is:  250.0
Q3 of data is:  375.0
Q4 of data is:  50.0
```

IQR:

- The interquartile range(IQR) is the difference between the 75th and 25th percentile of the data. It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers
- The interquartile range(IQR),also called as midspread or middle 50% or technically H-spread is the difference between the third quartile(Q3) and the first quartile(Q1).
- IQR is a amount of spread in the middle 50% of a dataset

$$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$$

Implementation in python:

In [14]:

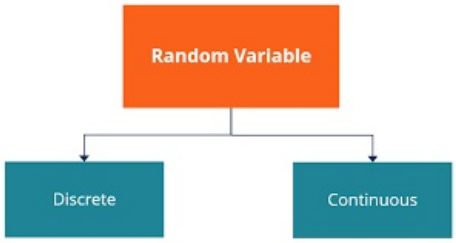
```
from scipy import stats

data = [32, 36, 46, 47, 56, 69, 75, 79, 79, 88, 89, 91, 92, 93, 96, 97, 101, 105, 112, 116]

IQR = stats.iqr(data)

print(IQR)
```

```
30.5
```

Random variable:

A random variable is a set of possible values from a random experiment.
We use a capital letters like X or Y for random variables.

Random variable are of two types discrete or continous Random variable:



Discrete Random Variable: A discrete random variable is one which may take on only a countable number of distinct values and thus can be quantified.
Example:

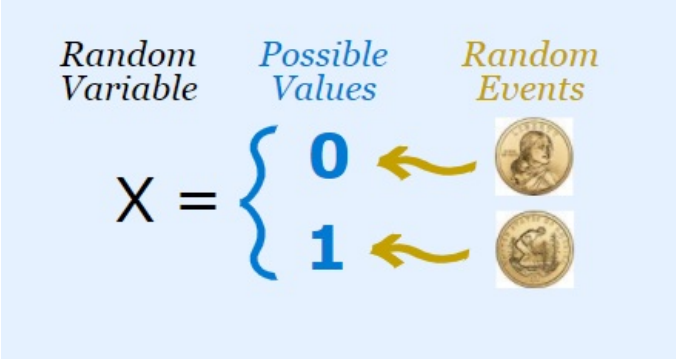
1. **Coin toss: you can define a random variable X to be the side which comes up when you toss a coin. X can take values : [H,T] and therefore is a discrete random variable.**
2. **Rolled fair dice:you can define a random variable X to be the number which comes up when you roll a fair dice. X can take values : [1,2,3,4,5,6] and therefore is a discrete random variable.**

Some examples of discrete probability distributions are Bernoulli distribution, Binomial distribution, Poisson distribution etc.

Continuous Random Variable: A random variable which can take infinite number of values in an interval is known as continuous random variable.
Example:

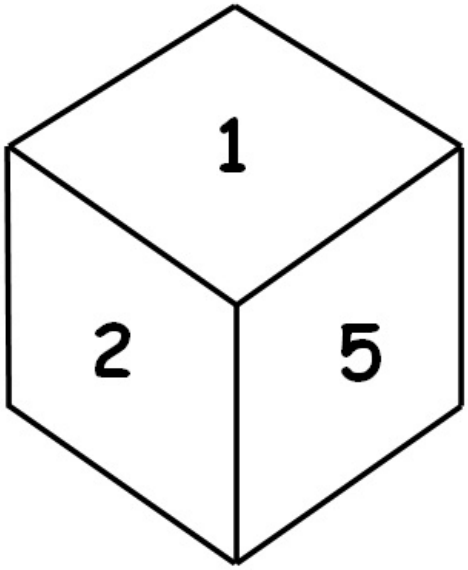
1. **Weight of a group of individuals: a random variable X to be the height of students in a class**
2. **Price of house: A random variable X to be the price of house in a city**

Example: Tossing a coin(discrete random variable).



Explanation: If we toss a coin we have only two possible outcome either we get "Head" or "Tail".so, our random variable "X" contains two possible outcome: X={H,T}

Example: Rolling a fair dice(discrete random variable).



we know in fair dice we have 1,2,3,4,5,6 on a single side of a dice.

If we roll a fair dice we have six possible outcomes either 1,2,3,4,5,6. So, our random variable "X" contain six possible outcome:

X={1,2,3,4,5,6}

Example: Height of a randomly picked student(continous random variable)

as we know that height lie between the intervals. so we taking the intervals is [130,170]

H={130.2,140.9,172.6,180.3,150,122.1}

Population and Sample:

Population Distribution: The population is the whole set of values, or individuals, you are interested in.
For example: If you want to know the average height of the residents of india, that is your population. i.e., the population of India

Population
Mean =μ
Variance $\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$
Standard Deviation $\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$

Sample Distribution: The sample is a subset of the population, and is the set of values you actually use in your estimation. Let us think 1000 individuals you have selected for your study to know about average height of the residents of India. This sample has computed from values

Sample
Mean= $\bar{x}$
Variance $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$
Standard Deviation $s = \sqrt{\frac{\sum(x - \bar{x})^2}{N-1}}$

Application of population and sample: Let us understand this with an example:

Question:We have to find out the average height of the mens in India?

Answer: Suppose you have to find out the average height of the mens in India. We have 532 millions mens in India.

The approach I get from aspiring data scientists is to simply calculate the average:

- **First, measure the weights of all the students in the science department**
- **Add all the weights**
- **Finally, divide the total sum of weights with a total number of students to get the average**

We have 532 million people. The size of the data is humongous. Does this approach make sense? Not really – measuring the height of all the students will be a very tiresome and long process. So, what can we do instead? Let's look at an alternate approach.

So, what we can do is randomly and independently pick the 10,000 people as sample and find the average height of 10,000 mens. Which picked up unbiasly.

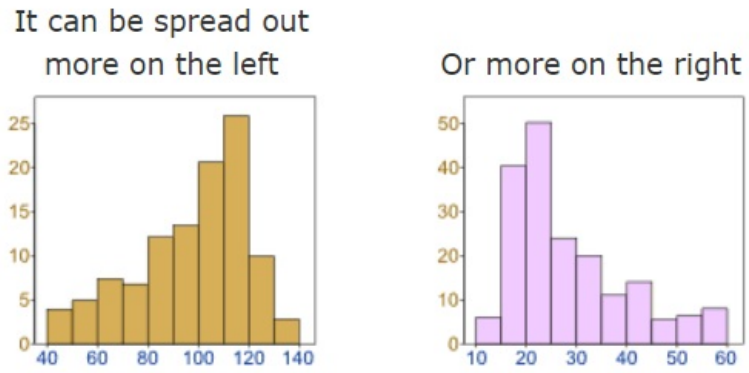
so the intuition is sample mean is roughly equal to the population means

$$\mu \approx \bar{x}$$

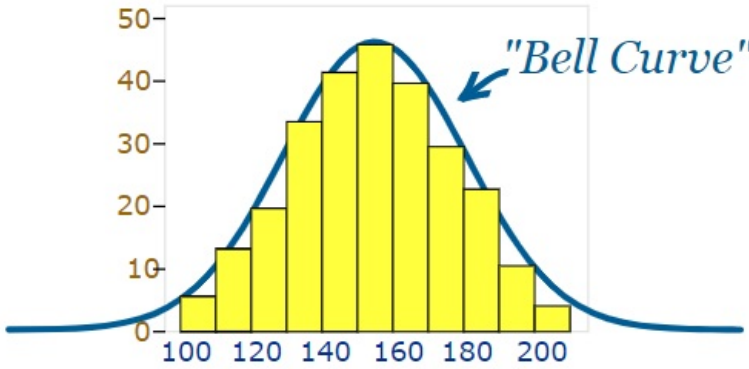
Note: When the sample size increases the mean of sample try to converge toowards μ(mean of population)

Gaussain/Normal Distribution:

Data can be "distributed"(spread out) in different ways:



But there are many cases where the data tends to be around a central value with no bias left or right and it gets close to a "Normal Distribution" like this:



A Normal Distribution

The Bell curve is a normal distribution and the histogram shows some data that follows it closely, but not perfecty(which is usual).

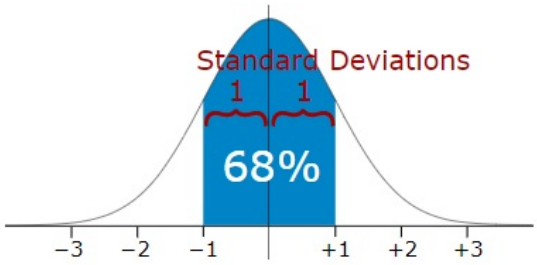
$X \sim N(\mu,\sigma^2)$
This notation represent as Random variable X follows Normal distribution with mean(μ) and variance(σ^2)

The Gaussain/Normal Distribution has:

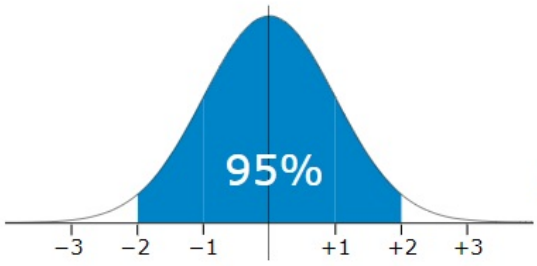
1. **mean=mode=median**
2. **symmetric about center**
3. **50% value less than mean and 50% value greater than mean**

The Gaussain/Normal Distribution property:

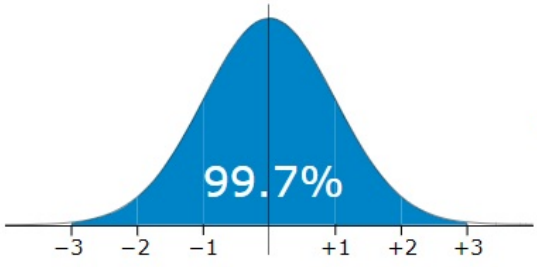
$pro[\mu-\sigma<=X<=\mu+\sigma]\approx68\%$
$pro[\mu-2\sigma<=X<=\mu+2\sigma]\approx95\%$
$pro[\mu-3\sigma<=X<=\mu+3\sigma]\approx99.7\%$



68% of values are within
1 standard deviation of the mean



95% of values are within
2 standard deviations of the mean



99.7% of values are within
3 standard deviations of the mean

Real life example that follows normal distribution:

1. Heights of people
2. Blood Pressure
3. Marks on exams
4. Size of things produced by machines

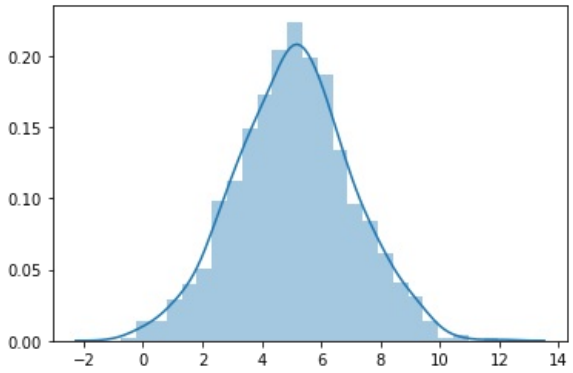
Implementation in python:

In [15]:

```
import numpy as np
import seaborn as sns

s=np.random.normal(5,2,1000)  #mean 5,standarad deviation 2,size 1000
sns.distplot(s)
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20d6bea6760>
```



Standarad Normal Distribution:

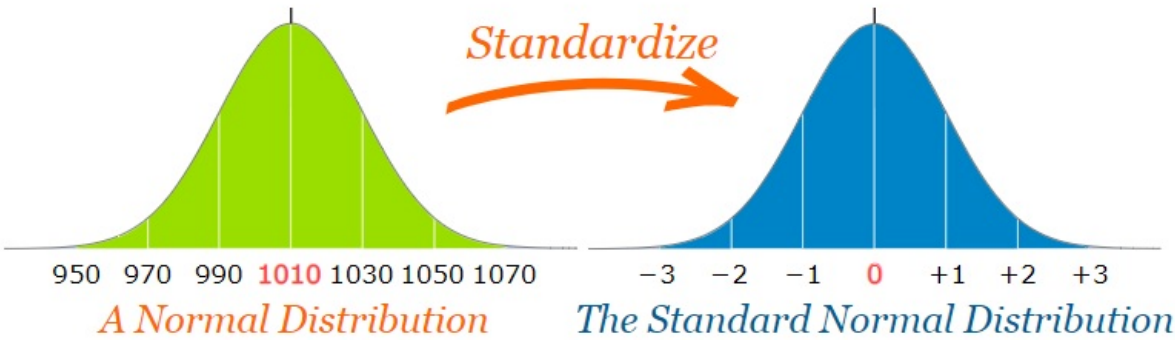
A standard normal distribution is a normal distribution with zero mean (μ=0) and unit variance (σ^2=1)

So, to convert a normal distribution to Standard normal distrubution we have to use Z-score standarization technique:

Z-score standarization: This technique consists of subtracting the mean of the column from each value in a column, and then dividing the results by the standard deviation of the column. The formulae to achieve this is the following:

$$z = \frac{x - \mu}{\sigma}$$

The result of standarization is that the features will be rescaled.So that they will have the properties of a standard normal distribution, as follows:
μ=0
σ^2=1



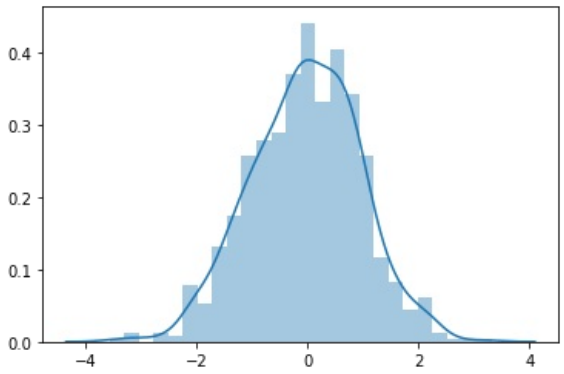
Implementation in python:

In [16]:

```
import numpy as np
import seaborn as sns

data=np.random.normal(0,1,1000) #mean 0,standarad deviation 1,size 1000

sns.distplot(data)
```

Out[16]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20d6c65a550>
```

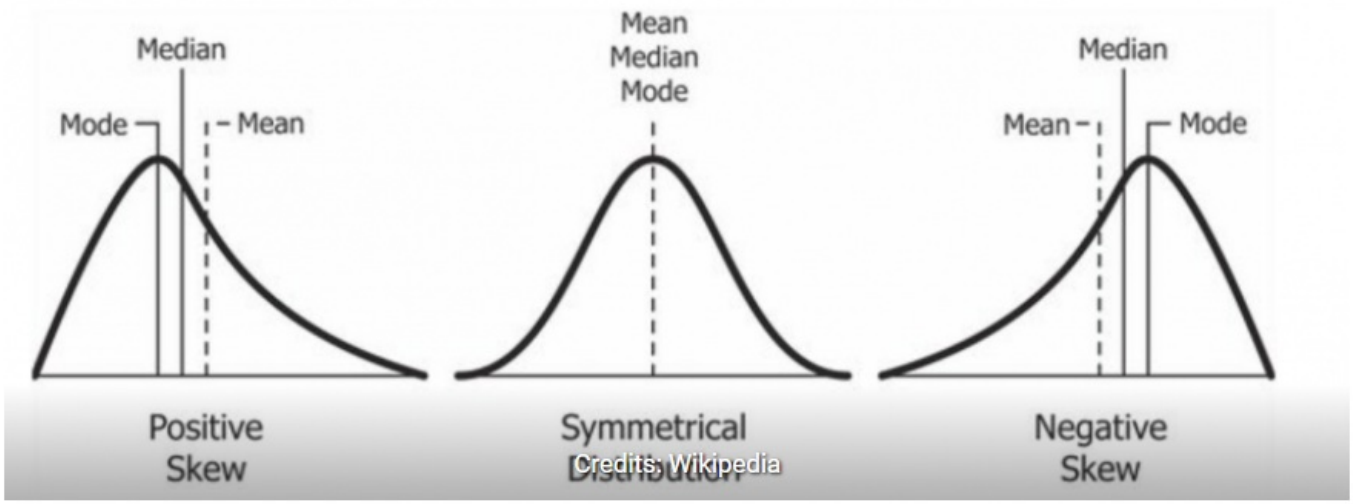


Skewness:

skewness is the measure of how much the probablity distribution of random variable deviates from the normal distribution.
well, the normal distribution is the probablity distribution without skewness.

Well, the normal distribution is the probability distribution without any skewness. You can look at the image below which shows symmetrical distribution that's basically a normal distribution and you can see that it is symmetrical on both sides of the dashed line. Apart from this, there are two types of skewness:

- Positive Skewness
- Negative Skewness



Formula For Skewness Calculate:

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

OR

$$\textbf{Skewness} = \frac{3\,(\textit{Mean} - \textit{Median})}{\textit{Standard Deviation}}$$

γ1 represents coefficient of skewness

xi represents ith value in data vector

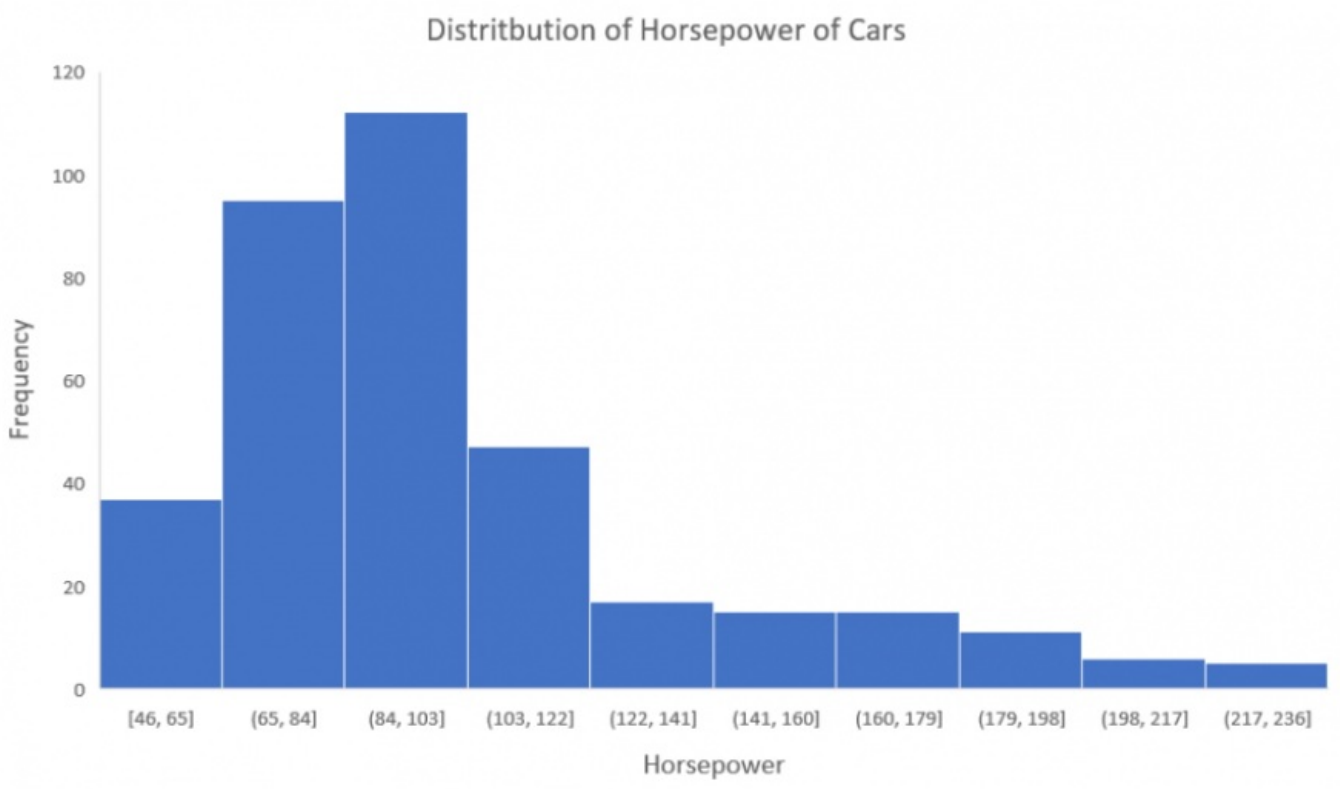
xbar represents mean of data vector

n represents total number of observations

The probability distribution with its tail on the right side is a positively skewed distribution and the one with its tail on the left side is a negatively skewed distribution.

Why is Skewness Important?

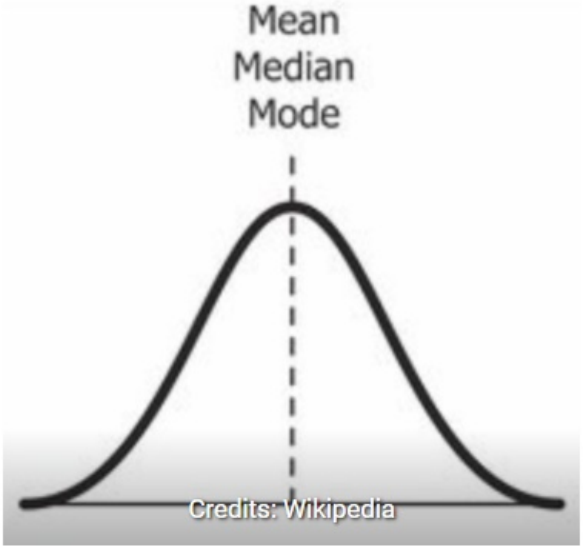
First, linear models work on the assumption that the distribution of the independent variable and the target variable are similar. Therefore, knowing about the skewness of data helps us in creating better linear models.
Secondly, let's take a look at the below distribution. It is the distribution of horsepower of cars:



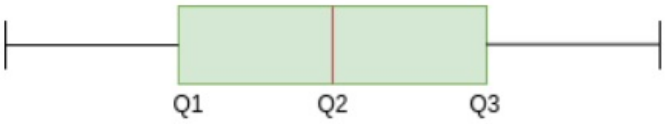
You can clearly see that the above distribution is positively skewed. Now, let's say you want to use this as a feature for the model which will predict the mpg (miles per gallon) of a car.
Since our data is positively skewed here, it means that it has a higher number of data points having low values, i.e., cars with less horsepower. So when we train our model on this data, it will perform better at predicting the mpg of cars with lower horsepower as compared to those with higher horsepower.
Also, skewness tells us about the direction of outliers. You can see that our distribution is positively skewed and most of the outliers are present on the right side of the distribution.

Note: The skewness does not tell us about the number of outliers. It only tells us the direction.

What is Symmetric/Normal Distribution?



Yes, we're back again with the normal distribution. It is used as a reference for determining the skewness of a distribution. As I mentioned earlier, the ideal normal distribution is the probability distribution with almost no skewness. It is nearly perfectly symmetrical. Due to this, the value of skewness for a normal distribution is zero.
But, why is it nearly perfectly symmetrical and not absolutely symmetrical?
That's because, in reality, no real word data has a perfectly normal distribution. Therefore, even the value of skewness is not exactly zero; it is nearly zero. Although the value of zero is used as a reference for determining the skewness of a distribution.
You can see in the above image that the same line represents the mean, median, and mode. It is because the mean, median, and mode of a perfectly normal distribution are equal.
So far, we've understood the skewness of normal distribution using a probability or frequency distribution. Now, let's understand it in terms of a boxplot because that's the most common way of looking at a distribution in the data science space.

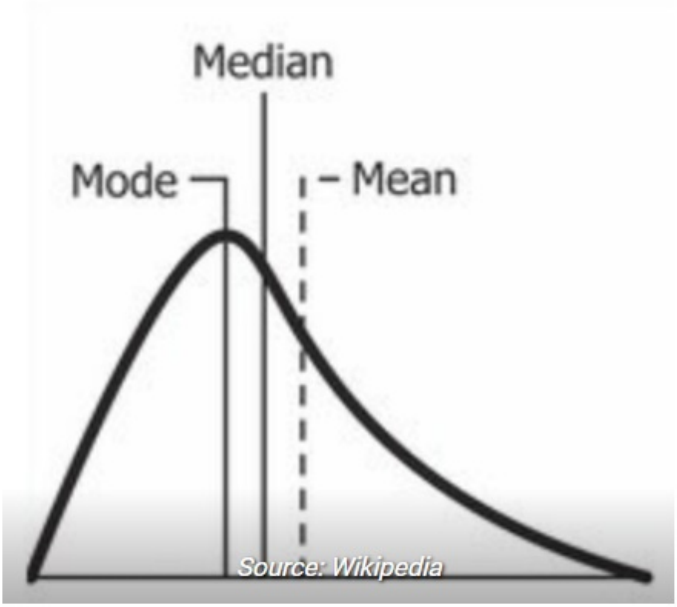


The above image is a boxplot of symmetric distribution. You'll notice here that the distance between Q1 and Q2 and Q2 and Q3 is equal i.e.:

$$Q_3 - Q_2 = Q_2 - Q_1$$

But that's not enough for concluding if a distribution is skewed or not. We also take a look at the length of the whisker; if they are equal, then we can say that the distribution is symmetric, i.e. it is not skewed.
Now that we've discussed the skewness in the normal distribution, it's time to learn about the two types of skewness which we discussed earlier. Let's start with positive skewness.

Understanding Positively Skewed Distribution



A positively skewed distribution is the distribution with the tail on its right side. The value of skewness for a positively skewed distribution is greater than zero. As you might have already understood by looking at the figure, the value of mean is the greatest one followed by median and then by mode.

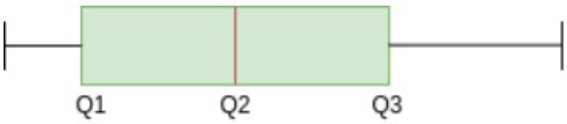
So why is this happening?

Well, the answer to that is that the skewness of the distribution is on the right; it causes the mean to be greater than the median and eventually move to the right. Also, the mode occurs at the highest frequency of the distribution which is on the left side of the median. Therefore, mode < median < mean</b?.

In the above boxplot, you can see that Q2 is present nearer to Q1. This represents a positively skewed distribution. In terms of quartiles, it can be given by:

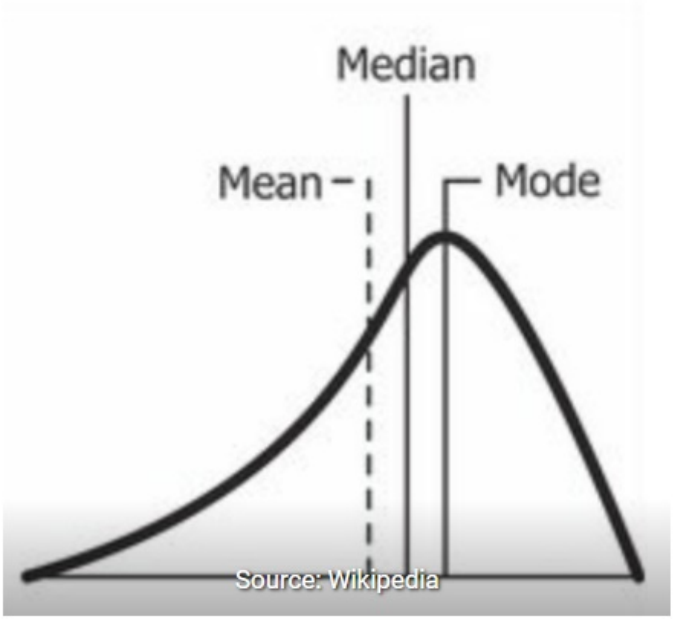
$$Q_3 - Q_2 > Q_2 - Q_1$$

In this case, it was very easy to tell if the data is skewed or not. But what if we have something like this:



Here, Q2-Q1 and Q3-Q2 are equal and yet the distribution is positively skewed. The keen-eyed among you will have noticed the length of the right whisker is greater than the left whisker. From this, we can conclude that the data is positively skewed.
So, the first step is always to check the equality of Q2-Q1 and Q3-Q2. If that is found equal, then we look for the length of whiskers.

Understanding Negatively Skewed Distribution



As you might have already guessed, a negatively skewed distribution is the distribution with the tail on its left side. The value of skewness for a negatively skewed distribution is less than zero. You can also see in the above figure that the mean < median < mode.

In the boxplot, the relationship between quartiles for a negative skewness is given by:

$$Q_3 - Q_2 < Q_2 - Q_1$$

Similar to what we did earlier, if Q3-Q2 and Q2-Q1 are equal, then we look for the length of whiskers. And if the length of the left whisker is greater than that of the right whisker, then we can say that the data is negatively skewed.

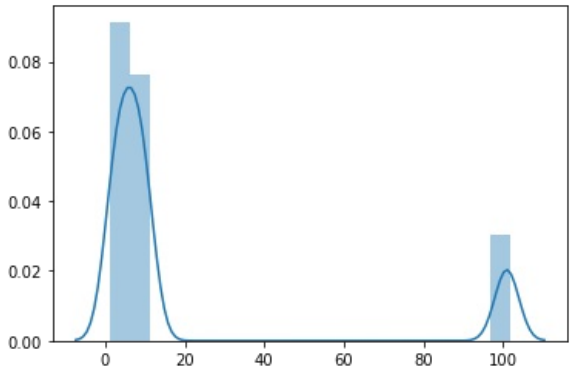
when
skewness=0 : normally distributed
skewness<0 : more weight on the left hand side / Negative skewness
skewness>0 : more weight on the right hand side / Positive skewness

implementation in python:

In [17]:

```
import numpy as np
import seaborn as sns
from scipy.stats import skew

a=[1,2,3,4,5,6,7,8,9,10,11,100,102]
sns.distplot(a)

print("Skewness of the data is:",skew(a))
```

```
Skewness of the data is: 1.8896904721397088
```



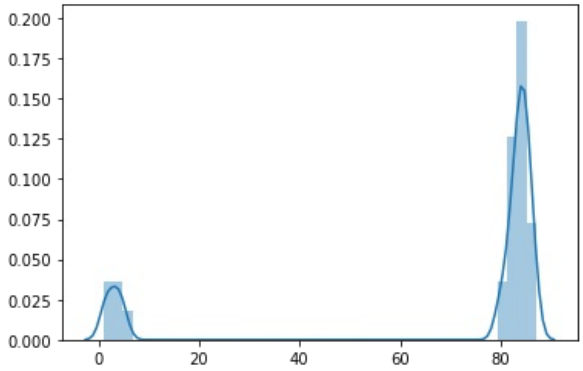
Explanation: As we can see this plot is positive skew because the value of skewness is greater than 1

implementation in python:

In [18]:

```
import numpy as np
import seaborn as sns
from scipy.stats import skew

a=[1,2,3,4,5,80,80,82,82,82,83,83,83,83,84,84,84,84,84,85,85,85,85,85,85,86,86,86,87]
sns.distplot(a)

print("Skewness of the data is:",skew(a))
```

```
Skewness of the data is: -1.7250964422584605
```



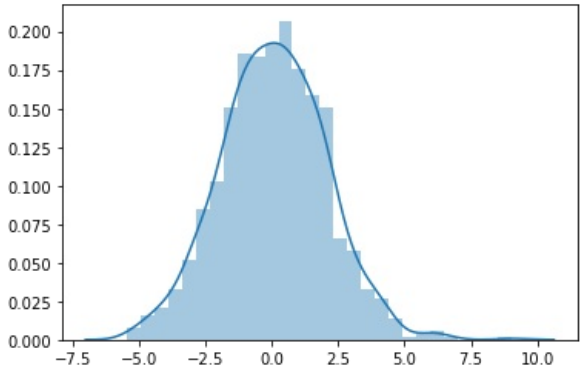
Explanation: As we can see this plot is negative skew because the value of skewness is less than 1

implementation in python:

In [19]:

```
import numpy as np
import seaborn as sns
from scipy.stats import skew

a=np.random.normal(0,2,1000)

sns.distplot(a)

print("Skewness of the data is:",skew(a))
```

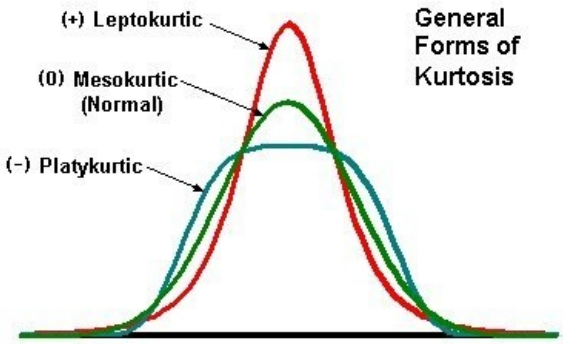
```
Skewness of the data is: 0.13467537582204764
```



Explanation: As we can see this plot follow nearly normal distribution that is why it have skew value nearly about 0(zero)

Kurtosis:

- Kurtosis is the measure of thickness or heaviness of the given distribution.
- Its actually represents the height of the distribution.
- Like skewness, Kurtosis is a descriptor of shape and it describes the shape of the of the distribution interms of height or flatness. Some of the types of Kurtosis are Leptokurtic, Platykurtic and Mesokurtic.



It Is three types

1. mesokurtic (distribution is normal)
2. platykurtic (negative excess of kurtosis)
3. leptokurtic (positive excess of kurtosis)

Formulae:

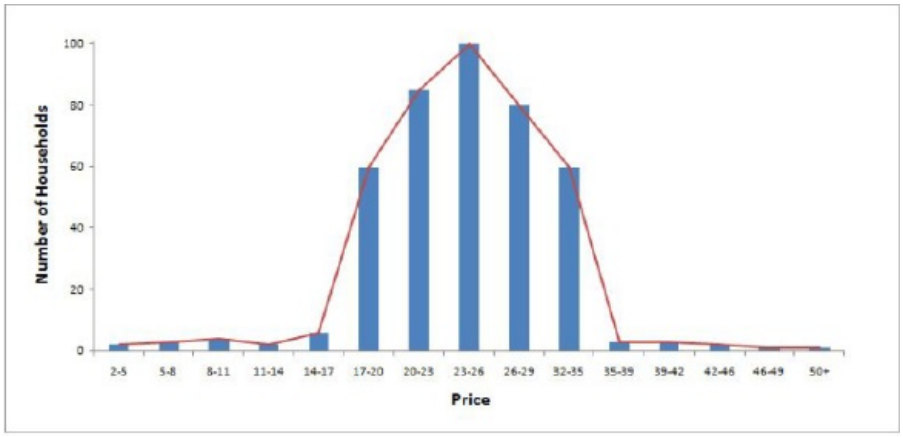
$$\text{Kurtosis (X)} = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

Important Parameter:

- **array : Input array or object having the elements.**
- **axis : Axis along which the kurtosis value is to be measured. By default axis = 0.**
- **fisher : Bool; Fisher's definition is used (normal 0.0) if True; else Pearson's definition is used (normal 3.0) ifset to False.**
- **bias : Bool; calculations are corrected for statistical bias, if set to False.**
- **Returns : Kurtosis value of the normal distribution for the data set**

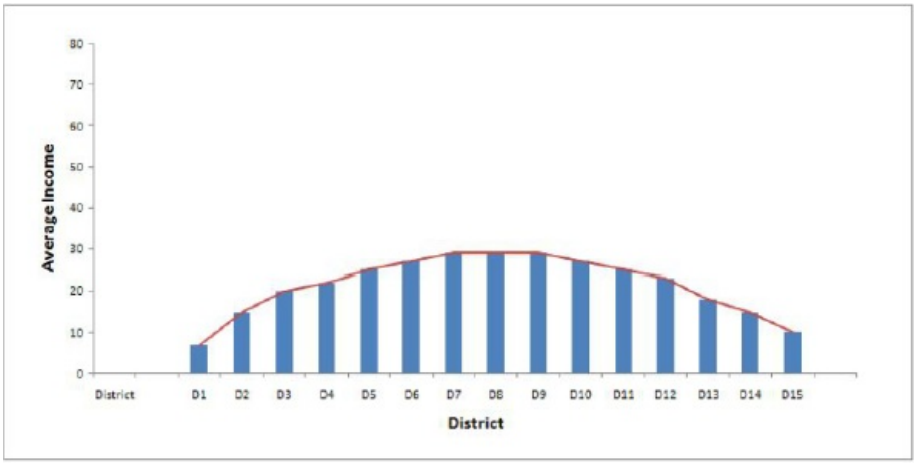
1. Leptokurtic

- **When there is a positive excess of kurtosis, the shape of the distribution is called Leptokurtic. To understand this in terms of shape, it has fatter tails and if compared to a normal distribution it has a similar peak (to be precise, such a distribution has higher peak than what is found in a normal bell-shaped distribution and significantly higher if compared to a Platykurtic distribution) and has values clustered around the centre (mean).**
- **Example- If you are asked to collect a sample to find out the average price of the car that people own in Delhi and you decide to go only in the upper-middle-class localities, then the shape of such a sample'sdistribution will be Leptokurtic.**



2. Platykurtic

- **When there is a negative excess of kurtosis, the shape of the distribution is called Platykurtic.**
- **The data points are highly dispersed along the X-axis that results in thinner tails when compared to a normal distribution and has very few values clustered around the centre (mean). Such a distribution will have little central tendency.**
- **Example- You are asked to go to 15 districts or localities of your city to collect a sample to find out the average income of the state. You decide to go to each district and find the average income of each district by randomly choosing 40 people and finding their annual income. But when the samples were finally plotted on a histogram, the shape of the distribution seemed to be Platykurtic, because all the 15 districts you chose to go were where government households were located where income of people fell in the middle-income band with all having more or less the same average income causing the shape of your distribution to become flatter than the normal.**



3. Mesokurtic

- **This is when the distribution is normal. Here the tails of the distribution are neither too thin nor they are too thickand also the scores are equally divided with scores neither being clustered around the centre nor being too scattered.**

Implementation in python

In [20]:

```
import numpy as np
from scipy.stats import kurtosis

Data=[1,2,3,4,5,80,80,82,82,82,83,83,83,83,84,84,84,84,84,85,85,85,85,85,85,86,86,86,87]

print("Kurtosis for the Data is: ",kurtosis(Data))
```

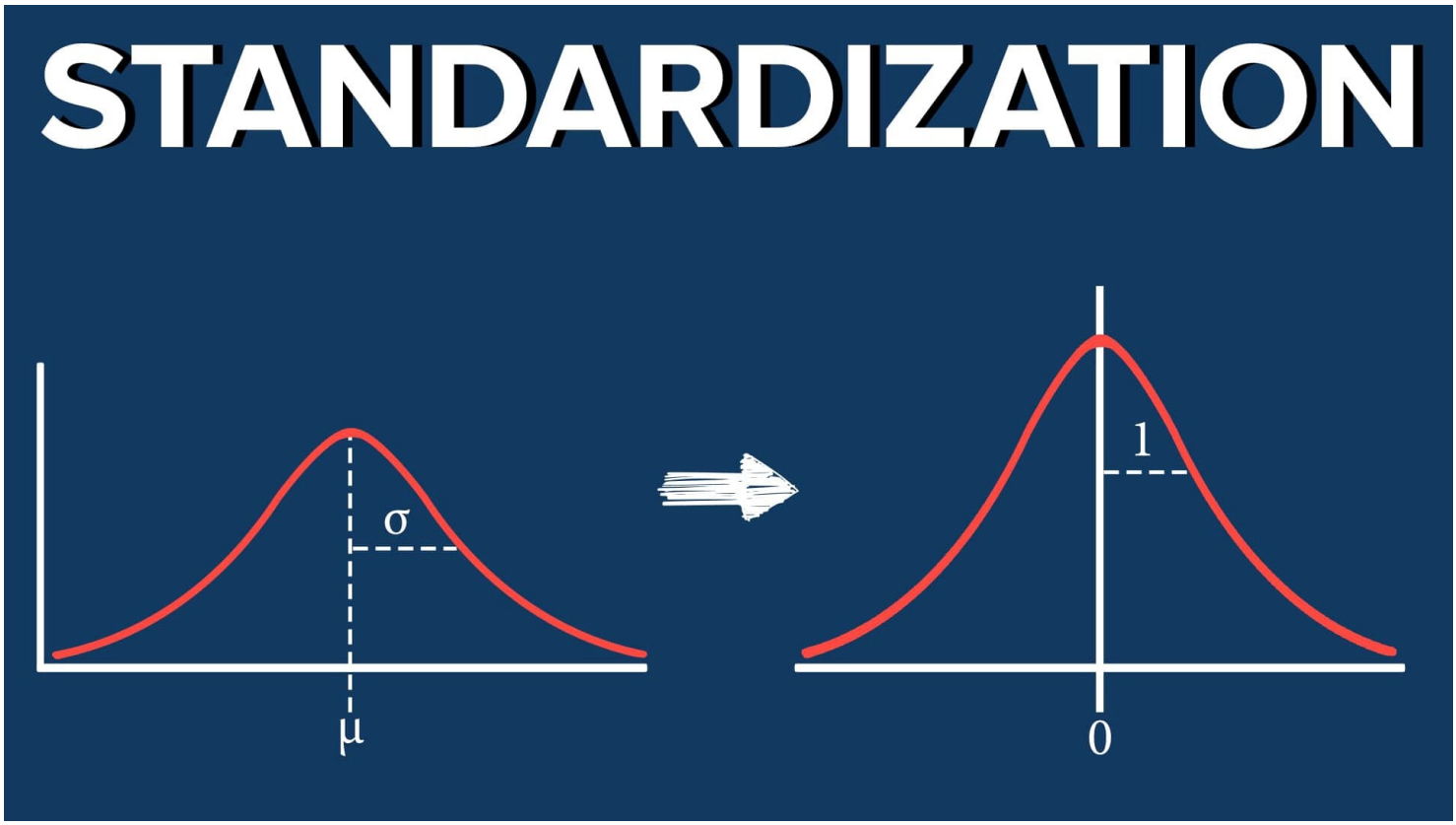
```
Kurtosis for the Data is:  0.9972558983082314
```

Standard normal variate:

It is the distribution that occurs when a normal random variable has a mean of 0(zero) and standard deviation of 1(one). The normal random variable of a standard normal distribution is called a stancard score or Z-score.

z~N(0,1)
This notation represent the random variable follow normal distribution with mean(μ) 0(zero) and standard deviation(σ) 1

Standarization: standarization is the process of transforming a variable with mean(μ) 0(zero) and standard deviation(σ) 1



Formulae:

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

let a variable with normal distribution with mean μ and standard deviation σ

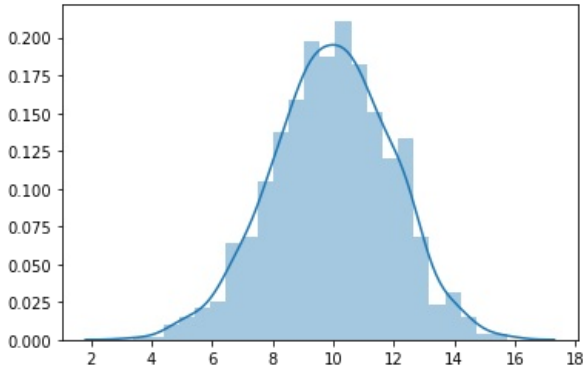
Implementation in python:

In [21]:

```
import numpy as np
import matplotlib as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler

data=np.random.normal(10,2,1000)   #create a normal distributon with mean is 10 and standard deviation is 1
sns.distplot(data)
```

Out[21]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20d6c9f81c0>
```

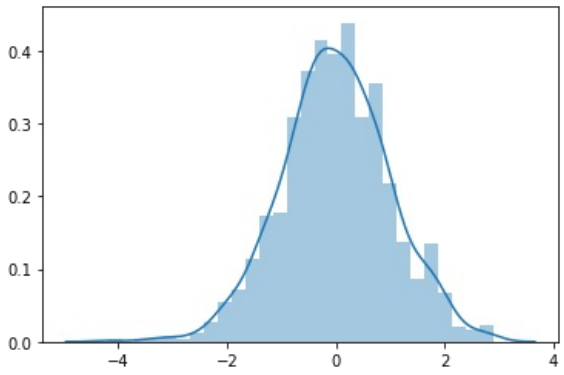


In [22]:

```
data=np.random.normal(10,2,1000).reshape(-1,1)
scaler=StandardScaler()
datascaled=scaler.fit_transform(data)            #applying standarization on data
sns.distplot(datascaled)
```

Out[22]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20d6ca8f7c0>
```

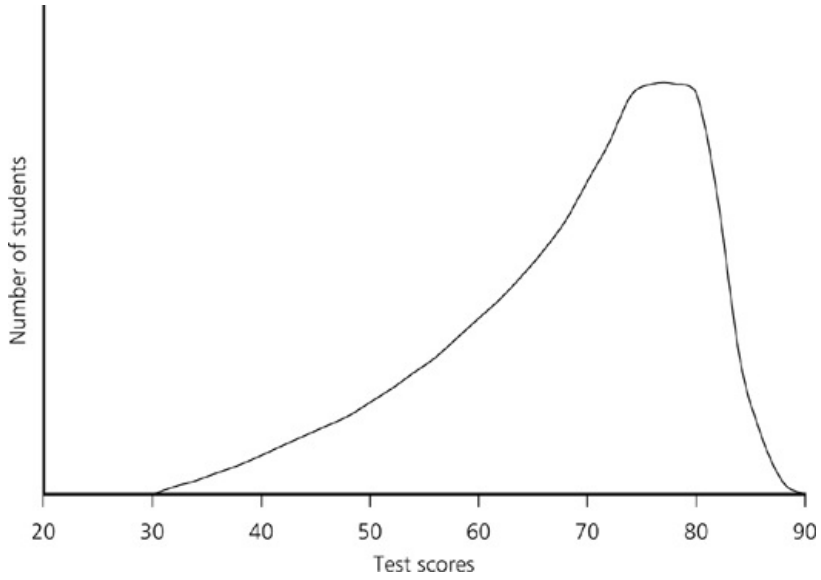


Central Limit Theorm(CLT):

Let us understand CLT therom which is one the best and eligent theorm in the whole statistics.
So, come to the theory first then we will understand via Real world problem and see the implementaion in python.

Definition: Given a dataset with unkmown distribution, the sample mean will be the Normal distribution Let us suppose X is a random variable follows any distribution. We do not know the distribution of X with mean(μ) and variance(σ^2)

X~Any_distribution(μ,σ^2)

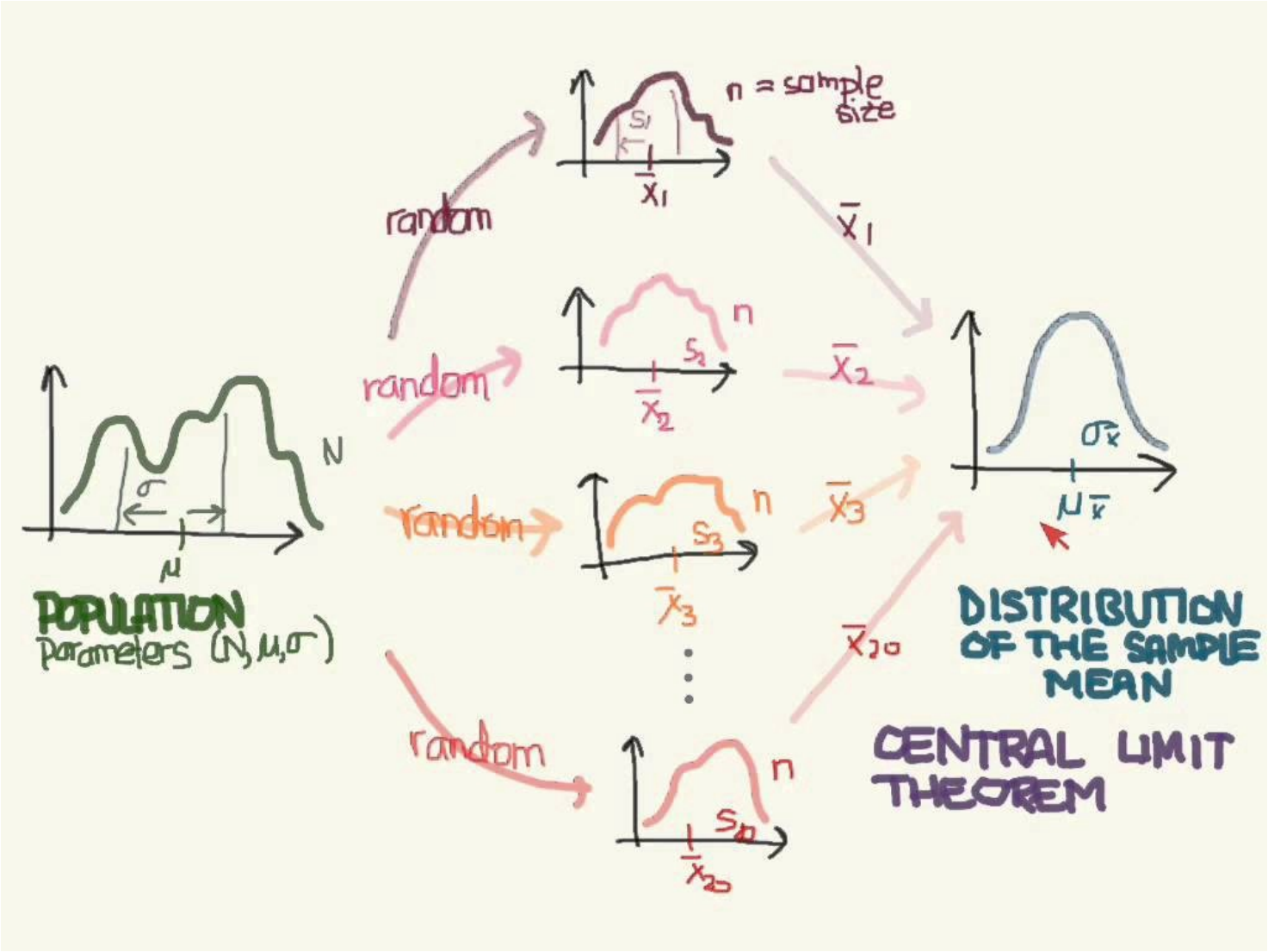


Take random sample of size (n=30)→s1→$\bar{x}1$
Take random sample of size (n=30)→s2→$\bar{x}2$
Take random sample of size (n=30)→s3→$\bar{x}3$
Take random sample of size (n=30)→s4→$\bar{x}4$
.
.
.
.
.
Take random sample of size (n=30)→sm→$\bar{x}m$

$\bar{x}1$, $\bar{x}2$, $\bar{x}3$, $\bar{x}4$, $\bar{x}5$, $\bar{x}6$,....., $\bar{x}m$

$x_i \sim N(\mu, \sigma^2/N)$

μ:This is same as population mean
σ^2: This is same as population variance
N:This is the same as sample size N



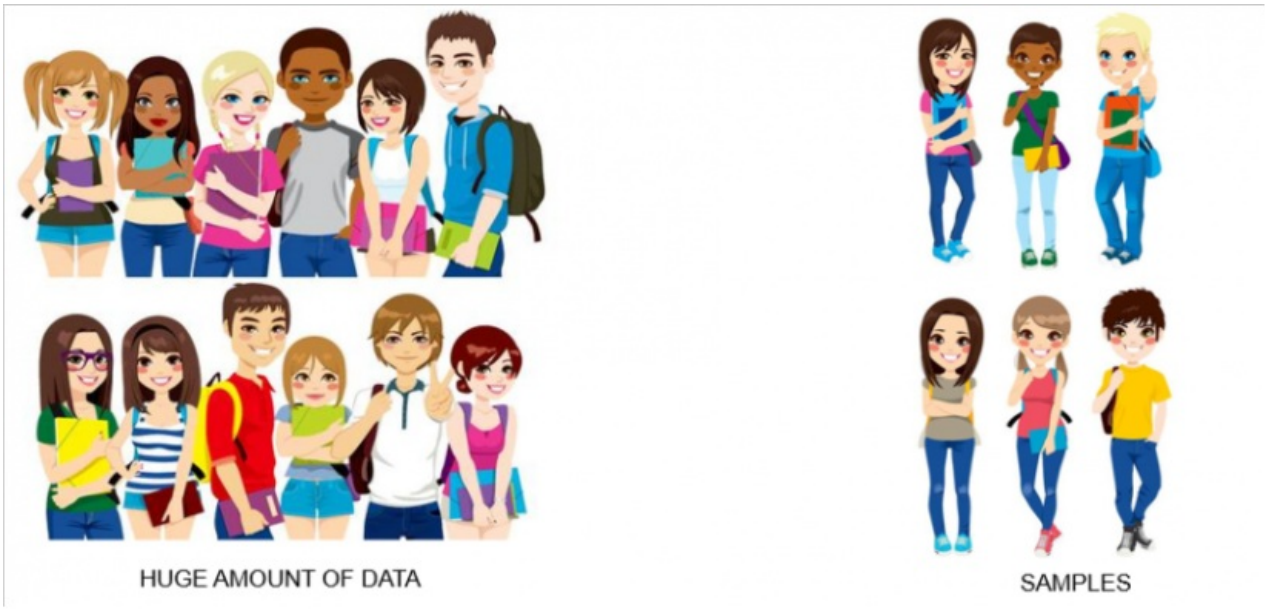
Real life scenario

Consider that there are 15 sections in the science department of a university and each section hosts around 100 students. Our task is to calculate the average weight of students in the science department. Sounds simple, right?
The approach I get from aspiring data scientists is to simply calculate the average:

- **First, measure the weights of all the students in the science department**
- **Add all the weights**
- **Finally, divide the total sum of weights with a total number of students to get the average**
-

But what if the size of the data is humongous? Does this approach make sense? Not really – measuring the weight of all the students will be a very tiresome and long process. So, what can we do instead? Let's look at an alternate approach.

- **First, draw groups of students at random from the class. We will call this a sample. We'll draw multiple samples, each consisting of 30 students.**



- **Calculate the individual mean of these samples**
- **Calculate the mean of these sample means**
- **This value will give us the approximate mean weight of the students in the science department**
- **Additionally, the histogram of the sample mean weights of students will resemble a bell curve (or normal distribution)**

Assumptions Behind the Central Limit Theorem

Before we dive into the implementation of the central limit theorem, it's important to understand the assumptions behind this technique:

1. **The data must follow the randomization condition. It must be sampled randomly**
2. **Samples should be independent of each other. One sample should not influence the other samples**
3. **Sample size should be not more than 10% of the population when sampling is done without replacement**
4. **The sample size should be sufficiently large. Now, how we will figure out how large this size should be? Well, it depends on the population. When the population is skewed or asymmetric, the sample size should be large. If the population is symmetric, then we can draw small samples as well**

Quantile-Quantile(Q-Q) plot:

Q-Q plot are the graphical representation to measure that two sets of data come from the same distribution. In Q-Q plot we graphically analyze and compare two probablity distribution by plotting their quantiles each other. If the two distribution which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on the same line.

How does it works:

Let us we have a random variable X.

X=x1 , x2 , x3 , x4 , x5 , x6 , x7 , x8 , ... , x500

Step 1 X=x1 , x2 , x3 , x4 , x5 , x6 , x7 , ... , x500
Sort all the xi's in ascending order

x1' , x2' , x3' , x4' , x5' , x6' , ... , x500'

Computer the percentile:

so, we have 500 variables
1st percentile: x5' : x(1)
2nd percentile: x10' : x(2)
3rd percentile: x15' : x(3)
4th percentile: x20' : x(4)
5th percentile: x25' : x(5)
.
.
.
.
.
.
.
100th percentile: x100' : x(100)

Step 2 Create a variable Y~N(0,1) knows as the theoretical quantile
Sort all the yi's in ascending order

y1' , y2' , y3' , y4' , y5' , y6' , ... , y500'

Computer the percentile:

so, we have 500 variables
1st percentile: y5' : x(1)
2nd percentile: y10' : y(2)
3rd percentile: y15' : y(3)
4th percentile: y20' : y(4)
5th percentile: y25' : y(5)
.
.
.
.
.
.
.
100th percentile: y100' : y(100)

Step 3 Plot the Q-Q plot using all the percentile:
x(1),x(2),x(3),x(4),x(5),x(6),.....,x(100)
y(1),y(2),y(3),y(4),y(5),y(6),.....,y(100)

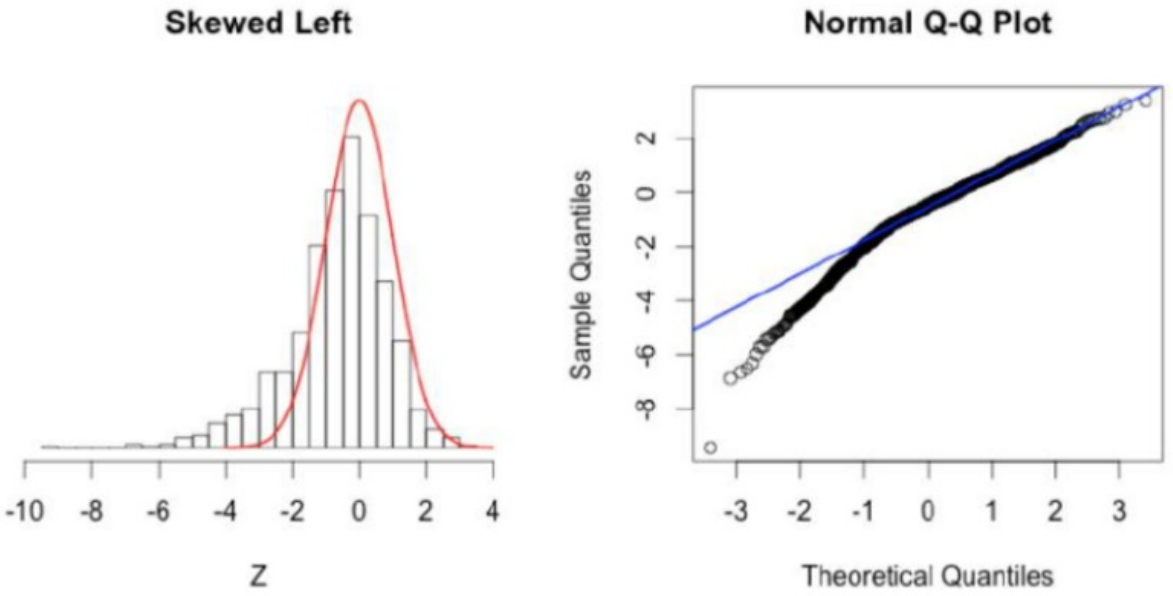
we will have 100 pair of (xn,yn)

Now we have to focus on the ends of the straight line. If the points at the ends of the curve formed should fall on the straight line.
If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normal distribution.

Skewed Q-Q plot

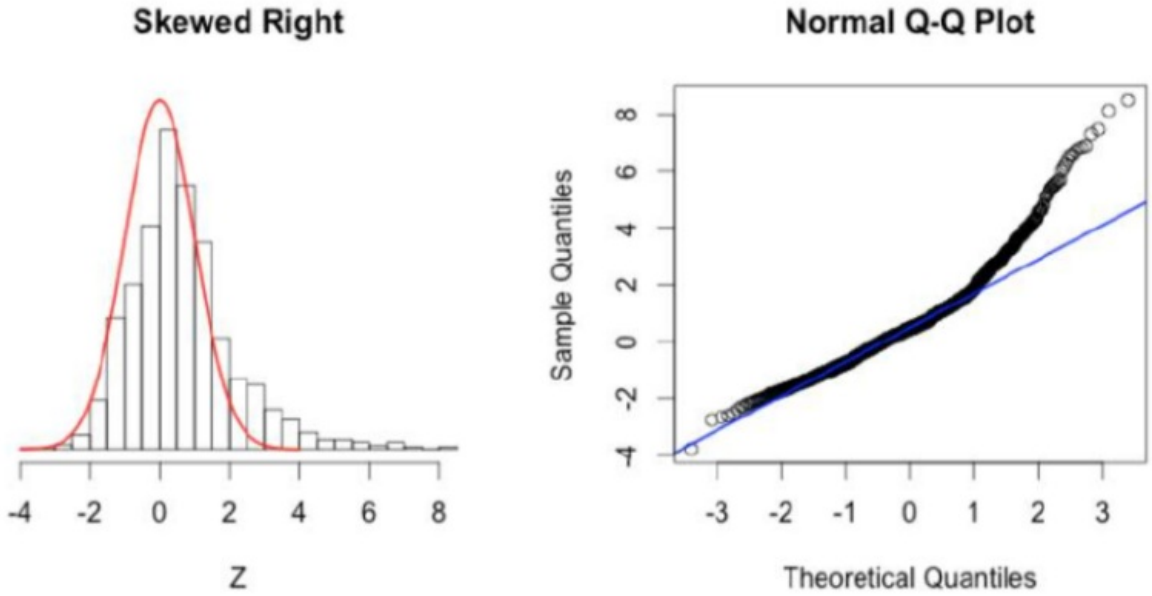
Q-Q plots are also used to find the skewnessof a distribution. When we plot theoratical quantiles on x - axis and the sample quantile whose distributio. We want to know on the Y-axis then we see a very pecular shape of a Normally distributed Q-Q plot for skewness.

- **If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not, then we can clearly say that the distribution has a longer tail to its left or simply it is left skewed.**



Left Skewed Q-Q plot for Normal DIstribution

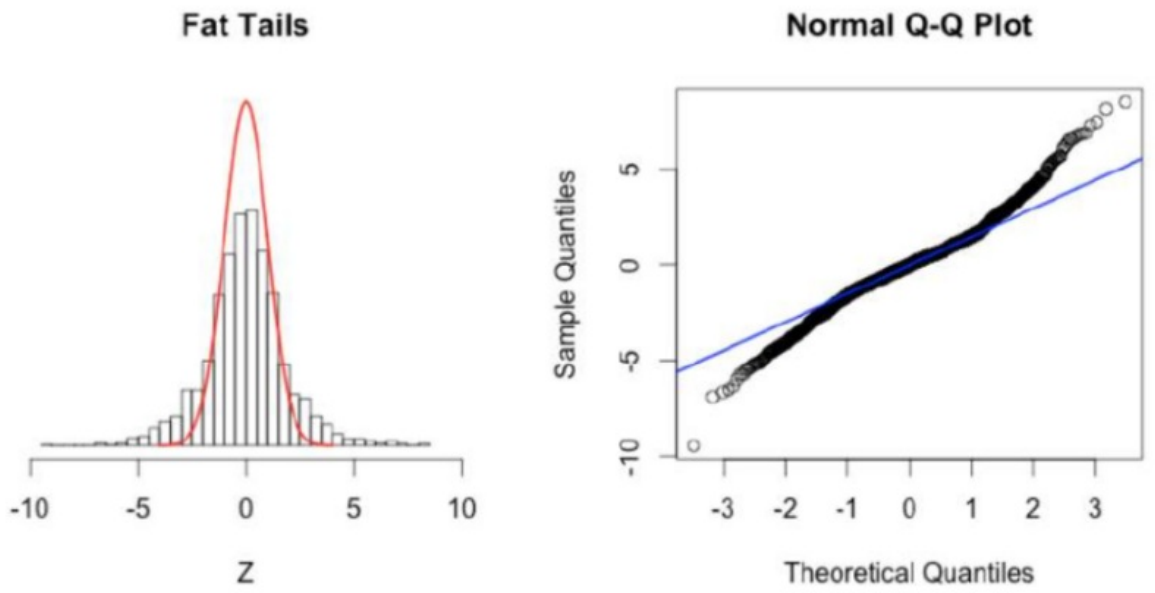
- **But when we see the upper end of the Q-Q plot to deviate from the straight line then the curve has a longer tail to its right and it is right-skewed/Positively skewed.**



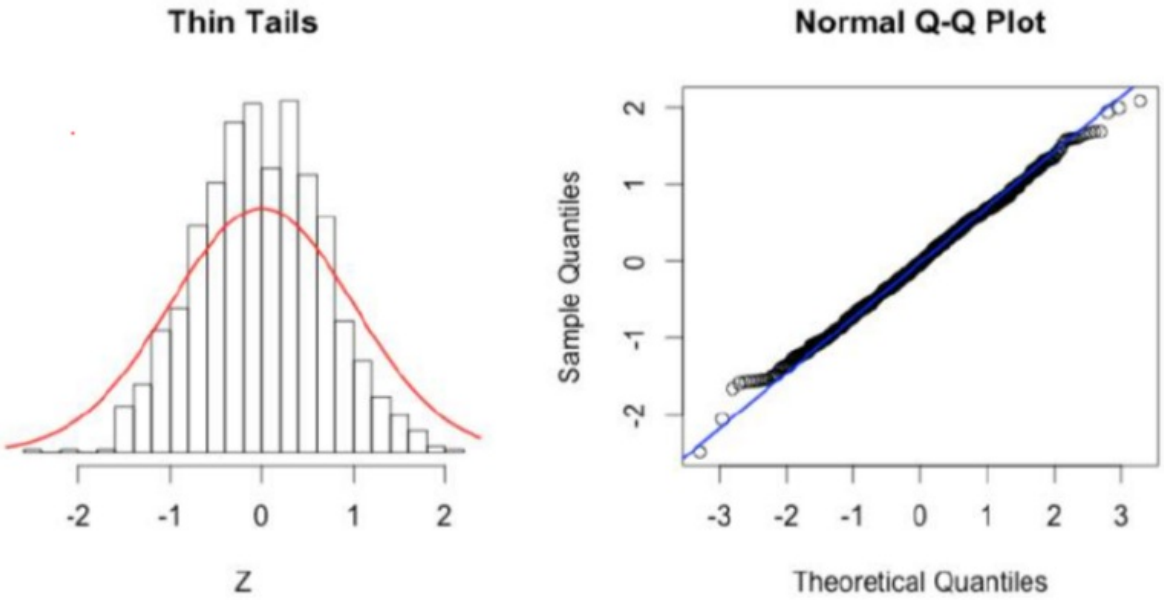
Right Skewed Q-Q plot for Normal DIstribution

Tailed Q-Q plot

Similarly, we can talk about the kurtosis of the distribution by simply looking at its Q-Q plot. The distribution with a fat tail will have both the end of the Q-Q plot to deviate from the straight line and its center follows a straight line, whereas a thin tailed distribution will form a Q-Q plot with a very less or negotiable deviaions at the end thus making it a perfect fir for the Normal distribution.



Fat-Tailed Q-Q plot for Normal DIstribution



Thin-Tailed Q-Q plot for Normal DIstribution

Implementation in python

In [23]:

```
import numpy as np
import pylab
import scipy.stats as stats

std_normal=np.random.normal(loc=0,scale=1,size=1000)

for i in range(1,101):
    print(i,np.percentile(std_normal,i))
```

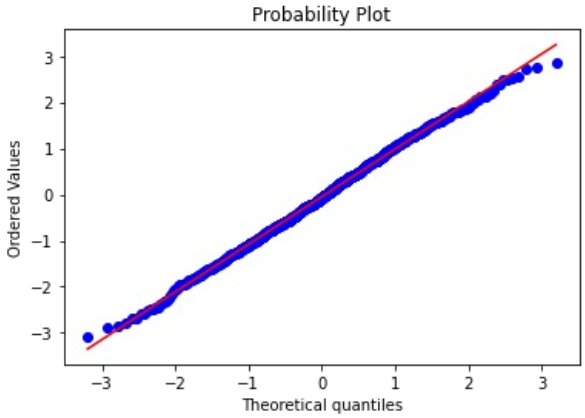
```
1 -2.482065887489731
2 -2.1451216271983897
3 -1.9363439603468986
4 -1.8172960190935723
5 -1.731422907595966
6 -1.6302542516584875
7 -1.5748723707436358
```

8 -1.5034661038755766
9 -1.453155724711262
10 -1.365919392150486
11 -1.297952223564574
12 -1.2690730132909092
13 -1.2258941193654933
14 -1.1744727111517315
15 -1.1287439527020788
16 -1.0796011137272248
17 -1.0429398512793748
18 -1.01314709727199
19 -0.9716253825322165
20 -0.9477368482620047
21 -0.9046708841498973
22 -0.8490908166668015
23 -0.8193311279744729
24 -0.7872761000801516
25 -0.7708932718256729
26 -0.717456174332018
27 -0.6980910661624181
28 -0.6698390507819495
29 -0.6430900407716886
30 -0.624363731459619
31 -0.5958282468661468
32 -0.5746971337399565
33 -0.5500330726672025
34 -0.5161723864726415
35 -0.49177875821989203
36 -0.457162656670487
37 -0.4299883331153871
38 -0.38088930456774656
39 -0.33896865604329374
40 -0.3178424208897661
41 -0.2843639404249636
42 -0.2660805336273624
43 -0.2478794324472404
44 -0.22741061722272013
45 -0.21365411842160664
46 -0.19125628402501393
47 -0.1598863924975108
48 -0.14369452157405754
49 -0.10738552332125208
50 -0.06963915133624504
51 -0.03924544932897052
52 -0.011775207742850076
53 0.014808760435810732
54 0.044923115930489976
55 0.06669142406532223
56 0.11269047935502417
57 0.13110349396785845
58 0.17602871672787607
59 0.20662341456360242
60 0.2398391442718054
61 0.2660880540866102
62 0.28282602442785015
63 0.3046457460664666
64 0.34979933121947626
65 0.3731685568313978
66 0.39106558715039696
67 0.41969358180200833
68 0.43882107826909256
69 0.46113633650596586
70 0.4926528704730457
71 0.5390350331098541
72 0.5787011925026659
73 0.626595598070074
74 0.6632510061416989
75 0.6802437648801951
76 0.7060862442586515
77 0.7322908808869916
78 0.7654046722381529
79 0.8024102669171669
80 0.8552764424515304
81 0.9095918242563404
82 0.9471925007967342
83 0.9875639176609977
84 1.0218668295189213
85 1.0726037284360013
86 1.1161898743833603
87 1.152849284173044
88 1.2152969642846416
89 1.2600668289934096
90 1.3005462583533378

```
91 1.341399097461625
92 1.3994936015024573
93 1.4964108328832983
94 1.554025334039373
95 1.5976936441875327
96 1.7008434748439434
97 1.7991708336018195
98 1.958993659432991
99 2.1781273233277822
100 2.8489971331130515
```

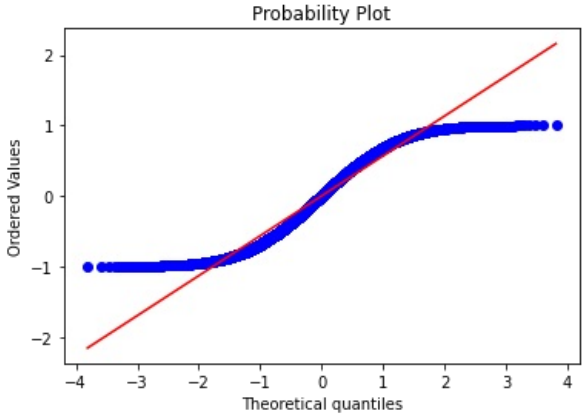
In [24]:

```
stats.probplot(std_normal,dist="norm",plot=pylab)
pylab.show()
```



In [25]:

```
# generate 100 sanples from N(20,5)
measurements = np.random.uniform(low=-1, high=1, size=10000)
#try size=1000

stats.probplot(measurements, dist="norm", plot=pylab)
pylab.show()
```



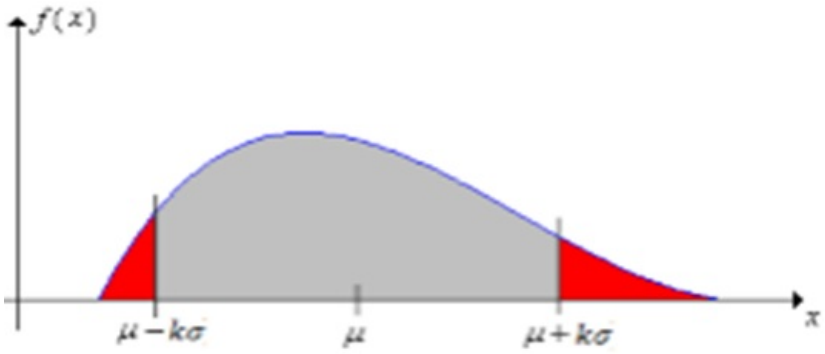
Chebyshev's inequality:

as we are familier with the Normal distribution concept and we know that if any random variable follow Normal distribution so we can make inference that within first standard deviation [μ-σ<=x<=μ+σ] 68% distribution lie and within second standard deviation [μ-2σ<=x<=μ+2σ] 95% distribution lie and within third distribution [μ-3σ<=x<=μ+3σ] 99.7% distribution lie.
But what if we don not know the distribution so how can i make inference about how much data lie in region.

The Chebyshev's inequality states that, for wide class of probability distribution, no more than a certain amount of values can be more than a certain distance from the mean, with the formula as follow.
chebyshevs inequality is used to find the proportion of observation you would expect to find within "k" standard deviation from the mean.

Condition on chebyshevs inequality

1. **We have a finite mean(μ)**
2. **We have standard deviation (σ) non-zero and finite.**



Formulae:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Where X is a random variable, μ is an expected value of X, σ is a standard deviation of X and k > 0.

We can write this in equation,
p(μ-kσ>=x>=μ+kσ) <= 1/k^2

or in other terms
p(μ-kσ<=x<=μ+kσ) <= 1-(1/k^2)

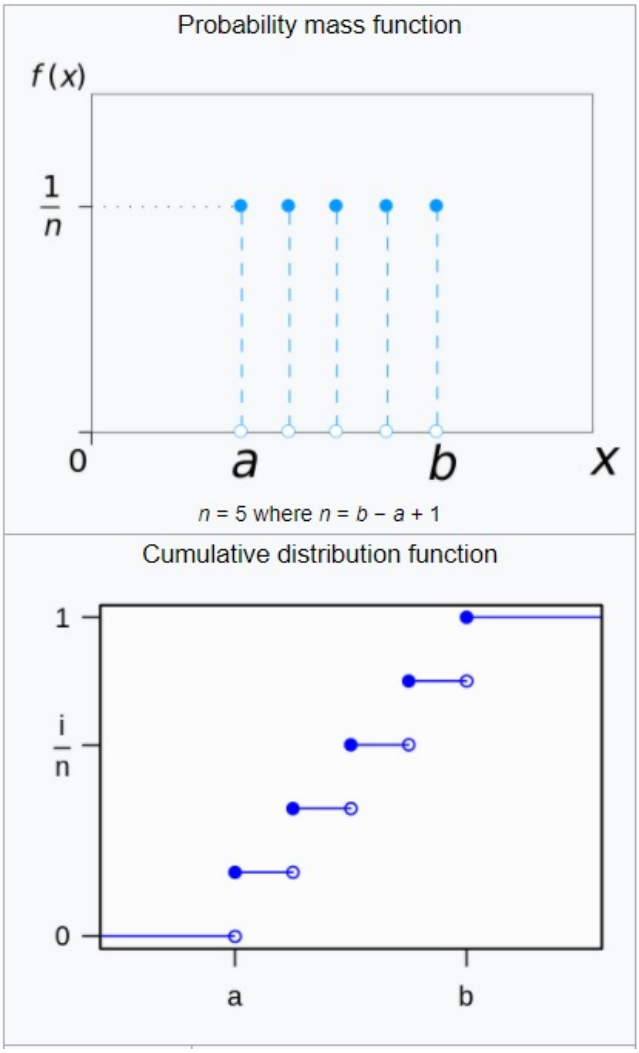
Uniform Distribution:

- **Perhaps one of the simplest and useful distribution is the uniform distribution uniform distribution is a distribution in which there equal probablities across all the values in the set.**
- **A uniform distribution, sometimes also known as a rectangular distribution, is a distribution that has a constant probablities**
- **The simplest probability distribution is the uniform distribution, which gives the same probability to any points of a set**
  ** A uniform distribution is of two types:**

  1. **Discrete uniform distribution**
  2. **Continous uniform distribution**

Discrete uniform distribution: The discrete uniform distribution is a symmetric probablities distribution. Where in a finite number of values are equally likely to be observed . Every one of n value has equal probablity 1/n.

discrete uniform



Notation	$\mathcal{U}\{a,b\}$ or $\operatorname{unif}\{a,b\}$
Parameters	a, b integers with $b \geq a$ $n = b - a + 1$
Support	$k \in \{a, a+1, \ldots, b-1, b\}$
PMF	$\frac{1}{n}$
CDF	$\frac{\lfloor k \rfloor - a + 1}{n}$
Mean	$\frac{a+b}{2}$
Median	$\frac{a+b}{2}$
Mode	N/A
Variance	$\frac{(b-a+1)^2 - 1}{12}$
Skewness	0

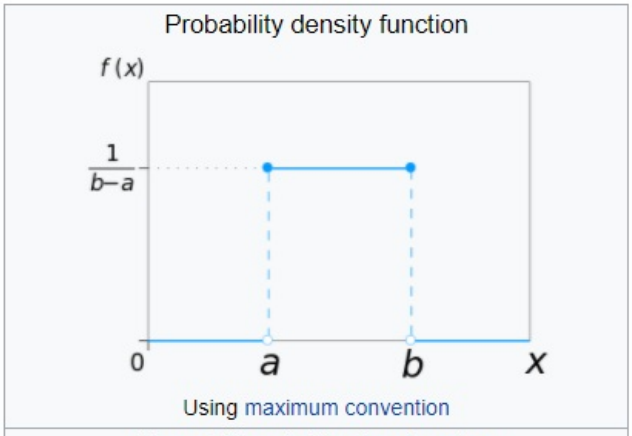
Example: Throwing a fair dice

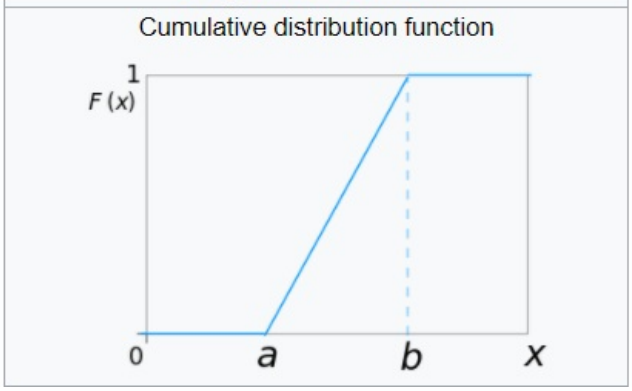
So, the possible values are 1,2,3,4,5 and 6 and each time the die is thrown the probablity of a given score is 1/6.

p(X=1)=p(X=2)=p(X=3)=p(X=4)=p(X=5)=p(X=6)=1/6

Continous uniform distributionDescribes an experiment where there is an arbitary outcome that lies between certain bounds are defined by the parameters, a and b, which are minimum and maximum values. The intervals can be either closed(e.g., [a,b]) or open (e.g.,(a,b)). The difference between the bounds define the interval length, all interval of same length on the distribution supports are equally probable.

Uniform





Notation	$\mathcal{U}(a, b)$ or $\operatorname{unif}(a, b)$
Parameters	$-\infty < a < b < \infty$
Support	$x \in [a, b]$
PDF	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
CDF	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$
Mean	$\frac{1}{2}(a + b)$
Median	$\frac{1}{2}(a + b)$
Mode	any value in (a, b)
Variance	$\frac{1}{12}(b - a)^2$
Skewness	0

Example: **Random number generator**

So the every variable in between the interval has equal chance of happening

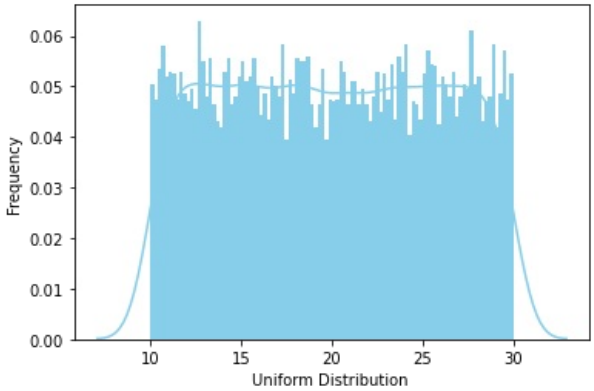
Implementation in python:

In [26]:

```
# import uniform distribution
from scipy.stats import uniform

data_uniform = uniform.rvs(size=10000, loc = 10, scale=20)

ax = sns.distplot(data_uniform,bins=100,kde=True,color='skyblue',hist_kws={"linewidth": 15,'alpha':1})
ax.set(xlabel='Uniform Distribution ', ylabel='Frequency')
```

Out[26]:

```
[Text(0, 0.5, 'Frequency'), Text(0.5, 0, 'Uniform Distribution ')]
```



Bernoulli distribution:

The Bernoulli distribution is the discrete probablity distribution of a random variable which takes a binary, boolean outout

- **1 with probablity p (Success)**
- **0 with probablity p-1 (Failure)**

The idea is that, whenever you are running an experiment which might lead either to a success or to a failure. You can associate with your success(labelled with 1) a probablity with p, while your failure (labelled with 0) will have probablity (1-p)

Formulae

$$P(n) = \begin{cases} 1-p & \text{for } n = 0 \\ p & \text{for } n = 1, \end{cases}$$

The probablity of the success p is the parameter of the bernoulli distribution, and if a discrete random variable X follows that distribution, we write:

X~BE(p)

Example: Imagine your experiment consists of flipping a coin and you will win if the output is tail. Since the coin is fair, you know that the probablity of having "tail" is p=1/2. Hence, once set "tail=1" and "head=0". You can compute the probablity of success as follows:

p(X=1)=f(1)=p=1/2

Example:Imagine you are about to toss a dice, and you bet your money on number 1: hence number 1 will be your success (labelled with 1), while any other number will be a failure(labelled with 0). The probablity of success is 1/6. If you want to compute the probablity of failure. You will do like this:

p(X=0)=f(0)=1-p=5/6

Binomial Distribution:

This distribution describes the behaviour of the output of "n" random experiments, each having a Bernoulli distribution with probablity P.

Example: flipping a fair coin. We said that our experiment consisted of flipping that coin once. Let us now modify it a bit and say that we are going to flip that coin 5 times. Among these trials, we will have some successes(tail,labelled as 1) and some failure(head, labelled as 0). Each trial has probablity 1/2 of success and 1/2 of failure. We might be interested in knowing what is the probablity of obtaining a given number X of successes. How shall we proceed?

Let us visualize this experiment:

```
H  H  H  T  T<br><br>
```

So, we flipped our coin 5 times and we lost and we lost in the first three trails, whilewe won in the last 2. Since we said that successes=tail=1 and failure=head=0, we can reframe it as follows:

```
0  0  0  1  1
```

Now, every trail is a bernoulli random variable, hence its probablity of occurence is P. If it is equal to 1, otherwise it is 0.Hence, if we want to compute the probablity of having the above situations(3 failure and 2 success). we will have something like that:

0(1-p) 0(1-p) 0(1-p) 1(p) 1(p)

and since trials are independent among each other:

```
p^2(1−p)^3
```

Generalizing this reasoning, if we had n trials with x successes:

0 0 0 0 0 0 0 . . . . . .0(n-x)times
1 1 1 1 1 1 1 1. . . . . .1(x times)

1-p 1-p 1-p 1-p 1-p 1-p. . . . . 1-p(n-x)times
p p p p p p p p. . . . . p(x)times

```
p^x(1−p)^n−x
```

Now a further concept need to be introduced. Indeed, so far we computed the probablity of having 2 successes exactly in the ordershown above. Neverthless since we are interested in having a given number of success regardless of the order they are given to us, we need to take into account all thepossible combination of having X successes

Namely, imagine we flip a coin 3 times and we want to compute the probablity of having 1 tail out of 3 trials. Hence, we will win in one of the following scenarious:

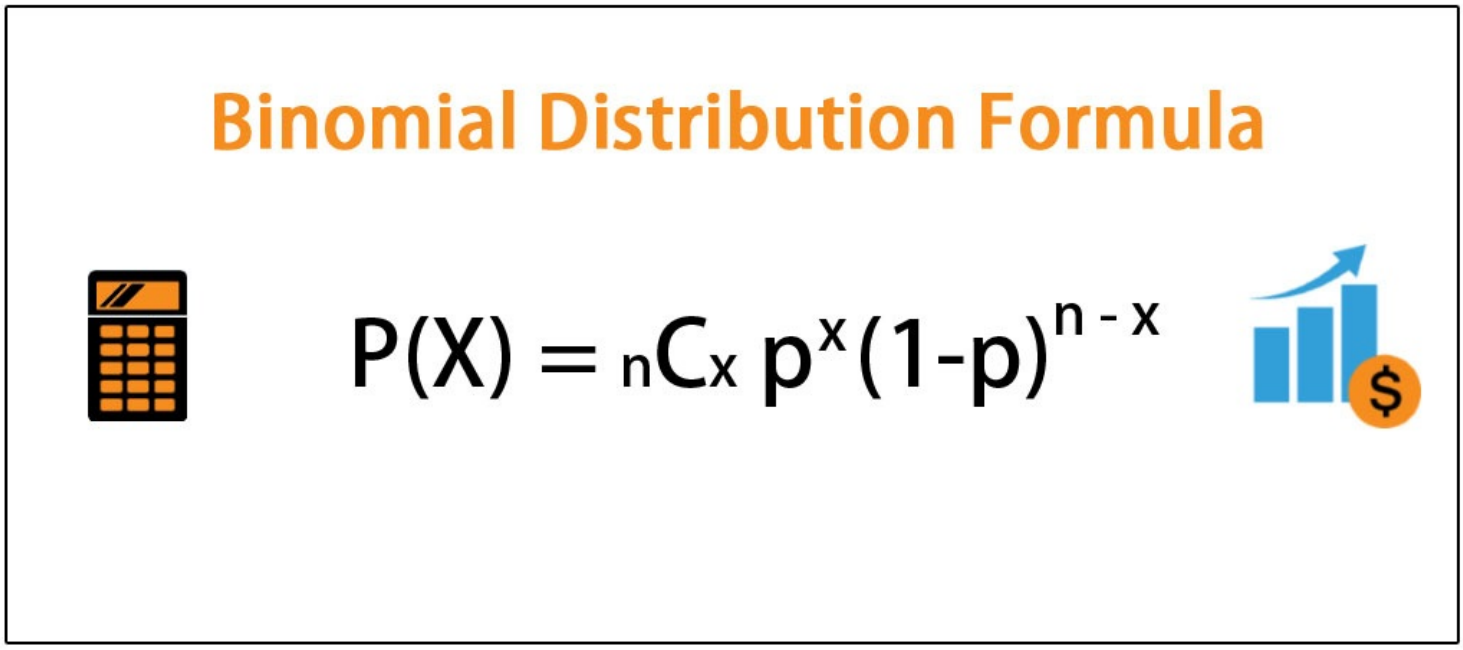
H H T
H T H
T H H
H T T
T H T
T T H

As you can see, there are three different combinations of outcomes which lead to a success. How can we incorporate this notion in our probablity function? The answer is the binomial coefficient is given by:

C(n,x)=n!/(x!(n-x)!)

Where n is the number of trials and X is the number of success of which we want to know the probablity of occurence.
So when we run n independent experiments, each having a bernoulli distribution with parameter p, and we want to know the probablity of having X success.

The probablity function will be:



Log Normal Distribution:

The log Normal distribution is a continous probablity distribution of a random variable whose logarithm is normally distributed.

In simpler words, if X is a log Normal distribution

X~N(μ,σ)
then y=ln(X) has a Normal distribution
If Y has a normal distribution, then the exponential function of Y
X=exp(Y)
has a Log Normal Distribution

Occurence and application

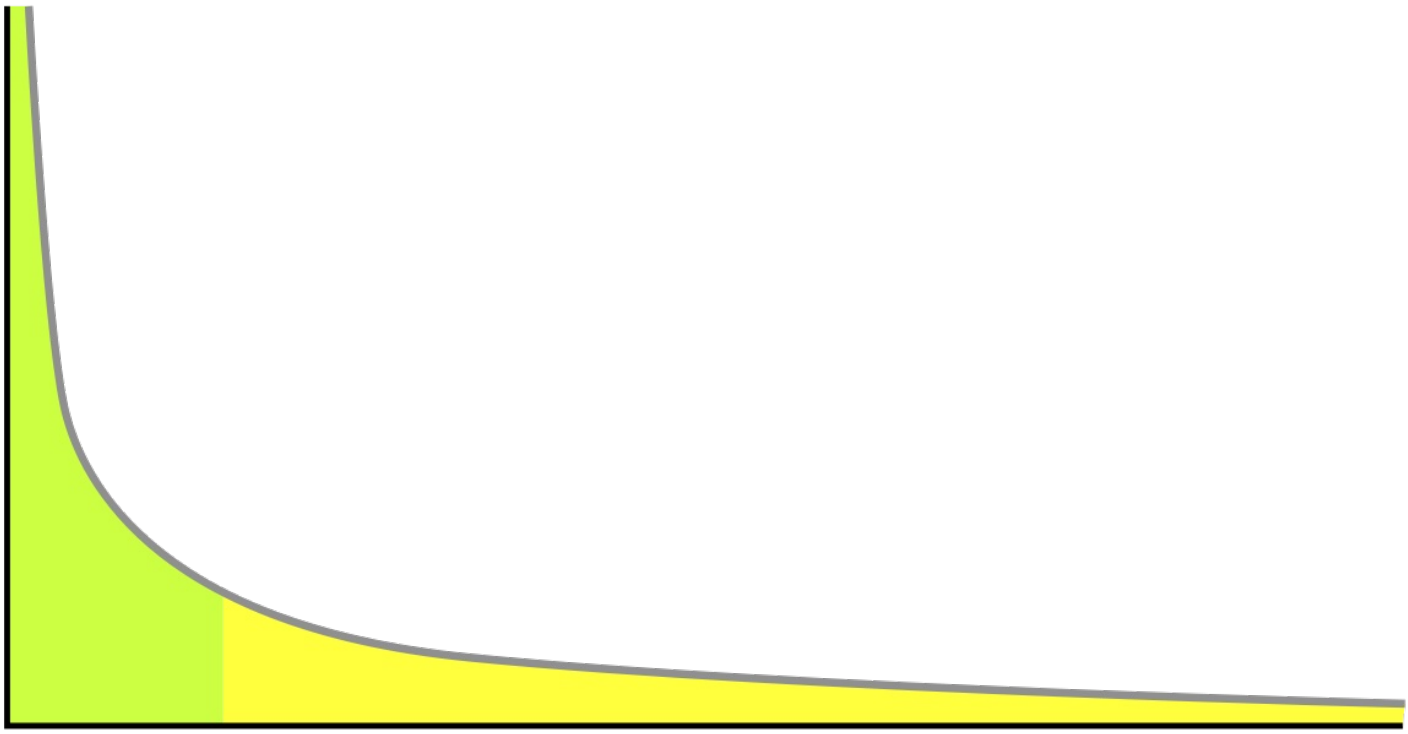
- **The length of commentsposted in internet discussion forums follows a log normal distribution**
- **The user's dwell time on the article**

How to check a random variable X follows log normal distribution:

In [ ]:

Power Law:

A power law is a functional relationship between two quantiles, where a relative change in one quantile result in a proportional relative change in other quantile, independent of initial size of those quantities. One quantities varies as a power of another.



Power law follows 80-20 rule means the 80% of distribution found in first 20% of range and 20% of distribution found in last 80%cof range

Co-variance:

- **covariance is used to measure relationship between two variables**
- **covariance is a measure of how much two random variable vary together. It is similar to variance, but variance tells you how a single variable varies, and co-variance tells you how two variable vary together**

Formulae:

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable a and y
x_i = data value of x
y_i = data value of y
$\bar{x}$ = mean of x
$\bar{y}$ = mean of y
N = number of data values

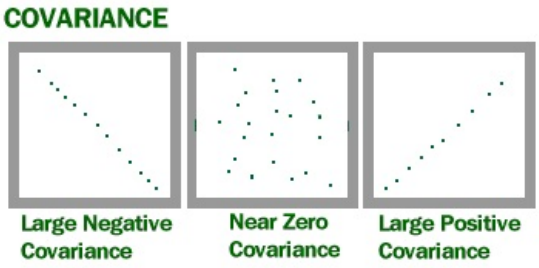
Types of Covariance:

Covariance can have both positive and negative values. Based on this, it has two types:

1. **Positive Covariance**
2. **Negative Covariance**

Positive Covariance
If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.

Negative Covariance
If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable corresponds to lesser values of another variable and vice-versa.



Limitations:

1. **If you change the unit of feature covariance will differ**
2. **covariance will only tells about the direction not the strength**

Mathematically solved

Question: The table below describes the rate of economic growth (xi) and the rate of return on the S&P 500 (y_i). Using the covariance formula, determine whether economic growth and S&P 500 returns have a positive or inverse relationship. Before you compute the covariance, calculate the mean of x and y.

Economic Growth % (x_i)	S&P 500 Returns % (y_i)
2.1	8
2.5	12
4.0	14
3.6	10

x = 2.1, 2.5, 4.0, and 3.6 (economic growth)

y = 8, 12, 14, and 10 (S&P 500 returns)

Find $\bar{x}$ and $\bar{y}$.

Solution:

$\bar{x} = \frac{\sum x_i}{n}$

$\bar{x} = \frac{2.1+2.5+4+3.6}{4}$

$\bar{x} = \frac{12.2}{4}$

$\bar{x} = 3.1$

$\bar{y} = \frac{\sum x_i}{n}$

$\bar{y} = \frac{8+12+14+10}{4}$

$\bar{y} = \frac{44}{4}$

$\bar{y} = 11$

Now, substitute these values into the covariance formula to determine the relationship between economic growth and S&P 500 returns.

xi	yi	xi − x̄	yi − ȳ
2.1	8	-1	-3
2.5	12	-0.6	1
4.0	14	0.9	3
3.6	10	0.5	-1

$$Cov(x,y) = \frac{(-1)(-3)+(-0.6)1+(0.9)3+(0.5)(-1)}{4-1} = \frac{3-0.6+2.7-0.5}{3} = \frac{4.6}{3} = 1.533$$

Pearson correlation coefficient(PCC):

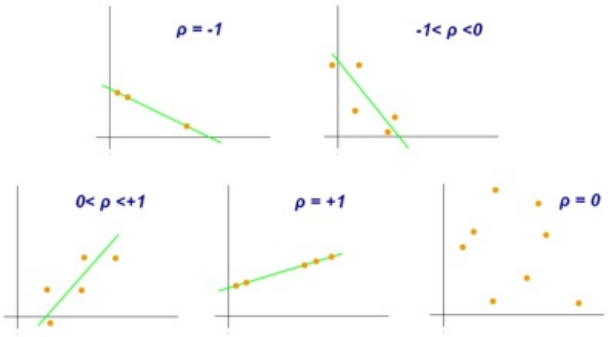
PCC are used in statistics to measure how strong a linear relationship is between two variables.
The range of the possible results of PCC is (-1,1) where:

1. **0 indicates no correlation**
2. **1 indicates a perfect positive correlation**
3. **-1 indicates a perfect negative correlation**

Formulae:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable



Limitation:

The PCC only follows linear relationship if in case your data follows linear method then PCC works amazingly but for non linear structure PCC gets fails.

Question 1: Calculate the linear correlation coefficient for the following data. X = 4, 8 ,12, 16 and Y = 5, 10, 15, 20.

Solution:

Given variables are,

X = 4, 8 ,12, 16 and Y = 5, 10, 15, 20

For finding the linear coefficient of these data, we need to first construct a table for the required values.

x	y	x^2	y^2	XY
4	5	16	25	20
8	10	64	100	80
12	15	144	225	180
16	20	256	400	320
Σ x = 40	Σ y =50	480	750	600

According to the formula of linear correlation we have,

$$r(xy) = \frac{(4\times600)-(40\times50)}{\sqrt{4(480)-40^2}\sqrt{4(750)-50^2}}$$

$$r(xy) = \frac{2400-2000}{\sqrt{1920-1600}\sqrt{3000-2500}}$$

$$r(xy) = \frac{400}{\sqrt{320}\sqrt{500}}$$

$$r(xy) = \frac{400}{17.89\times22.36}$$

$$r(xy) = \frac{400}{400} = 1$$

Therefore, r(xy) = 1

Spearman Rank correlation Coefficient:

SRCC covers some of the limitations of PCC>=. It does not carry any assumption about the distribution of the data. SRCC is a test that is used to measure the degree of association between two variables by assigning rank to the values of each random variable by assigning rank to the value of each random variable and computing PCC out of it.

X Y Rank(X) Rank(Y) 121 56 6 7 124 34 8 6 101 12 4 2 96 32 3 4 231 14 10 3 123 35 7 5 129 7 9 1 111 76 5 8 78 120 1 10 91 101 2 9

Given two random variable X and Y. Compute rank of each random variable, such that the least value has Rank1. Then apply the PCC on Rank(X),Rank(Y) to compute SRCC

$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}},$$

where

ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables,
$\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)$ is the covariance of the rank variables,
σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

$$r_R = 1 - \frac{6\Sigma_i {d_i}^2}{n(n^2-1)}$$

where n is the number of data points of the two variables and d_i is the difference in the ranks of the i^{th} element of each random variable considered. The Spearman correlation coefficient, ρ, can take values from +1 to -1.

- A ρ of +1 indicates a perfect association of ranks
- A ρ of zero indicates no association between ranks and
- ρ of -1 indicates a perfect negative association of ranks.
  The closer ρ is to zero, the weaker the association between the ranks.

SRCC range between -1 to +1 and works well with monotonically increasing and decreasing function.

Mathematically solved

Question: The following table provides data about the percentage of students who have free university meals and their CGPA scores. Calculate the Spearman's Rank Correlation between the two and interpret the result.

State University	% of students having free meals	% of students scoring above 8.5 CGPA
Pune	14.4	54
Chennai	7.2	64
Delhi	27.5	44
Kanpur	33.8	32
Ahmedabad	38.0	37
Indore	15.9	68
Guwahati	4.9	62

Solution: Let us first assign the random variables to the required data –

X – % of students having free meals
Y – % of students scoring above 8.5 CGPA

Before proceeding with the calculation, we'll need to assign ranks to the data corresponding to each state university. We construct the table for the rank as below –

State University	d_X = $Ranks_X$	d_Y = $Ranks_Y$	$d = (d_X - d_Y)$	d^2
Pune	3	4	-1	1
Chennai	2	6	-4	16
Delhi	5	3	2	4
Kanpur	6	1	5	25
Ahmedabad	7	2	5	25
Indore	4	7	-3	9
Guwahati	1	5	-4	16
				$\Sigma d^2 = 96$

Now, using the formula(with n = 7 here) –

$$r_R = 1 - \frac{6\Sigma_i {d_i}^2}{n(n^2-1)}$$

$$= 1 - \frac{6.96}{7.(49-1)}$$

$$= 1 - \frac{576}{336}$$

$$= -0.714$$

Such a strong negative coefficient of correlation gives away an important implication – the universities with the highest percentage of students consuming free meals tend to have the least successful results (and vice-versa). Similarly, we can solve all other questions.

In [ ]: