Statistics

Central limit Theorem:

In probability theory, the central limit theorem establishes that, in many situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.

Mean:

➔ Mean is the average value of total numbers of data and sometimes it is also called **Arithmetic Mean.**

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

➔ **Geometric Mean** is a type of average, usually used for growth rates like population and interest rates. While the Arithmetic mean adds the items where Geometric mean multiply the items. Also, you can get geometric mean only for positive numbers.

$$\bar{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = (x_1 x_2 \cdots x_n)^{1/n}$$

➔ **Harmonic Mean** very specific type of average. It is generally used when dealing with average of Units, like speed or other rates and ratios.

$$\bar{x} = n \cdot \left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$$

Python ex:

```
Import numpy as np
np.mean([1,2,3,4,5,5])
```

Median: is the centre value of given data.

In the data centre mean and median often tracked over tie to spot trend, which informs power costs predictions.

1. Arrange the values ascending order
2. Apply the formula
   (n+1)/2 if total count of numbers is odd
   Ex: 1,2,3,4,5
   Then, (5+1)/2 → 3rd value (3) of ordered number list the Median

   Python ex:
   Import numpy as np
   np.median([1,2,3,4,5])

3. **[(n/2)th term + {(n/2)+1}th]/2 if** total count of numbers is even.
   Ex: 1,2,3,4,5,5
   Then, [(6/2)+(6/2)+1]/2 → [3rd+4th]/2 → [3+4]/2 → 3.5 is the Median value.

   Python ex:
   Import numpy as np
   np.median([1,2,3,4,5,5])

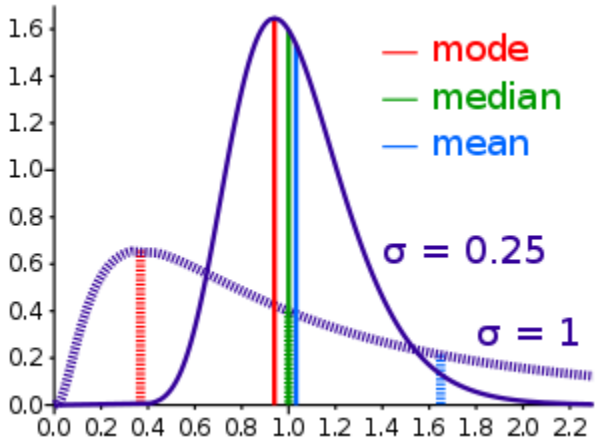
Mode: is the number that occurs most often with in a set of numbers.

- → Mode is the most common or occurred value in data set
- → Frequency of same value.
- → Maximum frequency of any value – Maximum number of frequencies
  1,2,3,**4,4,4**,5,5,6 Here Mode is 4 (Most common value)
  1,2,3,4,4,4,5,5,5,6 Here no Mode (No Most common value)

Python ex:

```
import statistics
list1 =[1, 2, 3, 3, 4, 4, 4, 5, 5, 6]
```

```
print(statistics.mode(list1))
```



Range: is the difference between the highest and lowest values within the set of data.

Range=Max number- Minimum number

Python ex:

```
for i in range(0,4):
    print(i, end='  ')
```

Standard deviation: Standard deviation used to quantify the amount of variation or dispersion of data values.

Or

How much the data is deviated from the mean is known a standard deviation.

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

Python formula: np.std([1,2,3,4,5])

Variance: is the expectation of squared deviation of random variables from its mean.

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ Sample Variance

Python formula: np.var([1,2,3,4,5])

Z –Score: is deviation of standard deviation from the mean, but more technically.

$$z_i = \frac{x_i - \bar{x}}{s}$$

S = Standard deviation

Co-Variance: is a measure of how changes in one variable are associated with change in second variable. Especially Covariance measures the degree to which variables are nearly associated.

Cov(x,y)= $\sum_{i=1}^{n}(\mathrm{Xi} - \mathrm{X})(\mathrm{Yi} - \mathrm{Y})/\mathrm{n}$

$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Correlation: is a statistical technic that can show whether and how strongly pairs of variables are related.

- ➔ Strength of association of 2 variables
- ➔ The direction of relation ship

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

Python Formula: from scipy.stats import pearsonr

Pearsonr("Variable name")

-1	0	1
Highly negatively Correlated	No Correlation	Highly Positively correlated

0 - 0.5 →weakly positively correlated

0.5 - 1 →Moderate positively correlated

0 - -0.5 →weakly Negative correlated

- 0.5 - -1 →Moderately Negative correlated

Quartiles:

- ➔ **Q1**
- ➔ **Q3**
- ➔ **IQR**
- ➔ **Skewness**
- ➔ **Kurtosis**
- ➔ **5 Number summary**
- ➔ **Box Plots**

Any value that 25% of total data

Q1= (n+1)/4 the value

6, 9, 11, 24, 36, 39, 43

Rule1: After dividing by 4 the whole total numbers take the value as quartile number.

Rule2: The value is ending with 2.5 the mean of previous and next value

Rule: The value is neither whole number or ending with 0.5, round the number by nearest value

1.33 = 1st value

1.75 = 2nd value

1.5 = 1st and 2nd values mean

Q3: ¾(n+1)

IQR: (Internal Quartile Range) is also called the mid spread or middle 50% is a measure of statistical dispersion being equal to the difference between 75% to 25%.

IQR=Q3-Q1

Python formula: iqr.("Variable name", "radius_mean")

Outlier: is an observation point that is distant from other observations.

Or

The value is disturbing the data from the data set or out of the range.

>Q3+1.5(IQR)

<Q1-1.5(IQR)

5 number summary:

- ➔ **Minimum**
- ➔ **Q1**
- ➔ **Median**
- ➔ **Q3**
- ➔ **Maximum**

Box Plots:

___ Maximum

----- Q3 (75 percentile)

------Median (50 percentile)

------Q1 (25 percentile)

----- Minimum

Skewness: is to check S**ymmetric/Asymmetric of a data**

Formula: 3(Mean-Median)/SD

Kurtosis: is a measure of whether the data are heavily tailed or lightly tailed to normal distribution

The kurtosis is the fourth standardized moment, defined as

$$\operatorname{Kurt}[X] = \operatorname{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{\operatorname{E}[(X-\mu)^4]}{(\operatorname{E}[(X-\mu)^2])^2},$$

Probability

Probability is chances of occurrence

P=X/E

Where E is total number of events

X is the desired outcome

Probability chances are between 0 to 1

Three type of probabilities

1. A Prior
2. Empirical
3. Subjective

A Prior: The probability of occurrence based on prior knowledge of the process involved. (Ex: 1/6 of kudos)

Empirical: The probability based on the observation (Ex: do survey)

Subjective: The probability based on your enthusiasm, mostly used on decision making. (Ex: Guessing 90% done, 95% done)

Coin → Head -- Probability = 1/2

Dice → 1 -- Probability = 1/6

Cards → k -- Probability = 4/52

	TV Price		
TV screen quality	High	Low	Total
HD	38	42	80
Normal	70	150	220
Total	108	192	300

➔ To get HD TV Probability = 80/300

Joint Probability: The probability of an occurrence involving two or more events.

➔ To get low price and Normal TV =150/300=0.5

Marginal Probability: The probability that consist of the set of joint probability

P= P(A^B)+ P(A1^B1)+ P(A2^B2)+------- P(An^Bn)

➔ HD ^Low, HD^High, Normal^High
42/100 + 38/300 + 70/300

General addition Rule Probability:

P(AUB)=P(A)+P(B)-P(A^B)

➔ HD TV screen or low screen
➔ 80/300 + 192/300 – 42/300

→What is the probability of a king or clubs while picking up a card from card desk?

4/52 + 13/52 - 1/52

→2 tomatos, 3 Potato, 4 onion, 5 brinjal

Potato or onion

3/14 + 4/14 -0/14

→in a dies >3 or even number , picking up a number >3 or even number

3/6 + 3/6 - 2/4

Compliment: The compliment of an event includes all events that are not part of the event.

A=X/E

A’ = 1 - X/E

If A=1/7 A’= 1 - 1/7 = 6/7

- ➔ Calculate compliment probability hearts from abck ofcards

  1 - 13/52=(52-13)/52=39/52

Conditional Probability:

The probability of an event A given information about the occurrence of another event B.

P(A/B)= P(A^B)/P(B)

P(B/A)= P(A^B)/P(A)

- ➔ P(Normal Screen Quality TV/low price)= 150/192
- ➔ P(HD Screen quality TV/High price)=38/108

Create a function for general addition rule probability

Distribution of a statistical data set as a listing of all function showing all possible values of the data and how often they occurred.

- ➔ Continues
- ➔ Decretive

1. Binomial Distribution
2. Poisson Distribution

Binomial Distribution: Any data set that will have two data sets that is binomial distribution.

P(X/n,II)=n!/X!(n-x)!.IIX(1-x)

Where n=total no of events

X= Desired outcome

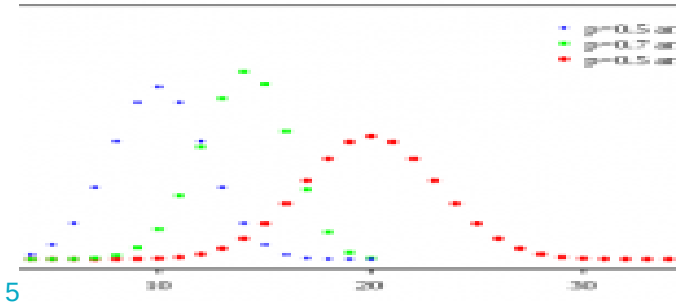
II = Possibility of desired out come

All the events in binomial distribution are independent one to another. Ex: Coin

The binomial distribution formula is:

$b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n-x}$
Where:
b = binomial probability
x = total number of "successes" (pass or fail, heads or tails etc.)
P = probability of a success on an individual trial
n = number of trials
Note: The binomial distribution formula can also be written in a slightly different way, because ${}_nC_x = n!/x!(n-x)!$ (this binomial distribution formula uses factorials (What is a factorial?). "q" in this formula is just the probability of failure (subtract your probability of success from 1).

$$P(X) = \frac{n!}{(n-X)!\,X!} \cdot (p)^X \cdot (q)^{n-X}$$



5

➔ If I toss a coin 10 times what is the probability of getting 5 tails?

$=10!/5!(10-5)!*(5/10)^5 *(1-5/10)^{10-5}$
$=252*0.3125*0.3125$
$=0.253$

➔ 80% of cosmetics are bought by women, what is the probability of 6 women of randomly pick 9 cosmetic byers?

$=9!/6!(9-6)!*(0.8)^6*(1-08)^{9-6}$
$=84*0.262*0.008$
$=0.17$

➔ What is the probability of getting 1 out of 5 times, if you roll a dice 10 times?

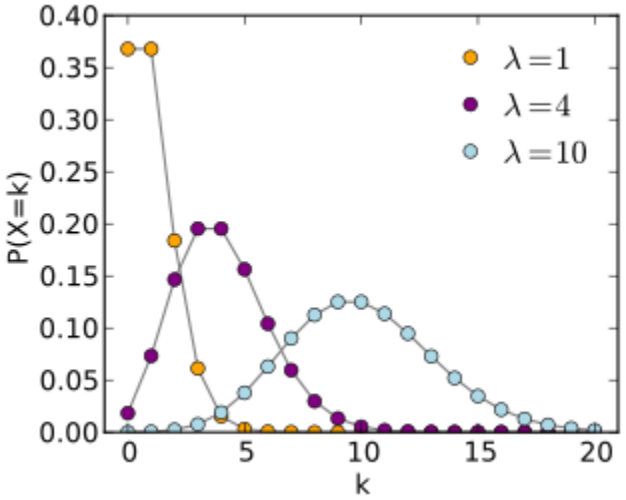
$=10!/5!(10-5)!*(1/5)^5 *(1-1/5)^{10-5}$

$=0.0264$

Poisson Distribution:

This probability gives us the probability of a given no of events happening in a fixed interval of time.

$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!}$$



Where e=2.71828, lambda=Known outcome, x=desired outcome

If it rains 3 times in Bangalore every one week, what is the probability of raining 4 times in coming week?

Normal Probability Distribution:

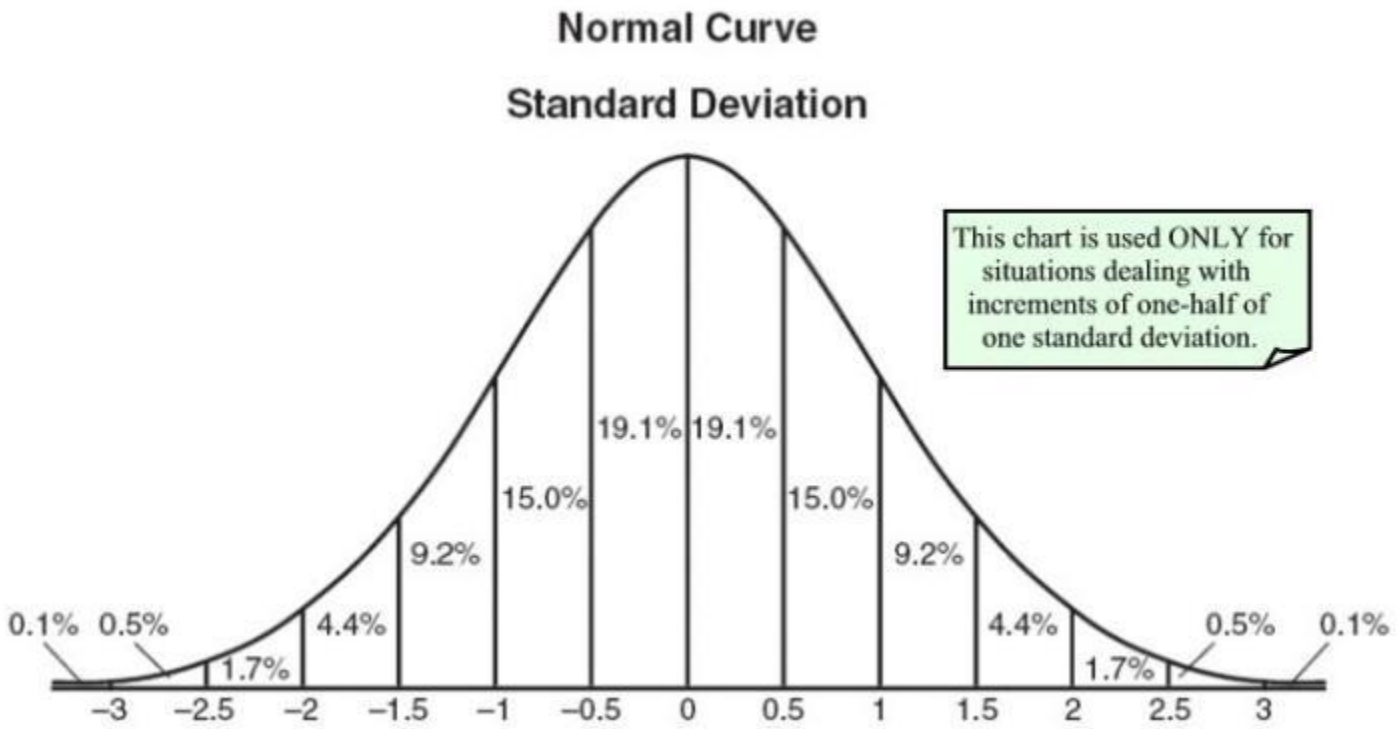
The Normal Probability Distribution is very common in the field of statistics.

Whenever you measure things like people's height, weight, salary, opinions or votes, the graph of the results is very often a normal curve.

The Normal Distribution

A random variable X whose distribution has the shape of a **normal curve** is called a **normal random variable**.

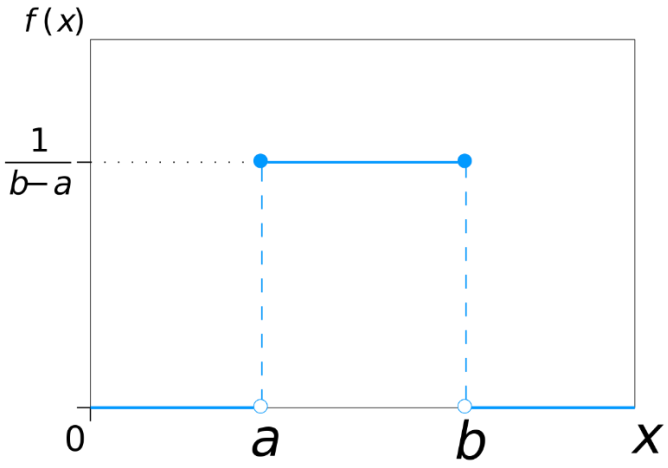
$$f(X)=\sigma 2\pi 1 e^{-(x-\mu)2/2\,\sigma 2}$$



Uniform Distribution:

In the uniform distribution value has the same probability of occurrence anywhere in the range between the smallest value and the largest value.

F(x)=1/b-a where b=largest value in data set and a=smallest value in data set



Exponential distribution:

This probability is used to model the length of time between two events.

The exponential distribution is one of the widely used continuous distributions. It is often used to model the time elapsed between events. We will now mathematically define the exponential distribution and derive its mean and expected value. Then we will develop the intuition for the distribution and discuss several interesting properties that it has

$F(x)= \lambda e^{-\lambda x}$

1. If avg rate of arrival of patients in a hospital is 30 patients per hour. What is the probability that next patient will arrive after 12 min?

   $=(2.71828)^{-30*(12/60)} *30$

$=0.074$