

So----- you actually did it. --- hmmm----- Damn Bro-----

Probability & Statistics for M.L3

Probability is a measure of how likely an event is to occur.
 $P(\text{---})$.

Complement of Probability

$$P(\text{Event}') = 1 - P(\text{Event})$$

$$P(\text{Event}') = 1 - P(\text{Event})$$

Sum of probabilities

for disjoint events (mutually exclusive)
for example $\therefore$

$$P(\text{Soccer}) = 0.3$$

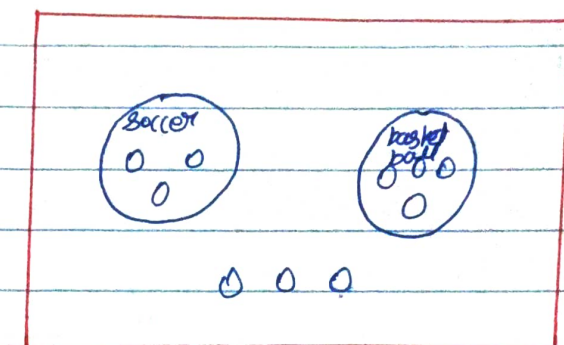
$$P(\text{Basketball}) = 0.4$$

$$P(\text{Soccer or Basketball}) = 0.7$$

$$= \frac{\text{Soccer or basketball}}{\text{Total}}$$

Imagine if in a school the probability that a kid plays soccer is 0.3 & the probability that a kid plays basketball is 0.4, then the probability of a kid playing soccer or basketball is actually the sum of the probabilities.

Using venn diagram



$$P(\text{Soccer} \cup \text{Basketball}) = P(\text{Soccer}) + P(\text{Basketball})$$

$$P(A \cup B) = P(A) + P(B)$$

Q $\Rightarrow$ A dice is rolled, what is the probability of obtaining an even number or a 5?

Solⁿ $\Rightarrow$ $P(A) = \text{obtaining an even number} = \frac{3}{6} \Rightarrow \frac{1}{2}$

$$P(B) = \text{obtaining 5} = \frac{1}{6}$$

$$P(A \cup B) = \frac{1}{2} + \frac{1}{6} = P(A) + P(B)$$

$$\Rightarrow \frac{8}{12} \Rightarrow \frac{2}{3}$$

For joint events or similar events. (non-mutually exclusive)
 $\Rightarrow$ Let's say the probability of rainy is 80% & probability of windy is 70% then what is the probability of the weather being rainy or windy, now if you do it by the disjoint definition you will get 150% which is wrong, so it is different for joint events.

Ex $\Rightarrow$ At a school, kids play Soccer & Basketball

$$P(S) = 0.6, \quad P(B) = 0.5$$

What is the probability that a kid plays soccer or basketball?

$$|S| = 6 = \text{no of kids playing soccer}$$

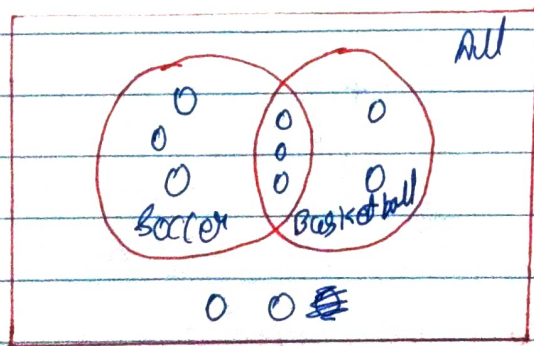
$$|B| = 5 = \text{no of kids playing basketball}$$

$$|S \cap B| = \text{No of kids who play both soccer \& basketball} = 3$$

The total no of kids = ~~10~~ 10

$$\begin{aligned} \text{No of kids who play soccer or basketball} \\ = 6 + 5 - 3 \Rightarrow \boxed{8} \end{aligned}$$

Venn diagram



Now,

$$|S \cup B| = |S| + |B| - |S \cap B|$$

$$= 6 + 5 - 3 = 8$$

So, $P(S \cup B) = P(S) + P(B) - P(S \cap B)$

$$= 0.6 + 0.5 - \frac{3}{10}$$

$$\Rightarrow \boxed{0.8} \text{ Ans}$$

Independence

Independence happens when the occurrence of one event doesn't affect the probability of the occurrence of another event.

eg $\Rightarrow$ If you toss a coin twice whatever happened on the first throw doesn't affect the outcome of the second one.

Quiz $\Rightarrow$

Imagine a school where there are 100 kids, 40 of them play soccer & 60 of them play nothing,

Shuffle the kids & put them in a room one containing 30 kids & other⁽²⁾ containing 70 kids.

what is the probability of finding a randomly selected kid in room 1 who plays football

Solⁿ $\Rightarrow$ we split the kids in 40, 60 & 0 each room will have 40% football players so,

in room one there are 30 players
Hence,

$$40 = \frac{?}{30} \times 100 = \boxed{12}$$

~~12 of them play football.~~

$$P(E) = \frac{12}{30} \Rightarrow \boxed{\frac{2}{5}}$$

$$P(\text{Soccer} \cap \text{Room 1}) = P(S) \cdot P(R_1) \\ \Rightarrow P(\text{Soccer and Room 1}) \Rightarrow \boxed{0.4 \times 0.3} \Rightarrow \boxed{0.12}$$

Product rule

$$P(A \cap B) = P(A) \cdot P(B)$$

Q $\Rightarrow$ what is the probability of landing on heads five times?
Solⁿ $\Rightarrow$

$$P(5 \text{ heads}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \boxed{\frac{1}{32}}$$

Q $\Rightarrow$ what is the probability of getting 10 sixes?

$$P(10 \text{ sixes}) = \left(\frac{1}{6}\right)^{10}$$

Birthday Problem

Quiz $\Rightarrow$ You have 30 friends at a party.
What do you think is more likely.

- That there exist's two people with same birthday with the same birthday.
- That no two of them have the same birthday.
- Assume that year has 365 days, nobody has a birthday on feb 29)

Solⁿ ⇒

Probability that everyone has a different birthday.

$$\Rightarrow \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \cdots \frac{336}{365}$$

$$\Rightarrow \boxed{0.294}$$

Probability that two of them have same birthday

$$\text{is } \Rightarrow 1 - 0.294$$

$$\Rightarrow \boxed{0.706}$$

Conditional Probability

The Probability of an event ^{happening} given that another event has already occurred.

Q ⇒ What is the probability of landing on heads twice?
if the first toss is already heads.

$$P(HH | 1^{st} \text{ is } H) = \frac{1}{2}$$

We know that the first one is head so second one can either be heads or tails
so the probability is 0.5.

→ it says the probability of ~~##~~ Heads, Heads
given that 1st toss is head.

Dice example $\therefore$ 2 die

$\Rightarrow$ find the probability that on first roll we get 6 & the sum = 10?

So, $P(\text{1st is 6} \cap \text{sum} = 10) =$
using product rule

$$P(\text{1st is 6}) = \frac{1}{6}$$

$$P(\text{sum} = 10 | \text{1st is 6}) = \frac{1}{6} \left\{ \begin{array}{l} \rightarrow \text{This means The probability} \\ \text{of the first sum is 10 but} \\ \text{among the outcomes where} \\ \text{first digit is 6.} \end{array} \right.$$

$$\text{So, } P(\text{1st is 6} \cap \text{sum} = 10) \Rightarrow P(\text{1st is 6}) \cdot P(\text{sum} = 10 | \text{1st is 6})$$
$$= \frac{1}{6} \cdot \frac{1}{6} \Rightarrow \boxed{\frac{1}{36}}$$

in general,

$$\boxed{P(A \cap B) = P(A) \cdot P(B|A)} \rightarrow \text{Product rule.}$$

$\hookrightarrow P(B|A) = P(B)$ when the case is independent.

Bayes Theorem

Intuition $\Rightarrow$

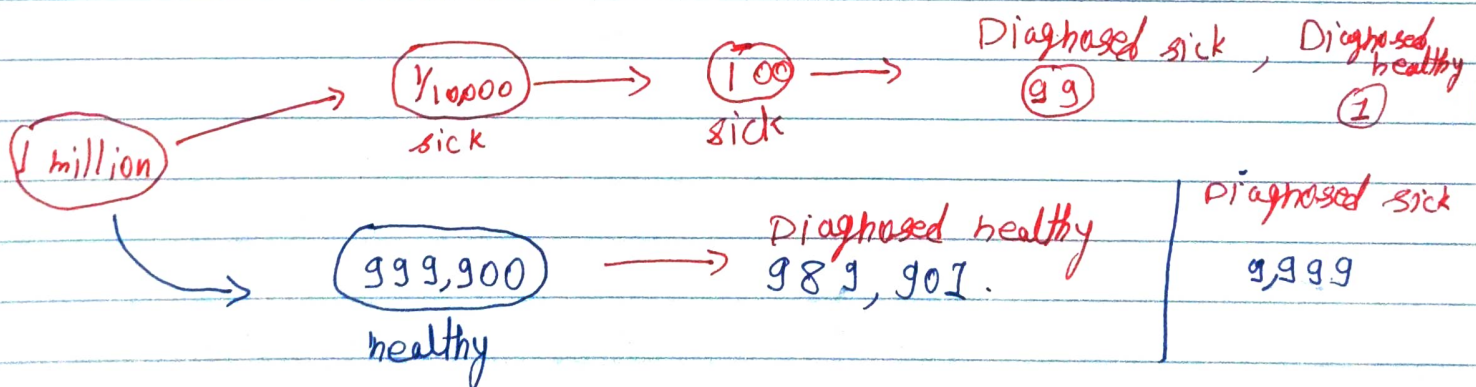
So, let's say there is a rare disease going on & you want to get tested for it.

The test works 99% of the times & miss diagnose 1% of the time

$\Rightarrow$ Now, you are tested +ve, how you have to calculate the probability that, you are +ve given that ~~the~~ you tested +ve.

Let's say we have 1 million people & the illness only affects 1/10,000 people. Now, the test is 99% effective.

So,



Now, what is the probability, that you are actually Sick given that you are actually Sick?

$$\text{So, } P(\text{sick} | \text{diagnosed sick}) = \frac{99}{99 + 9999} = \boxed{0.0098}$$

$$= \frac{\text{sick \& diagnosed sick}}{\text{sick \& diagnosed sick} + \text{healthy \& diagnosed sick}}$$

Bayes theorem formula

$$P(A/B) = \frac{P(A) \cdot P(B/A)}{P(A) \cdot P(B/A) + P(A') \cdot P(B/A')}$$

Bayes Theorem - Spam example

data $\Rightarrow$

Total emails = 100

Spam emails = 20

Spam emails that contain word
lottery = 14

Not spam emails that contain
word lottery = 10

Q $\Rightarrow$ $P(\text{Spam} | \text{lottery})$, what is the probability that an email is spam given that it has the word lottery in it?

According to Bayes Theorem,

$$P(A/B) = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A') \cdot P(B|A')}$$

$$\Rightarrow \frac{P(\text{Spam}) \cdot P(\text{lottery} | \text{Spam})}{P(\text{Spam}) \cdot P(\text{lottery} | \text{Spam}) + P(\text{Not spam}) \cdot P(\text{lottery} | \text{Not spam})}$$

$$P(\text{Spam}) \Rightarrow \frac{20}{100} \Rightarrow \boxed{0.2}$$

$$P(\text{lottery} | \text{Spam}) = \frac{14}{20} \Rightarrow \boxed{0.7}$$

$$P(\text{Not spam}) = \frac{80}{100} = \boxed{0.8}$$

$$P(\text{lottery} | \text{Not spam}) = \boxed{0.125}$$

Putting these in formula.

$$P(\text{spam} | \text{lottery}) = \frac{(0.2) \times (0.7)}{(0.2 \times 0.7) + (0.8 \times 0.125)}$$

$$= \frac{0.14}{0.24} = \boxed{0.583}$$

Bayes Theorem

prior & Posterior

Prior

Event

Posterior

$P(A)$

E

$P(A|E)$

Here the Prior is the probability that an event happens E of something then an event haven & the posterior is the probability that ^{something} an event happens with regards to that event.

for example in the spam email,

we found the probability of an email being spam then an event E happens that says there are emails with word lottery in them that are spam. So after that we found the probability of a spam email given it has the word lottery in it, which was better than simply the probability of an email being spam.

Bayes Theorem :- The naive Bayes model

looking at the email spam classification again :-

Prior

Event

Posterior

$$P(\text{spam}) = \frac{\text{spam envelope}}{\text{spam envelope} + \text{ham envelope}}$$

Email contains word
"lottery"

$$P(\text{spam} | \text{lottery}) = \frac{\text{lottery spam envelope}}{\text{lottery spam envelope} + \text{lottery ham envelope}}$$

To, make a classifier that gives the probability that an email is spam or not, we need many it to work with many words, not only a single word "lottery". Let's add another example

Email contains word
"winning"

$$P(\text{spam} | \text{winning}) = \frac{\text{winning spam envelope}}{\text{winning spam envelope} + \text{winning ham envelope}}$$

$$\text{Email containing "lottery" \& "winning"} \quad P(\text{spam} | \text{lottery} \& \text{winning}) = \frac{\text{lottery \& winning spam envelope}}{\text{lottery \& winning spam envelope} + \text{lottery \& winning ham envelope}}$$

so, according to Bayes theorem :-

$$P(\text{spam} | \text{lottery} \& \text{winning}) = \frac{P(\text{spam}) \cdot P(\text{lottery} \& \text{winning} | \text{spam})}{\text{Numerator} + P(\text{ham}) \cdot P(\text{lottery} \& \text{winning} | \text{ham})}$$

Here $\Rightarrow$ ham = Not Spam

Now, $P(\text{lottery \& winning} | \text{spam}) = \frac{\text{Spam emails with words "lottery" \& "winning"}}{\text{Spam emails.}}$

Well, now imagine if there are 100 words.

Email contains $(w_1, w_2, w_3, \dots, w_{100})$

$$P(\text{spam} | w_1, \dots, w_{100}) = \frac{P(\text{spam}) \cdot P(w_1, \dots, w_{100} | \text{spam})}{P(\text{spam}) \cdot P(w_1, \dots, w_{100} | \text{spam}) + P(\text{ham}) \cdot P(w_1, \dots, w_{100} | \text{ham})}$$

$$P(w_1, \dots, w_{100} | \text{spam}) = \frac{\text{Spam emails with } w_1, \dots, w_{100}}{\text{Spam emails}}$$

But there is a caveat, asking for an email to contain a hundred spam words is ~~just~~ ludicrous, maybe there is no email like that;

So, what are we gonna do?

So, we can't calculate $P(w_1, \dots, w_{100} | \text{spam})$ but we can estimate it,

by taking a naive assumption,

"Let's say The appearance of word 'lottery' & 'winning' are independent from each other"

This is not true because words are dependent on each other, even though we are wrong let's just assume it & ~~save~~ do the maths.

at applying conditional probability when the events are independent we can ~~say~~ write

$$P(\text{lottery winning} | \text{spam}) \text{ as}$$

$$P(\text{lottery} | \text{spam}) \cdot P(\text{winning} | \text{spam})$$

because

$$P(A \cap B) = P(A) \cdot P(B)$$

So for "n" number of words

$$P(\text{spam} | w_1, \dots, w_n) = \frac{P(\text{spam}) \cdot P(w_1 | \text{spam}) \dots P(w_n | \text{spam})}{\text{Numerator} + P(\text{ham}) \cdot P(w_1 | \text{ham}) \dots P(w_n | \text{ham})}$$

And this is the Naive Bayes Algorithm.

Now, Solving the spam email example :-

$$P(\text{spam}) = 0.2 \quad | \quad P(\text{lottery} | \text{spam}) = 0.7 \quad | \quad P(\text{winning} | \text{spam}) = 0.75$$

$$P(\text{not spam}) = 0.8 \quad | \quad P(\text{lottery} | \text{ham}) = 0.125 \quad | \quad P(\text{winning} | \text{ham}) = 0.1$$

$$P(\text{spam} | \text{lottery}) = \frac{P(\text{spam}) \cdot P(\text{lottery} | \text{spam}) \cdot P(\text{winning} | \text{spam})}{\text{Numerator} + P(\text{ham}) \cdot P(\text{lottery} | \text{ham}) \cdot P(\text{winning} | \text{ham})}$$

$$= \frac{0.2 \times 0.7 \times 0.75}{(0.2 \times 0.7 \times 0.75) + (0.8 \times 0.125 \times 0.1)}$$

$$\Rightarrow \frac{0.105}{0.105 + 0.2} = \frac{0.105}{0.305} = \boxed{0.344}$$

Probability in machine learning

Models like sentiment analysis model that will tell you whether a text is happy or sad, image recognition or where the model's tell you the probability of an image being cat or dog, or Email spam classifier that tells you the probability of an email being spam or not.

For example in a face generation model we generate a group of pixels such that the resulting image looks like a human face, what we do is generate pixels such that $P(\text{Face}|\text{Pixels})$ is high.

meaning the probability of the image being a face given some pixels

Random variables

Random variables are different they can take many values.
"We are quantifying the random processes"

ex $\Rightarrow$ The no of heads you will get, when you toss a coin 3 times. it may ~~at~~ vary everytime.

ex $\Rightarrow$

$X =$ no of heads

flip a coin

1 heads

$x=1$

0 heads

$x=0$

$$P(x=1) = 0.5$$

$$P(x=0) = 0.5$$

Here, x is a random variable.

e.g

let's say

X = Number of heads in 10 coin tosses

then, we can use it as,

$$P(X=3)$$

→ This means "that is
"The probability of 3 heads in 10 coin
tosses"

Discrete random variables

Finite no of values

let's say you flip a coin & the probability of getting heads is your random variable, so it can be in between (0-10) it's finite.

Continuous random Variables

Infinite no of values

let's say you are waiting for your food to arrive in restaurant, the time it takes can be 1.031 minutes, 1.5 minutes etc it's not defined it's ∞.

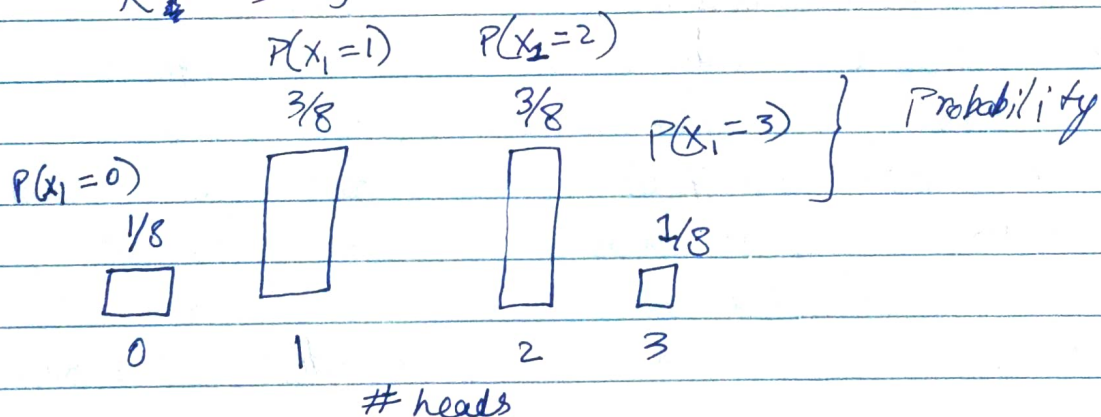
Random variable has uncertain outcomes.

Probability distribution (discrete)

Take all the possible scenarios that can happen put them on this horizontal axis & for each one of them, look at the probability that they happen, This forms a Probability distribution.

X_1 : number of heads in 3 coin tosses

$X_1 = 0$
 $X_1 = 1$
 $X_1 = 2$
 $X_1 = 3$ } you can get (0, 1, 2, 3) heads
so let's calculate the probability
& plot it on the graph.



Probability Mass function.

in the given example above for each x from (0-5) you have the probability that $X_3 = x$.

$$P_x(x) = P(X_3 = x), \quad x = 0, 1, 2, 3, 4, 5$$

it must satisfy

$$P_x(x) \geq 0$$

$$\sum_x P_x(x) = 1$$

Binomial Distribution

when you plot the value & Probability of each one of the random variable in a histogram you get Binomial Distribution. eg $\Rightarrow$

$\Rightarrow$ what is the probability that if I flip 5 coins, 2 of them land in heads?

$$\frac{10}{32}$$

Binomial Coefficient

in general :

$\binom{n}{k}$ counts all the combination for landing k heads in n coin tosses.

Property :

$\binom{n}{k} = \binom{n}{n-k}$. This is why PMF of a fair coin is symmetrical.

General PMF for $X = \text{no of heads in 5 coin tosses}$?

Your coin has $P(H) = P$

Event : $X = x$: $x \rightarrow$ heads in 5 tosses

All the possible orders $\rightarrow \binom{5}{x}$ $P^x (1-P)^{5-x}$ $x = 0, 1, 2, 3, 4, 5$.
Probability of seeing $5-x$ tails
 $\rightarrow$ Probability of seeing x heads

$$P_X(x) = \binom{5}{x} p^x (1-p)^{5-x}, \quad x = 0, 1, 2, 3, 4, 5$$

$\hookrightarrow x$ follows a binomial distribution.

We denote it as binomial of 5 & p,

~~$X \sim (5, p)$~~

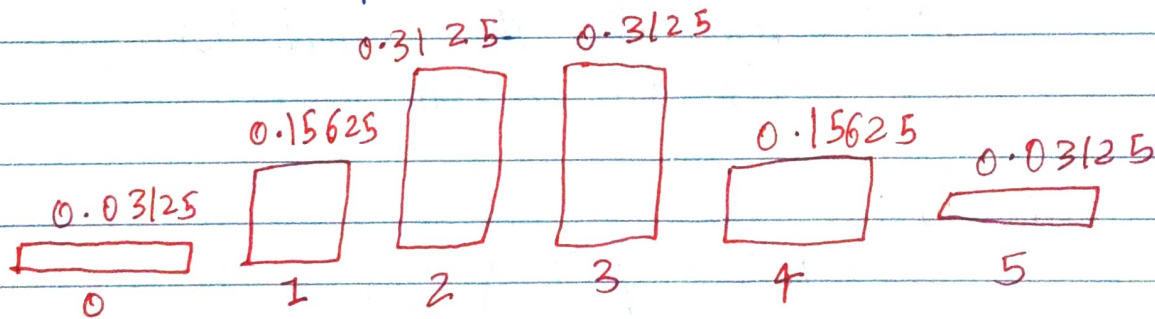
$X \sim \text{Binomial}(5, p)$

$\rightarrow P(H)$

$\rightarrow$ Number of flips

$\hookrightarrow$ in general this "tilde" symbol means that the variable X follows the distribution to the right of the expression.

PMF graph $\Rightarrow$



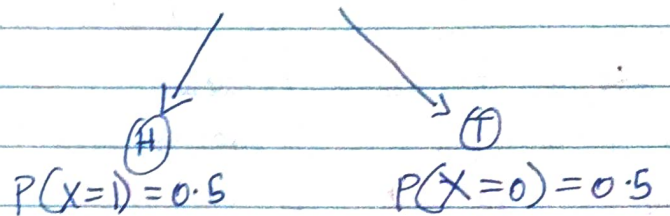
In general PMF for X : no. of heads in n coin tosses?
 Your coin has $P(H) = p$
 Event: $X = x$: x heads in n tosses.

$$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

$X \sim \text{Binomial}(n, p)$

Bernouli Distribution

Let's say, X = Random variable, X = No. of heads
you flip a coin



⇒ We're going to call it an experiment where we get a **Success** if it lands on heads & **failure** if it lands on tails.

$X \sim \text{Bernouli}(P)$ ⇒ here P is the probability of success

"The above given example is of Bernouli Distribution which has one parameter

Probability Distribution (Continuous)

The difference btw continuous & discrete distribution is

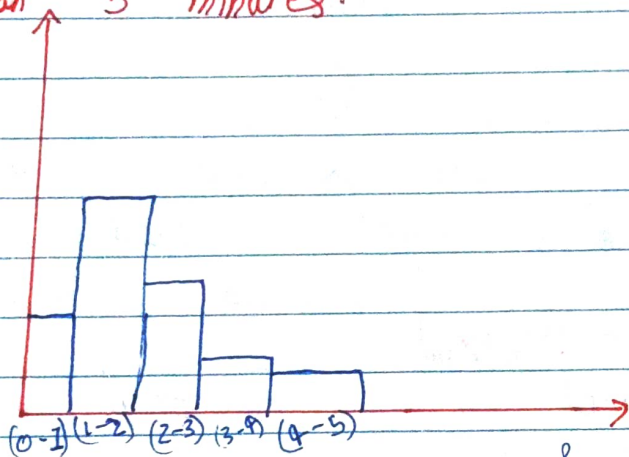
⇒ In the Discrete distribution your events always form a list, eg if you toss a coin you can get heads $[0, 1, 2, 3]$ times

⇒ In the Continuous distribution your events don't form a list, they are actually intervals, e.g if your random variable is the amount of time that I wait to get my food, it can be 1.002 minutes, 1.234 minutes or 3.14 minutes it is not fixed, see, $\mathbb{Z}$ Real numbers have ∞ no's between them, so there are ∞ possibilities for Probability & Random variable

So, insted of asking a call to take fixed amount of time let's look at them like windows or "interval"

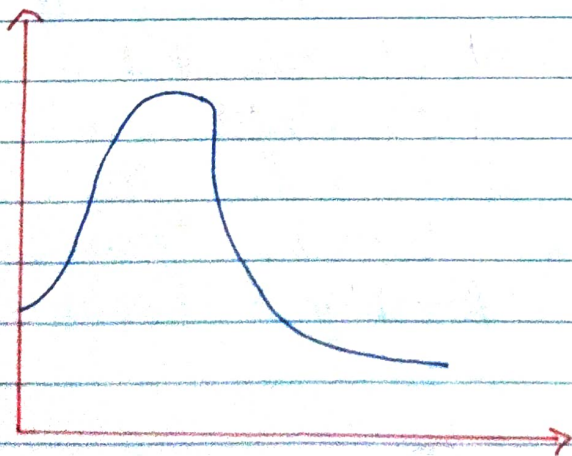
eg say,

what's the probability that the call last's some time's between a certain window of 0 & 1 mins, 1-2 mins, 2-3 mins, 3-4 mins, 4-5 mins. & the call never takes more ~~th~~ than 5 minutes.



& now we ~~th~~ have sorta discrete distribution. & the sum of the height of these bars add to 1.

Now if we make these intervals small like very small like infinitely times, then we get is the graph below & this is how continuous distribution looks like. imagine it as a collection of very very skinny bars.

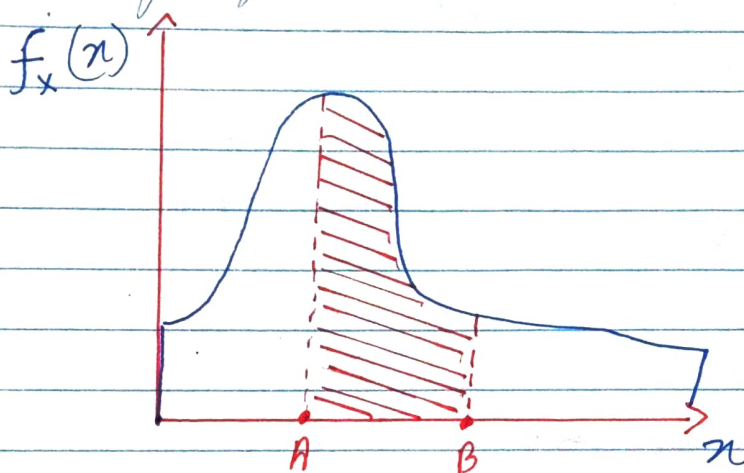


Now, in discrete distribution we know that sum of height's is equal to 1, but in continuous distribution the area under the curve is 1.

Probability Density function ($f_x(x)$)

it's usually denoted by $f_x(x)$

Tells you the rate you accumulate probability around each point. Only defined for continuous variables.



To calculate probabilities, you just need to calculate area under the curve between two points.

$$P(a < x < b) = \text{area under } f_x(x)$$

$f_x(x)$ needs to satisfy:-

- It is defined for all numbers
- It needs to be +ve for all values, $f_x(x) \geq 0$
- The Area Under $f_x(x) = 1$, The area under the curve should be equal to 1 as it is sum of all probabilities.
- The probability that the variable takes on a particular value is always 0, that's why we take intervals $P(X=x) = 0$

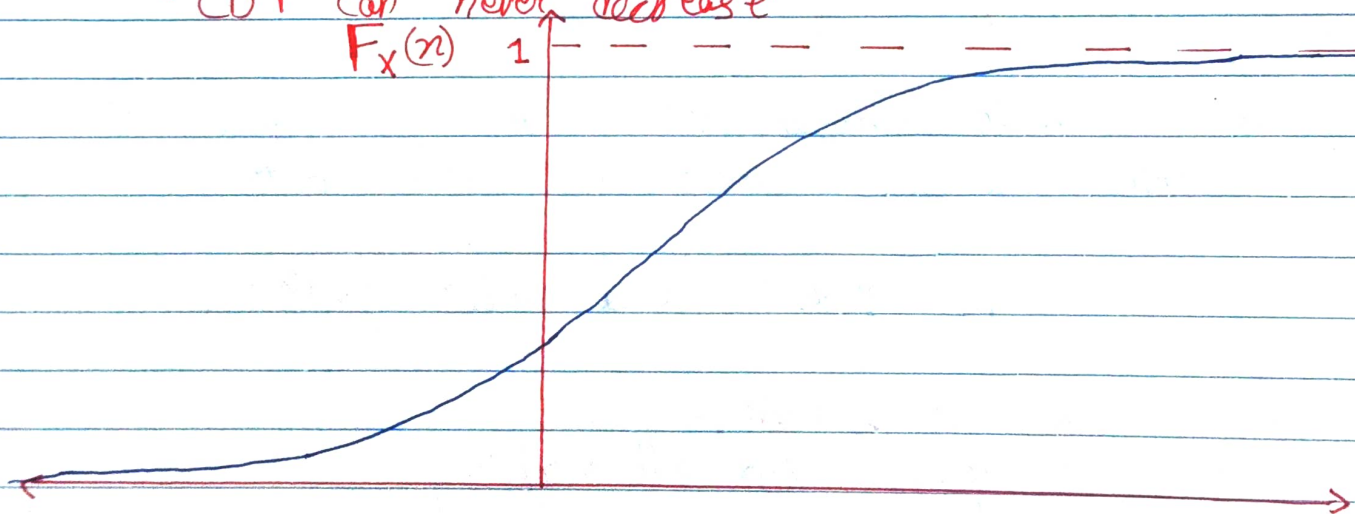
Cumulative Distribution function ($F_X(x)$)

The CDF shows how much probability the variable has accumulated until a certain value.

That means that, " $CDF(x) = P(X \leq x) = F_X(x)$
It is defined for every real no."

Properties

- $0 \leq F_X(x) \leq 1$
- Left "endpoint" is 0
- Right "endpoint" is 1
- CDF can never decrease



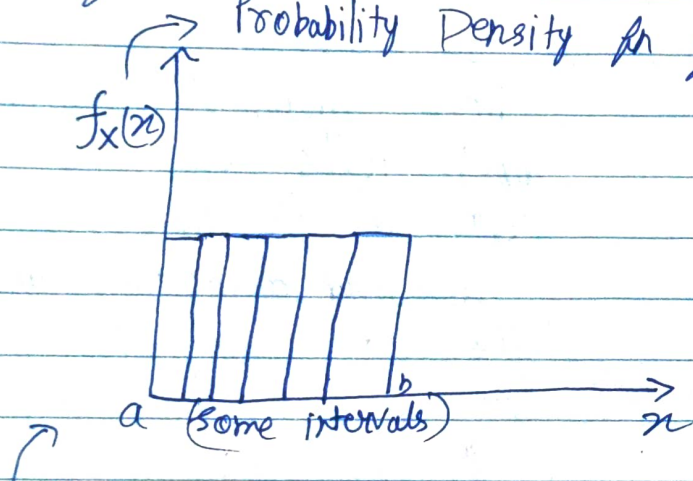
"Cumulative distribution function tells you the probability till a certain point, let's say 2. The height will tell you the probability from 0-2, & if say 3 then the height will tell you the probability from 0-3."

Uniform distribution

⇒ A type of continuous distribution where the probability of occurrence of an event is same throughout the sample space or variable X .

for ex →

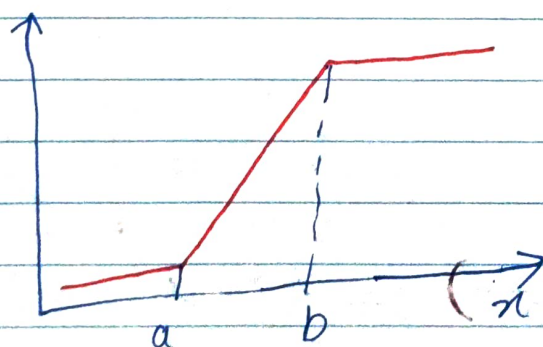
Probability Density fn graph $f_X(x)$



"A continuous random variable can be modeled with a uniform distribution if all possible values lie in an interval & have the same frequency of occurrence."

CDF graph

$F_X(x)$



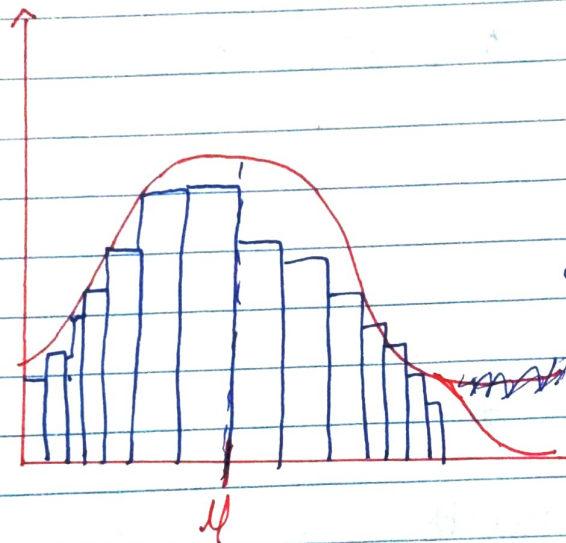
Since, the PDF has uniform distribution, the cumulative distribution fn forms a steep curve

in general

$$F_X'(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & b \leq x \end{cases}$$

Normal distribution (Gaussian distribution)

$\Rightarrow$ if "n" is very large in a Binomial distribution it can be approximated by a gaussian distribution pretty well.



Here mean = μ
which tells us the centre of the data,
 σ = standard deviation, which tells us how wide the data is or thickness of the bell curve.

The equation for gaussian or normal distribution is

$$\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$



$X \sim N(\mu, \sigma^2)$ $\rightarrow$ Here N stands for normal
 $\hookrightarrow$ Here this eqn says that x is distributed as " N " of μ, σ^2

Standard Normal distribution :-

when parameters :-

$$\mu : 0$$

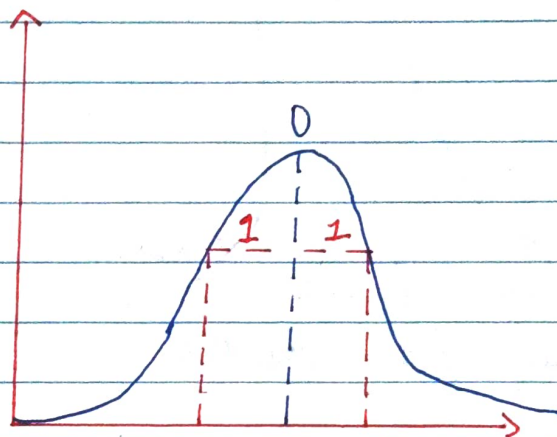
$$\sigma : 1$$

, it's called Standard Normal distribution

using these parameters in the eqn we get,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}$$

$$= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} x^2} \rightarrow \text{eqn (i)}$$



Standardization

There's a really easy way to convert any normal distribution to the standard one.

ex $\Rightarrow$

X distributes normally with
 $\mu = 2, \sigma = 2.5$

$$Z = \frac{X - 2}{2.5}$$

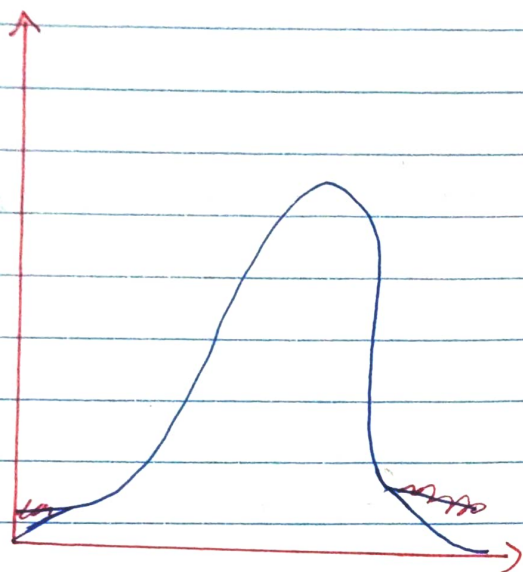
$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

you, can put the value of Z in eqn (i) from above, so

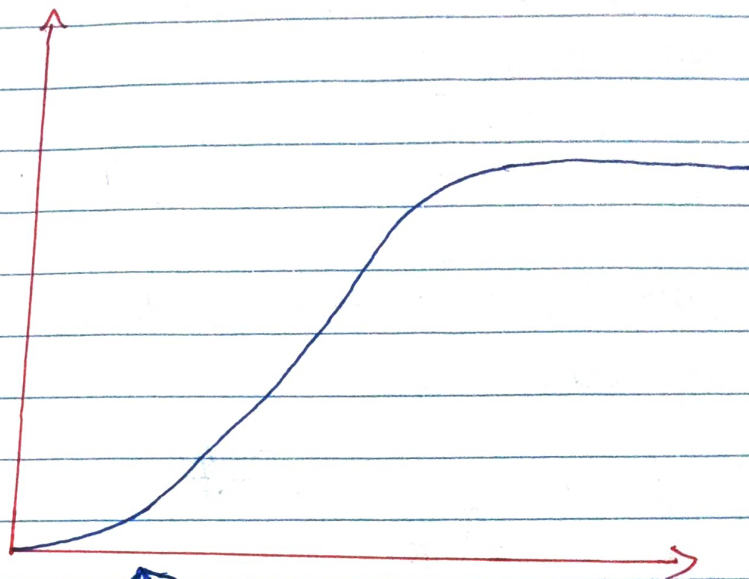
$$\Rightarrow \boxed{\frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(Z)^2}}$$

Standardization is crucial to compare variables of different magnitudes!

what does the CDF Look Like?



Gaussian ~~uniform~~ distribution

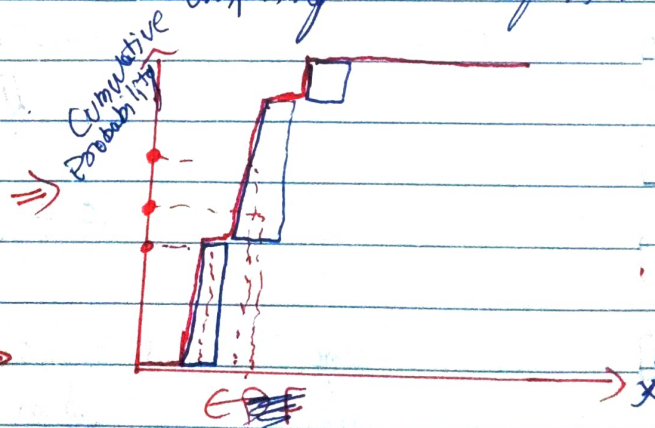
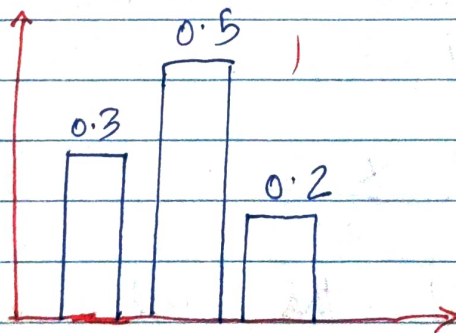


CDF

Sampling from a distribution.

⇒ Imagine you have a dataset but you need it to be larger & for that you can just take synthetic data by constructing a distribution & sampling out of it.

ex ⇒



step ① ∴ Generate a random no. btw 0 & 1.

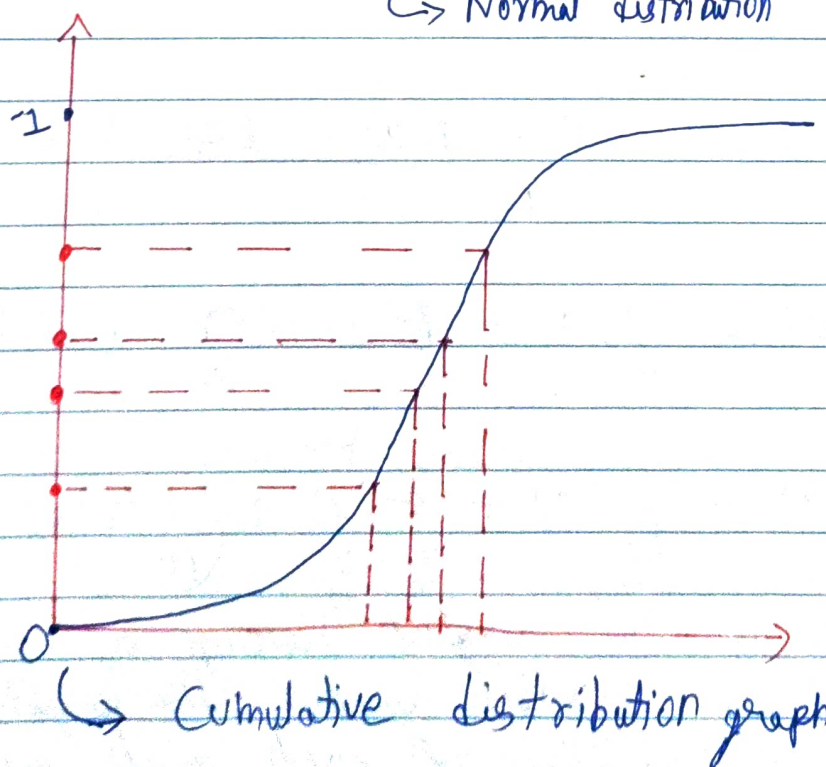
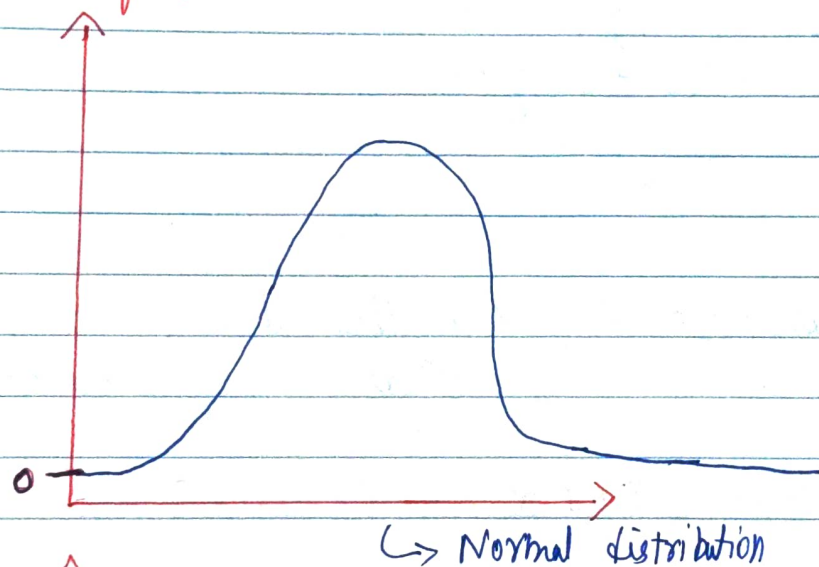
Cumulative Probability Distribution

Step 2 : find out which interval the number belong to

- $[0, 0.3]$
- $[0.3, 0.8)$
- $[0.8, 1]$

Step 3 : Assign an outcome based on the interval

Sampling from a normal distribution.



⇒ If you have a normal distribution then, what you do is draw a cumulative distribution graph for it & then then take the samples, let's say, the red dots on the vertical axis are the randomly generated numbers (0-1) then where they hit the distribution & where they hit's the horizontal axis.

So CDF both for discrete & continuous case is useful when you sample from a distribution.

Describing distributions

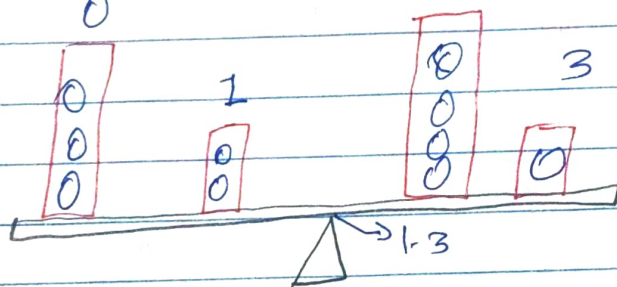
Expected values

Mean & Expected values both are the same term

Mean $\therefore$

It is (intuitively) the point where distribution balances or balancing point

Ex $\Rightarrow$



Mean (mathematically) $\Rightarrow$ Sum the values & divide by total no of values

$$\Rightarrow \frac{0+0+0+1+1+2+2+2+2+3}{10}$$

$$\Rightarrow \frac{3(0) + 2(1) + 4(2) + 1(3)}{10}$$

$$\Rightarrow \frac{3(0)}{10} + \frac{2(1)}{10} + \frac{4(2)}{10} + \frac{1(3)}{10} = 1.3$$

$$\Rightarrow \frac{3}{10}(0) + \frac{2}{10}(1) + \frac{4}{10}(2) + \left(\frac{1}{10}\right)(3) = 1.3$$

This can be the weighted average of the values of the variable where the weights are the probabilities of each each possible value

Expected value = $E[X]$

X = variable representing age of a child

C.g $\Rightarrow$

let's say some one offers you to play a game, ~~and~~ you flip a 3 coins. for each heads you win \$1. what is the amount you will be willing to pay to play the game.

$\Rightarrow$ So we have to find the mean of the data or expected value.

~~$P(H)$~~ $\Rightarrow$ P

~~X = Prob of getting~~

X = no. of heads

So, ~~n~~ $n = 0, 1, 2, 3$

$$\text{when, } (n=0) P(X) = \frac{1}{8}$$

$$\text{when, } (n=1) P(X) = {}^3C_1 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \Rightarrow \frac{3}{8}$$

$$\text{when, } (n=2) P(X) = {}^3C_2 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \Rightarrow \frac{3}{8}$$

$$\text{when, } (n=3) P(X) = \frac{1}{8}$$

So, expected value will be

$$E[X] = \frac{1}{8} \times 0 + \frac{3}{8} \times 1 + \frac{3}{8} \times 2 + \frac{1}{8} \times 3$$

$$E[X] \Rightarrow \frac{3}{8} + \frac{6}{8} + \frac{3}{8} \Rightarrow \frac{12}{8} \Rightarrow \boxed{1.5}$$

So, you should only pay \$1.5 or below.

Expected Value $\therefore$ Discrete Case
in general

X a discrete
Random Variable $\rightarrow$ PMF of X
 $P_X(x) = P(X=x) \Rightarrow$

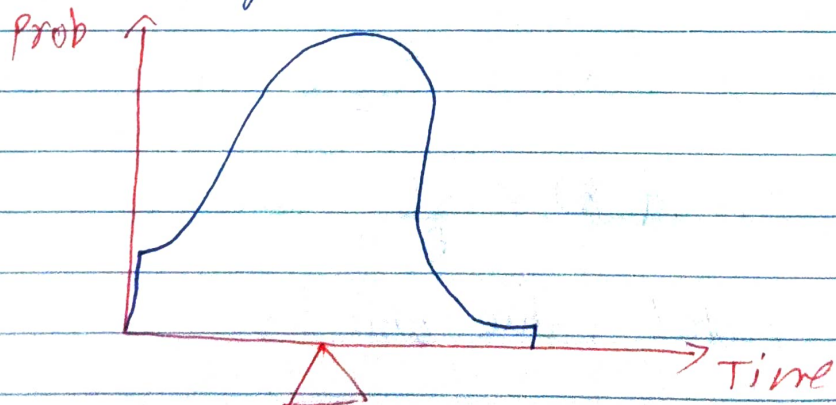
$$E[X] = \sum_x x \underbrace{P_X(x)}_{\text{weighted using probability mass function}}$$

The $E[X]$ expected value here is x time
Probability of x , Summed for all possible values
of x

Expected Value $\therefore$ Continuous Case
Probability density function

$$\int_{-\infty}^{\infty} x \cdot \underbrace{f_X(x)}_{\text{Probability density function}} dx \quad \left. \vphantom{\int_{-\infty}^{\infty}} \right\} \text{Probability using weighted using PDF (Prob Density fn)}$$

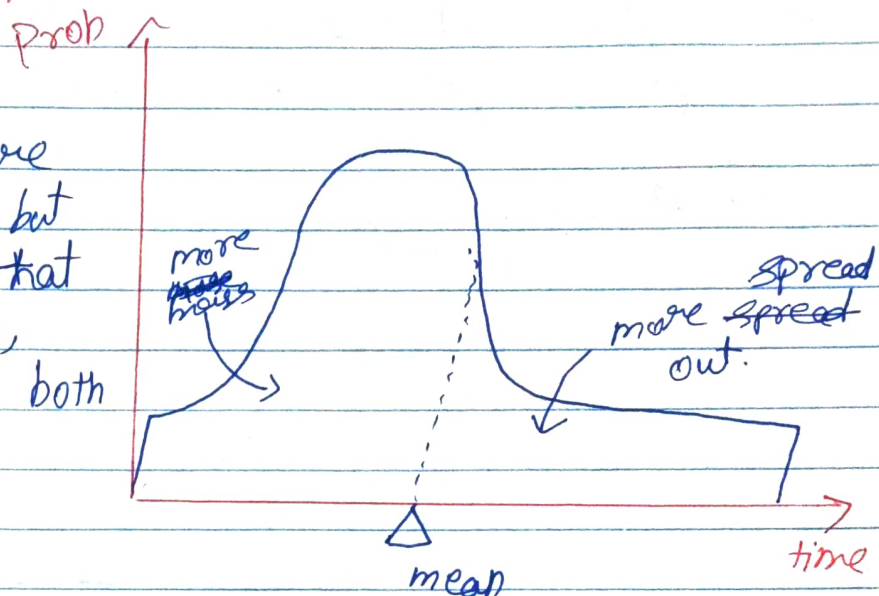
$\hookrightarrow$ The idea is that you are summing the length of
the bars which get thinner & thinner.



$\hookrightarrow$ Here mean is the weighted average of
the Blue probability density function

Expected value: Common misconception

⇒ It might seem natural to think, Mean is a place where data is split in half, but that's not the case. That is actually called median, it depends upon how both mass & how data is spread out.



Other measures of Central tendency: median & mode

Median :-

Imagine you have a lot of data points but there is one data that is super super big & it single handedly inflates the mean.

$$\begin{aligned} \text{ex } \Rightarrow \quad \text{Data } &\Rightarrow 2, 3, 4, 5, 8, 2000 \\ \text{mean } &\Rightarrow \frac{2022}{6} \Rightarrow 337 \end{aligned}$$

from the above example it is super clear that "337 average" value is very deceiving, so what we do we go to "median".

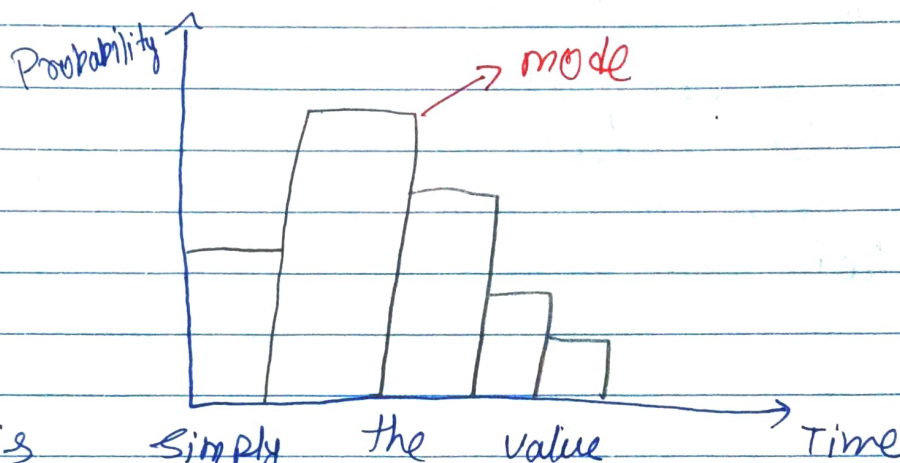
So, what does median do, it stacks up all the given data points in (increasing order) & finds the middle one.

Ex,

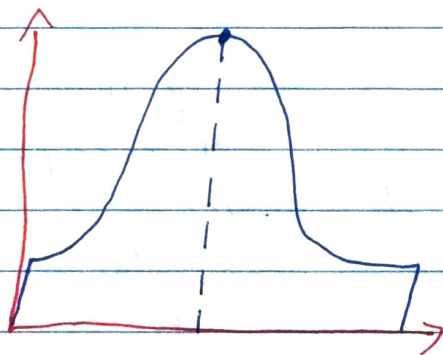
$$\begin{aligned} &\Rightarrow 2, 3, 4, 5, 8, 2000 \Rightarrow \text{The data points are even so we will average out the two centre data points} \\ &\Rightarrow \frac{4+5}{2} \Rightarrow \boxed{4.5} \end{aligned}$$

Now, Notice that the value "2000" doesn't affect the value that determines central tendency. & 4.5 is a pretty good value for central tendency

Mode :-



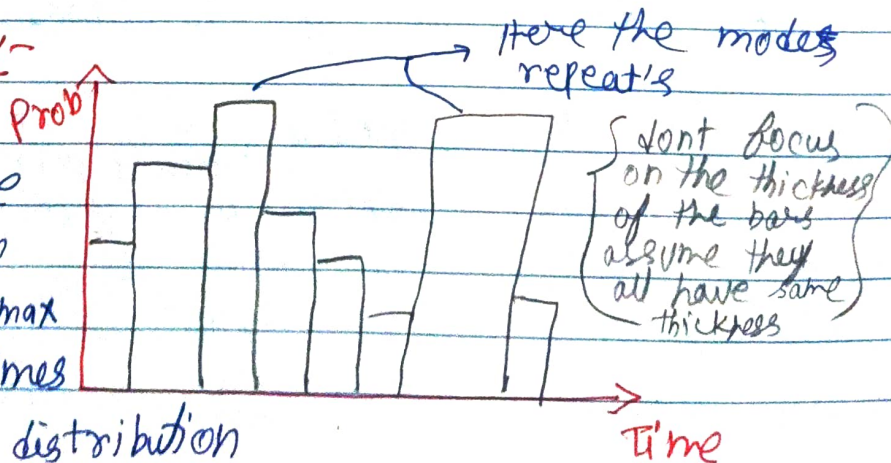
mode is simply the value or interval with the highest probability



For a continuous function mode is the value that has the highest height.

A multimodal distribution:-

The modes may not be unique you maybe in a distribution like on the right where the max value get's repeated many times & that's called a Multimodal distribution



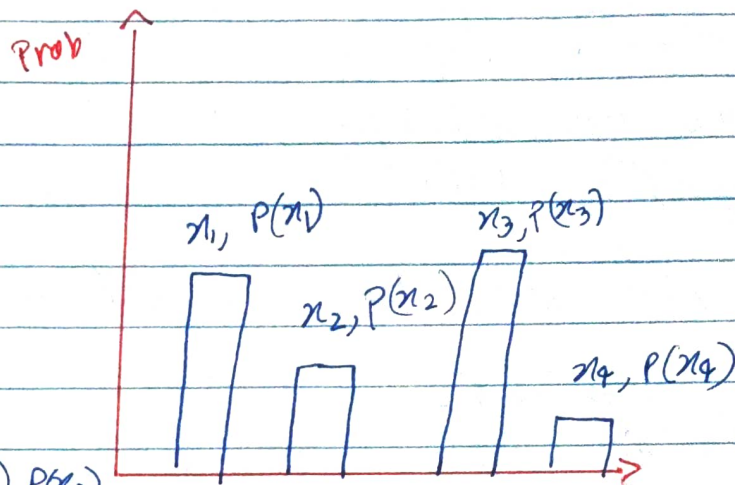
Expected value of a function.

So, the simple expected value will be

$$E[X] = x_1 P(x_1) + x_2 P(x_2) + x_3 P(x_3) + x_4 P(x_4)$$

for any other function $\Rightarrow$

$$E[g(x)] = g(x_1) P(x_1) + g(x_2) P(x_2) + g(x_3) P(x_3) + g(x_4) P(x_4)$$



Ex $\Rightarrow$

You wanna play a game of rolling a dice, everytime you roll a # number you will be given the square of the amount of number. find the amount you should pay to play the game, just find expected value broh!

Solⁿ $\Rightarrow$ using the formula above

$$(1)^2 \cdot \frac{1}{6} + (2)^2 \cdot \frac{1}{6} + (3)^2 \cdot \frac{1}{6} + (4)^2 \cdot \frac{1}{6} + (5)^2 \cdot \frac{1}{6} + (6)^2 \cdot \frac{1}{6}$$

$$\Rightarrow 0.16 + 0.66 + 1.5 + 2.66 + 4.16 + 6$$

$$\Rightarrow \boxed{15.19}$$

takeaways $\therefore$

$$E[ax] = a E[x]$$

$E[b] = b \} \Rightarrow$ Expected value of a constant is the constant itself.

Sum of Expectations

ex $\Rightarrow$ you flip a coin, if heads you win \$1, else you win nothing, Then roll a die. You win the amount you roll.

What are the expected winnings

(H) $\rightarrow \$1$	/	(1) (2) (3) (4) (5) (6)
(T) $\rightarrow \$0$		

expected wining $\Rightarrow$ for coin

$$\Rightarrow E[X] \Rightarrow \frac{1}{2} \times 1 + \frac{1}{2} \times 0$$

$$\Rightarrow \boxed{0.5}$$

$\Rightarrow$ for die $\Rightarrow$

$$\frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6}$$

$$\Rightarrow \boxed{\cancel{21/6}} \quad 21/6 \Rightarrow \boxed{3.5}$$

$$\text{Total expected winnings} = 3.5 + 0.5 \Rightarrow 4$$

So, The takeaway is $\therefore$

$$E[X_1 + X_2] = E[X_1] + E[X_2]$$

$\hookrightarrow$ Expectation of X_1 sum X_2 is expectations of individual X_1 & X_2 's

Variance

Variance tells us how spread out a distribution is

ex $\Rightarrow$ Two distributions can have same expected value but their one of them can be very narrow & the other one can be very wide.

$\Rightarrow$ you play a game with a friend
Heads $\rightarrow$ you win 1 dollar
Tails $\rightarrow$ you lose 1 dollar

$\Rightarrow$ The fair amount to play the game which can be calculated by Expected value is

$$E[X] = \frac{1}{2} \times 1 + \frac{(-1) \times 1}{2} \Rightarrow -0.5 + 0.5 \Rightarrow 0$$

$\Rightarrow$ You play another game with a friend,
Heads $\rightarrow$ you win 100 dollar
Tails $\Rightarrow$ you lose 100 dollar

$\Rightarrow$ The fair amount to play the game can be calculated by expected value is

$$E[X] = \frac{1}{2} \times 100 + \frac{(-100) \times 1}{2} \Rightarrow 50 - 50 \rightarrow 0 \Rightarrow \boxed{0}$$

Both of these games have same expected values but clearly there's a big difference between them. The second one has much larger spread of possible outcomes where you're losing or winning a hundred dollars at a time. If you want to quantify this spread you will have to use variance because Expected values will not work.

Variance is also called expected squared deviation

$$\text{Var}(x) = E[(x - E[x])^2]$$

$E \rightarrow$ it just essentially means that you're finding an average value.

Steps to calculate variance $\Rightarrow$

- ① Find x 's mean $\Rightarrow E[x]$
- ② Find the deviation from that mean for every value of x . $(x - E[x]) \rightarrow$ for every x .
- ③ Square those deviations $(x - E[x])^2 \rightarrow$ for every value of x .
- ④ Average those squared deviations
 $E[(x - E[x])^2]$

Then you get "Average squared deviation"

Variance can also be given by the formula $\therefore$

$$\text{Var}(x) = E[x^2] - E[x]^2$$

Property of Variance $\Rightarrow$

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

b or addition of a constant doesn't matter in Variance.

& If you multiply any constant by x in the variance then the result will be equal to squared of the constant multiplied by variance of x .

a general problem i wanted to skip it but skipping it is making me anxious, so here it is, i am only doing it to get rid of this anxiety.

⇒ Imagine, you play a game you roll a dice, & you win double of whatever you rolled.

However the game costs 5 dollars to play

so	(1)	(2)	(3)	(4)	(5)	(6)
wins	2	4	6	8	10	12
	-5	-5	-5	-5	-5	-5
net win	-3	-1	1	3	5	7

Expected value ⇒

$$\Rightarrow \frac{(-3) \times 1}{6} + \frac{-1 \times 1}{6} + \frac{1 \times 1}{6} + \frac{3 \times 1}{6} + \frac{5 \times 1}{6} + \frac{7 \times 1}{6}$$

$$\Rightarrow \frac{-0.5}{6} + \frac{-0.16}{6} + \frac{0.16}{6} + \frac{0.5}{6} + \frac{0.83}{6}$$

$$\Rightarrow 1.99 \text{ or } \boxed{2} \text{ dollars.}$$

Standard Deviation

$$\text{Var}(X) = E[(X - \mu)^2] = E[X^2] - E[X]^2$$

Say X is measured in meters

Then $E[X]$ is measured in meters

Then $\text{Var}(X)$ is measured in meters²

Then $\sqrt{\text{Var}(X)}$ is measured in meters

Let's call $\text{std}(X) = \sqrt{\text{Var}(X)}$, The standard deviation of X .

Sum of Gaussians: an Example

Ex $\Rightarrow$ imagine you are studying

(R) Total response time of a Computer system

(T) Processing Time

(L) Network Latency

Total response time

$$(R) = T + L$$

Imagine you can model the processing time in milliseconds with a gaussian distribution $T \sim N(10, 2^2)$, & Total latency, you can also measure it in milliseconds also a gaussian with mean 5 & standard deviation 1, $L \sim N(5, 1^2)$

& Suppose these two variables T & L are independent to each other.

Now, $R = T + L$ } is a gaussian too

$$\mu_R = E[R] = E[T+L] = E[T] + E[L]$$

$$= \mu_T + \mu_L = 10$$

$$\text{So } \mu_R = 15$$

$$\sigma_R^2 = \text{Var}(R) = \text{Var}(T+L)$$

$$\Rightarrow \text{Var}(T) + \text{Var}(L)$$

$$\Rightarrow \sigma_T^2 + \sigma_L^2$$

$$\Rightarrow 2^2 + 1^2$$

$$\sigma_R^2 \Rightarrow 5$$

$$\sigma_R = 2.236$$

In general,

if you have a linear combination of variables where X & Y are both Gaussians with parameters μ_x, σ_x & μ_y, σ_y then,

In general,

$$W = aX + bY$$

$$X \sim N(\mu_x, \sigma_x^2)$$

$$Y \sim N(\mu_y, \sigma_y^2)$$

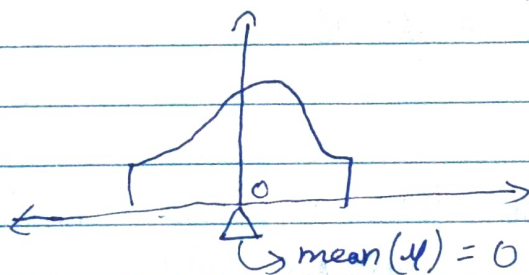
$$\rightarrow W \sim N(a\mu_x + b\mu_y, \frac{a^2\sigma_x^2 + b^2\sigma_y^2}{a^2\sigma_x^2 + b^2\sigma_y^2})$$

Standardizing a distribution

⇒ When you have a distribution with mean (μ) it's always nicer when the distribution has mean zero, so one thing you can do is subtract mean from the (x), to get a new distribution with mean zero

So, New " x " for all x 's is

$$(x - \mu)$$



⇒ & it's also nicer to have a standard deviation of one so what do we do, we simply divide the standard " x ", by the standard deviation to get the standard deviation of (1)
So,

$$x \rightarrow \frac{x}{\sigma}$$

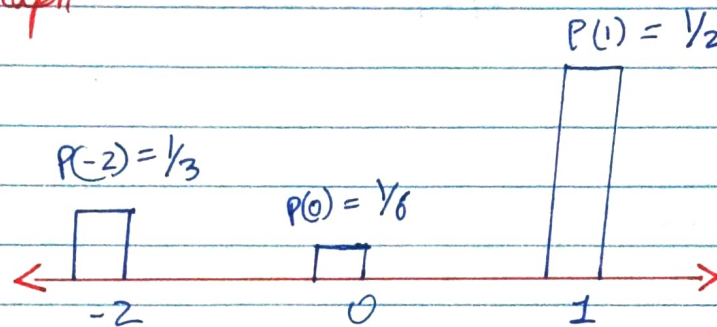
& Now, To standardize it we perform the both given operations given above simultaneously

$$x \rightarrow \left(\frac{x - \mu}{\sigma} \right)$$

Skeuiness & kurtosis

Moments of a Distribution :-

Let's say you have a random variable & its ~~prob~~ probability graph



So expectations,

→ first moment ^{or}

$$E[X] = \frac{1}{3}(-2) + \frac{1}{6}(0) + \frac{1}{2}(1)$$

→ second moment

$$E[X^2] = \frac{1}{3}(-2)^2 + \frac{1}{6}(0)^2 + \frac{1}{2}(1)^2$$

→ Third moment

$$E[X^3] = \frac{1}{3}(-2)^3 + \frac{1}{6}(0)^3 + \frac{1}{2}(1)^3$$

⋮

⋮

⋮

⋮

$$E[X^k] = \frac{1}{3}(-2)^k + \frac{1}{6}(0)^k + \frac{1}{2}(1)^k$$

In general,

$$E[X^k] = \cancel{\frac{1}{n} \sum_{i=1}^n} \quad \sum_{i=1}^n (p_i X_i^k) \quad \sum_{i=1}^n (p_i (x_i)^k)$$

Here $n \Rightarrow$ no. of elements in sample space, or number of elements random variable "X" can take

p_i = Probability of an element "i"

x_i = ~~P~~ The element x_i

k = moment of the distribution, the eqn given on the previous page can also be written as,

$$E[X^k] = p_1 x_1^k + p_2 x_2^k + \dots + p_n x_n^k$$

Skewness

⇒ Let's do it by example, of two games

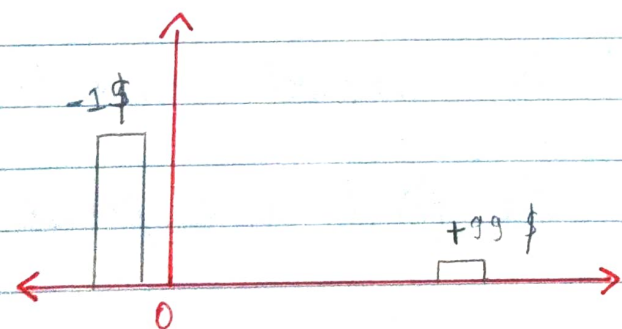
lottery vs Insurance

lottery

if you play lottery, you buy a ticket, worth 1\$ & win jackpot worth 100\$

You win \$99 with 1% prob
You loose \$1 with 99% Prob

graph



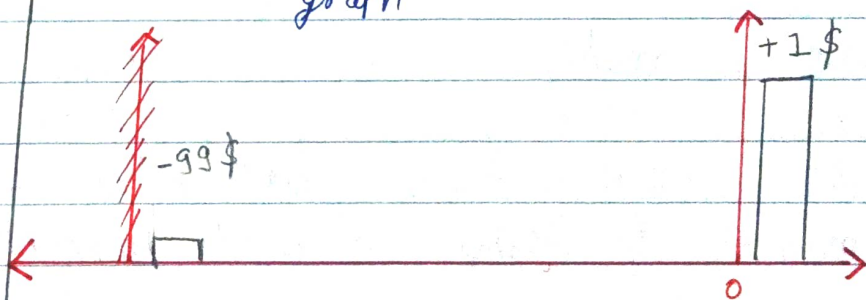
Insurance

Let's say you sell car insurance & the cost of insurance is \$1 & a client buys insurance, the prob. You have to pay 100 insurance money if the client crashes his car.

{ You win \$1 with 99% Prob
You loose \$99 with 1% Prob }

Considering the fact that the client crashes his car with 1% Prob.

graph



⇒ These two scenarios are tremendously different & let's see if that difference can be detected either by expected value or the variance

$$E[X_1] = -1 \cdot 0.99 + 99 \cdot 0.01 = 0 \quad E[X_2] = -99 \cdot 0.01 + 1 \cdot 0.99 = 0$$

These two games has expected values of zero.

So, with expectations we didn't succeed to at telling these two scenarios apart, so let's try "variance".

$$\text{Var}(X_1) = (-1)^2 \times 0.99 + (99)^2 \times 0.01 = 99$$

$$\text{Var}(X_2) = (-99)^2 \times 0.01 + (1)^2 \cdot 0.99 = 99$$

Now, the variance of these two scenarios turns out to be same, so how are we gonna tell the difference btw these two scenarios?

The answer is skewness.

$$\text{Var}(X_1) = E[X_1^2] - \overset{\rightarrow 0}{E[X_1]^2} = 99$$

$$\text{Var}(X_2) = E[X_2^2] - \overset{\rightarrow 0}{E[X_2]^2} = 99$$

Since $E[X_1] = 0$,

$E[X_2] = 0$,

we were only looking at the $E[X_1^2]$ & $E[X_2^2]$, because expectations of X_1 & X_2 are 0.

$$E[X_1^2] = 99$$

$$E[X_2^2] = 99$$

So in reality we were looking at the second moment.

So, these two distributions have the ~~first~~ same first moment & same second moment.

$$E[X_1] = 0$$

$$E[X_1^2] = 99$$

$$E[X_2] = 0$$

$$E[X_2^2] = 99$$

Let's look at the third moment

$$E[X_1^3] = (-1)^3 \times 0.99 + (99)^3 \times 0.01 = 9702$$

$$E[X_2^3] = (1)^3 \times 0.99 + (-99)^3 \times 0.01 = -9702$$

Now, the third moment tells us the difference it tells if you have numbers to the left & numbers to the right.

The lottery scenario has value far right side or +ve side so it is +vely skewed.

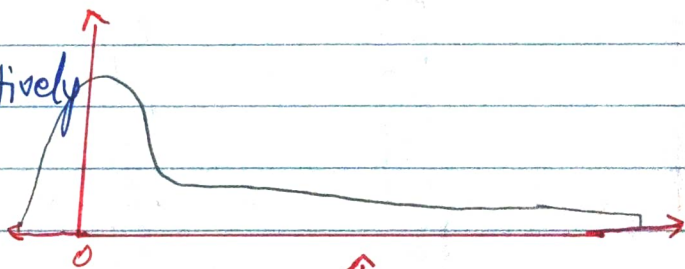
The insurance scenario has values far left or -ve side so it is -vely skewed.

Notice:- that before finding the skewness you have to standardize the distribution.

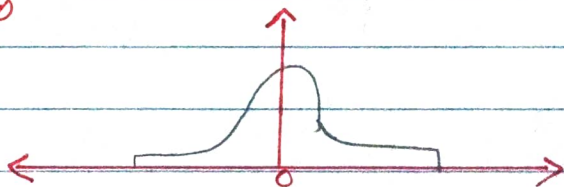
$$\text{Skewness} = E \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right]$$

We have 3 cases in skewness.

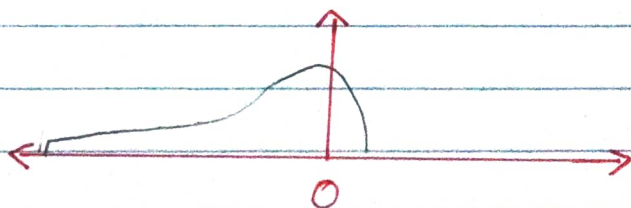
$$E \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right] > 0, \text{ so positively skewed}$$



$$E \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right] = 0, \text{ No skewness}$$



$$E \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right] < 0, \text{ Negatively skewed}$$



Kurtosis

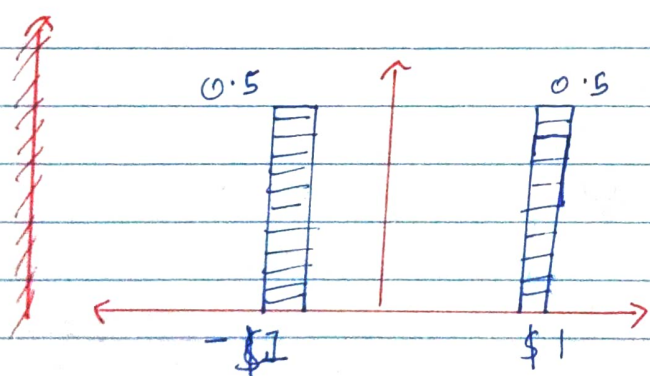
Let's see an example :-

Game - 1

Probability $\frac{1}{2}$: You win 1 dollar

Probability $\frac{1}{2}$: You lose 1 dollar

Graph $\Rightarrow$



Game - 2

Probability $\frac{100}{202}$: You win 10 cents

Probability $\frac{100}{202}$: You lose 10 cents

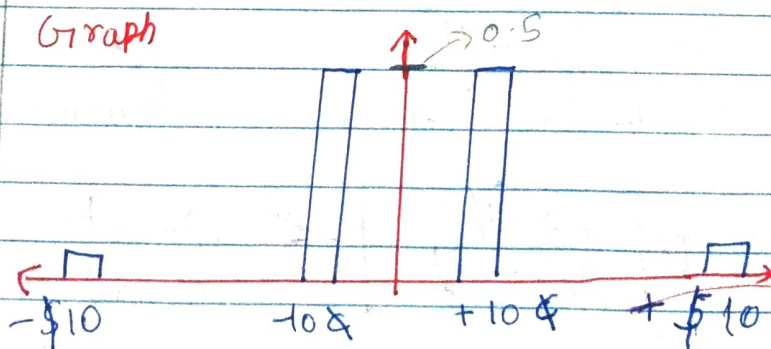
Probability $\frac{1}{202}$: you win 10 dollars

Probability $\frac{1}{202}$: you lose 10 dollars

In this game, there is $\frac{100}{202}$ prob that you win 10 cent's & same if you have to lose 10 cent's (&) but there is also a $\frac{1}{202}$ percent

chance that you will win 10 dollars & you will lose 10 dollars

Graph



So to find the difference between these two, what should we do, expected value?, standard deviation?, skewness? let's check

Game 1

$$E[X_1] = 0$$

$$\text{Var}(X_1) = 1$$

$$\text{Skew}(X_1) = 0$$

Game 2

$$E[X_2] = 0$$

$$\text{Var}(X_2) = 1$$

$$\text{Skew}(X_2) = 0$$

These two graphs have the same expectations, Variance & their Skewness is all the same but we know that one game is riskier than the other. ~~There is~~ There is a significant difference between these two games.

Well one way to describe them is tell that there are values which are really far away from zero. in the game 2 let's see it mathematically,

Let's check the moments:-

$$E[X_1] = 0$$

$$E[X_1^2] = 1$$

$$E[X_1^3] = 0$$

OK, so let's see the fourth moment,

$$E[X_1^4] = \frac{1}{2}(-1)^4 + \frac{1}{2}(1)^4$$

$$\Rightarrow 1$$

$$E[X_2] = 0$$

$$E[X_2^2] = 1$$

$$E[X_2^3] = 0$$

$$E[X_2^4] = \frac{100}{202}(-0.1)^4 + \frac{100}{202}(0.1)^4 + \frac{1}{202}(-10)^4 +$$

$$\frac{1}{202}(10)^4$$

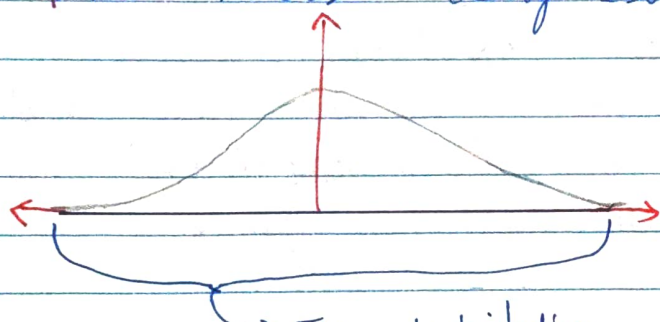
$$\Rightarrow 99.01$$

So, When The distribution has very large numbers very far away from centre, so $E[X^4]$ captures this & that is "kurtosis".

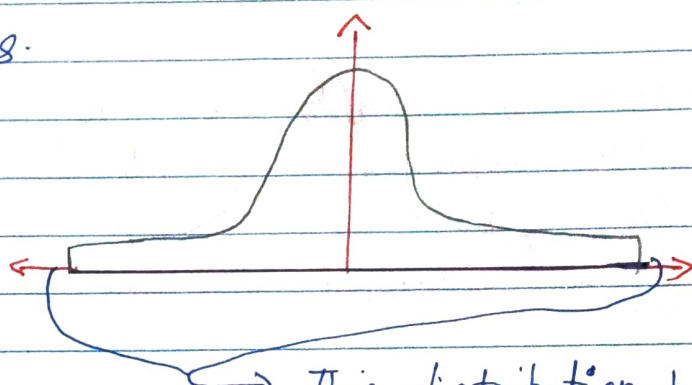
But before doing kurtosis we need to standardize the distribution.

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

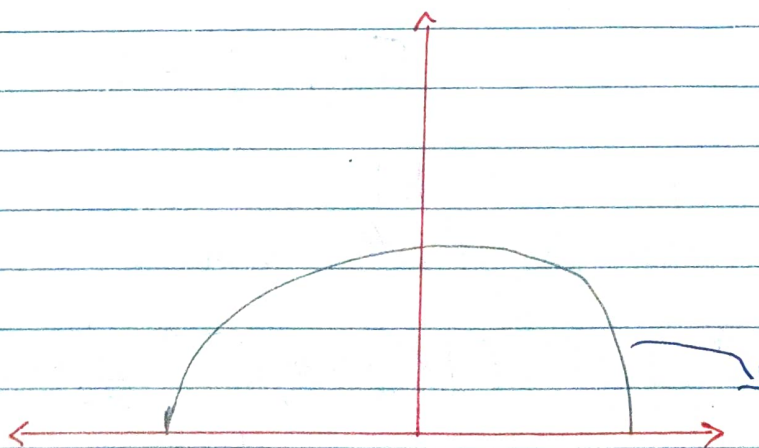
What kurtosis really says.



→ This distribution has thin tails
So, small kurtosis
"Even if they have the same variance"



→ This distribution has thicker tails so, the kurtosis here is bigger "tve kurtosis"



"-ve" kurtosis

distribution with "tve" kurtosis is pointy at the top & distribution with (-ve) kurtosis is round at the top.

Quantiles

ex $\Rightarrow$ let's consider an advertising sales data set & the newspaper ad budget.

18.3	18.4	23.2	51.2
35.2	29.7	75	8.7
65.9	14.2	54.7	25.9

So let's say you want to infer the median from the data set above so you arrange it in increasing order & calculate the median,

median for even number is just the average of the two values in between. The point that splits your data in half

So, $\frac{25.9 + 29.7}{2} \Rightarrow 27.8$

8.7	14.2	18.3	18.4
23.2	25.9	29.7	35.2
51.2	54.7	65.9	75

& 27.8 which is median, is also the 50% quantile or the second quartile.

{ So, what about the data that leaves $\frac{1}{4}$ of your data to the left & $\frac{3}{4}$ to the right? }

The point is

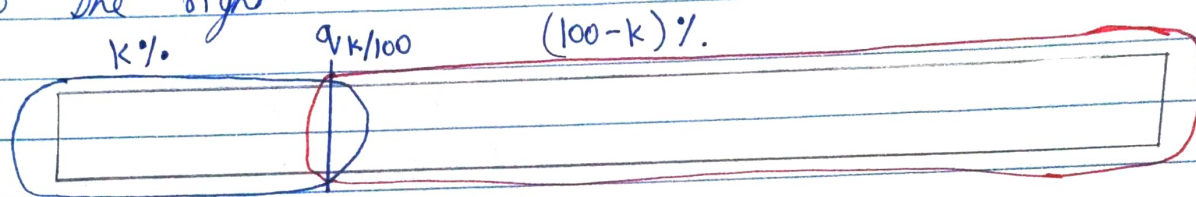
$q_{0.25} = Q_1 = \frac{18.3 + 18.4}{2} = 18.35$

8.7	14.2	18.3	18.4
23.2	25.9	29.7	35.2
51.2	54.7	65.9	75

Here, 18.35 is 25% quantile or first quartile (Q_1)
 $\hookrightarrow (q_{0.25})$

In general, you can do this for any percentage.

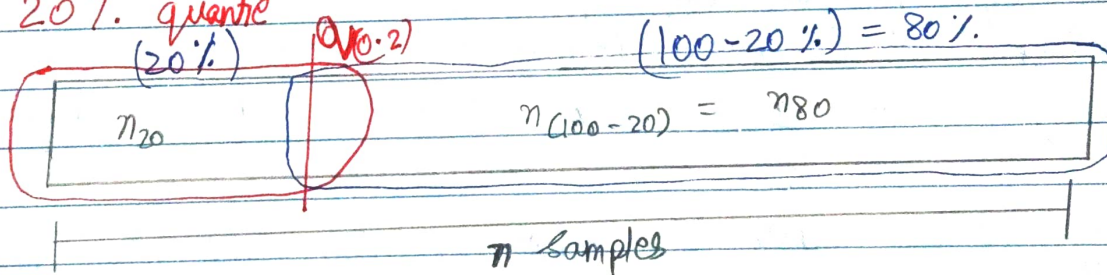
The $k\%$ quantile ($q_{k/100}$) is the value that leaves $k\%$ of your data to the left & $(100-k)\%$ of your data to the right



Some common quantiles:-

- 25% quantile (first quantile - Q_1)
- 50% quantile (median or second quantile - Q_2)
- 75% quantile (third quantile - Q_3)

e.g. $\Rightarrow$ for 20% quantile



$$80, \frac{20}{100} = \frac{n_{20}}{n} \approx P(X \leq q_{0.2})$$

Here, $\frac{n_{20}}{n}$ is The probability that the variable

you're measuring x is below the k percent quantile in this case it is $q_{0.2}$ because that 20%.

$k\%$ quantile ($q_{k/100}$) is the value such that $P(X \leq q_{k/100}) = \frac{k}{100}$
 you can say that the $k\%$ quantile is simply the value such that the probability of the variable being below Q is simply $(\frac{k}{100})$.

Visualizing Data: ~~Box~~ Box plots

To create a Box plot you need to calculate, all the three quartiles (25%, 50%, 75%)

$$Q_1 = q_{0.25} = \frac{18.3 + 18.4}{2} = 18.35$$

$$Q_2 = q_{0.50} = \frac{25.9 + 29.7}{2} = 27.8$$

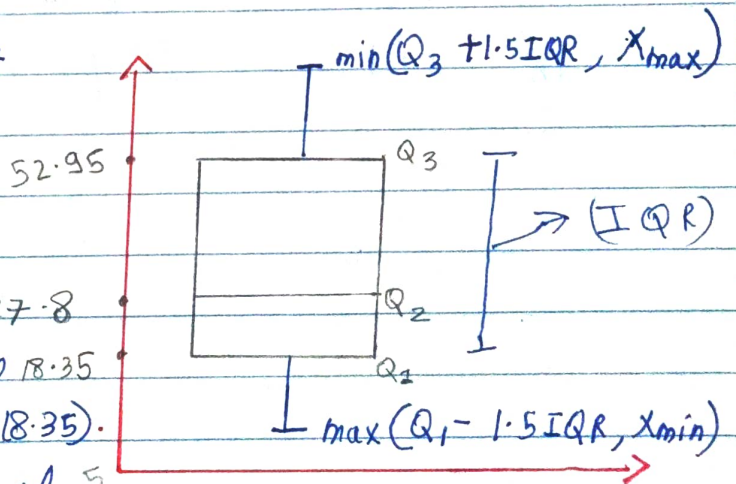
$$Q_3 = q_{0.75} = \frac{51.2 + 54.7}{2} = 52.95$$

Interquartile Range (IQR)

$$IQR = Q_3 - Q_1 = 52.95 - 18.35 = 34.6$$

"That's where 50% of your data lies"

min value of dataset = 8.7
max " " " " = 75



B Drawing the Box plot :-

$\Rightarrow$ It should go from the first to third quartile (Q_3, Q_1) , $(52.95, 18.35)$.

$\Rightarrow$ Then we draw line at the value of 5 the median (Q_2) (27.8).

⇒ Now, we draw the whiskers, draw two lines going from the end of the Box, until $(Q_1 - 1.5(IQR))$, & another one going from $(Q_3 + 1.5(IQR))$

⇒ However the whiskers should never go beyond the max & the minimum values of the data set, so if that happens like in this example simply cut the whisker at those values

What can you tell from this Box plot?

- Data is skewed (because $Q_2 - Q_1$ is way smaller than $Q_3 - Q_2$)
- No outliers meaning there are no extreme points (whiskers were cut at max & min value)
- Analyze dispersion; How dispersed or spread out the data is it can be measured by (IQR)

$$IQR = Q_3 - Q_1,$$

so you're just seeing how far apart Q_3 & Q_1 are, that shows how tightly or loosely the middle half of your data is packed

- if IQR is smaller, your data is concentrated in middle
- if IQR is large, your data is spread out in middle

Kernel density estimation: visualising data ("KDE")

It is used to visualize the "Probability density of function for a distribution".

"KDE" is your best estimate of the PDF, based on ~~on~~

KDE is used when we don't know the real PDF & only have data.

Can we gain any idea how PDF looks like from the Data?

⇒ well you can do that with histograms,

⇒ It represents a density fn

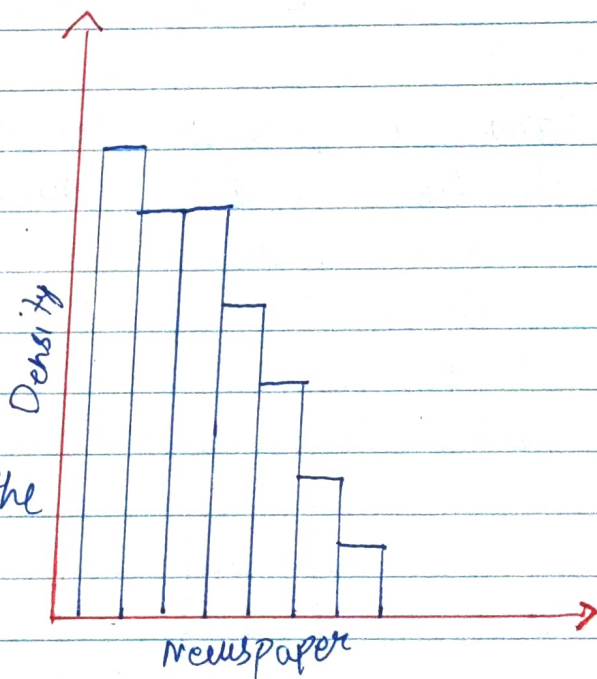
⇒ It is +ve

⇒ Area under the curve is 1.

But ---

⇒ PDFs are usually smooth fn

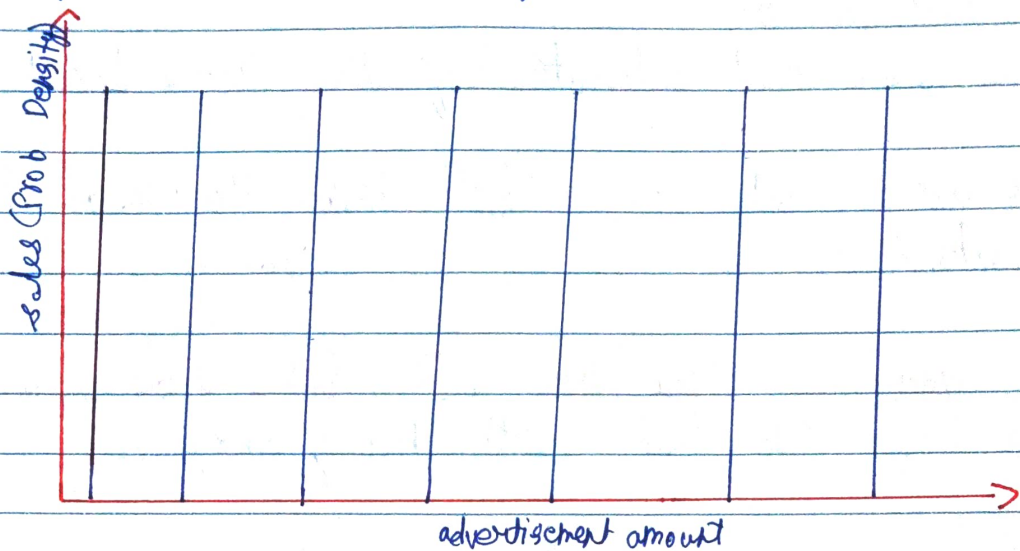
⇒ The discontinuities come from the method & not the data.



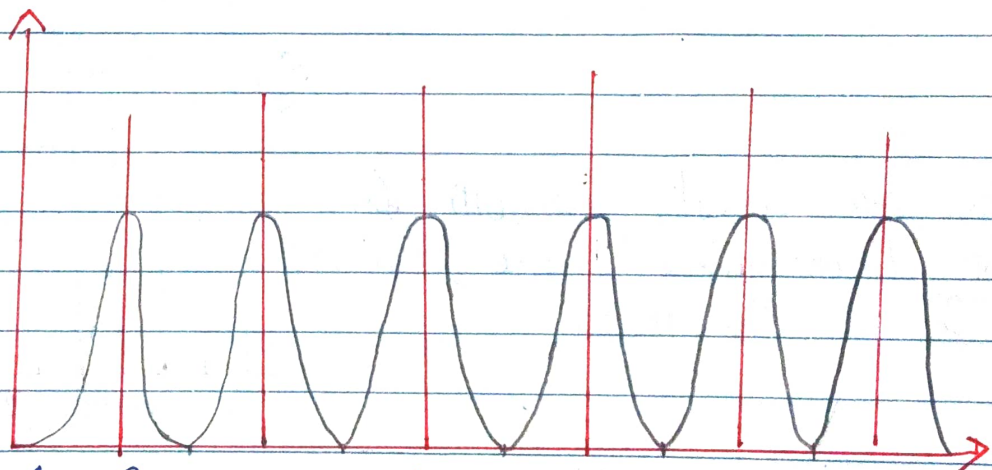
⇒ You can approximate the PDF of your data from this histogram, & the method to do that is kernel density estimation

Here are the steps for kernel density estimation.

① $\Rightarrow$ draw your observations along the x-axis.



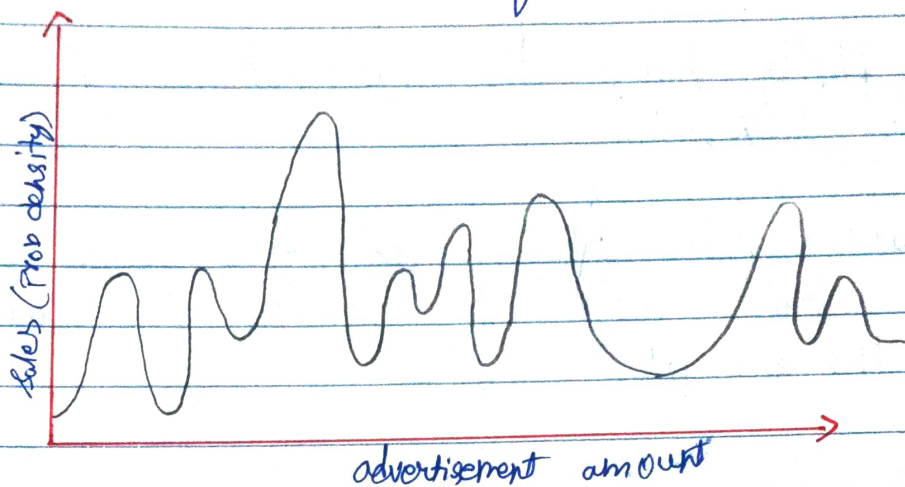
② Draw a gaussian centered at each observation, & that will be called kernel



The value of sigma you choose for a single gaussian density will determine how far the effect of each point spreads it could be thin or it could be wide.

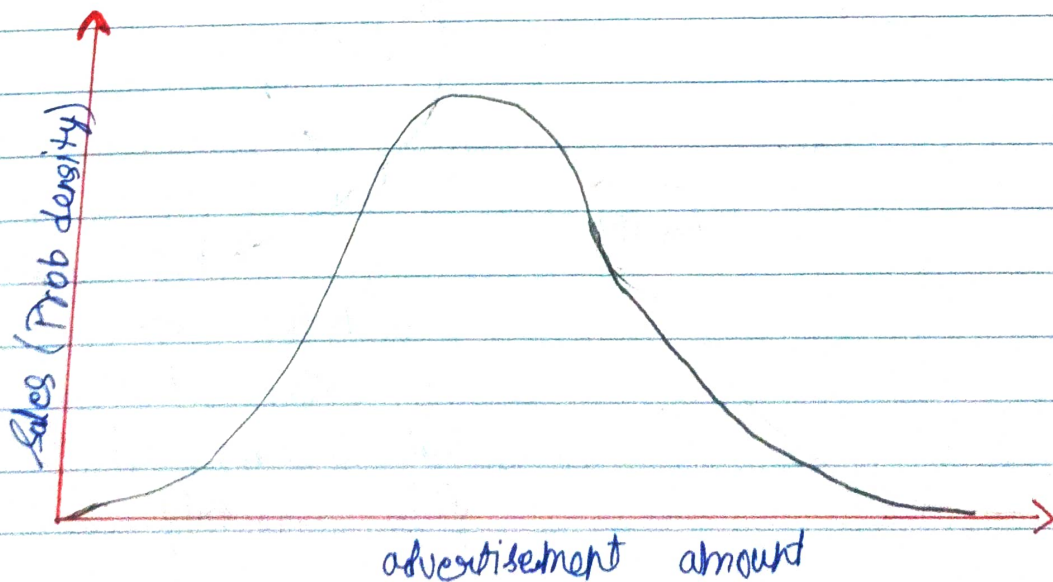
③ Multiply everything by $1/n$ & sum all the blue curves

& the graph will look something like this



⇒ Since the curve you see above represents Probability density then ~~sum~~ area under the curve ~~is~~ should be 1 always

⇒ The plot above doesn't look so great because you tried to plot the PDF from just a few observations from the dataset, to see the full picture PDF fully lets plot all the observations & after doing that it will look smooth.

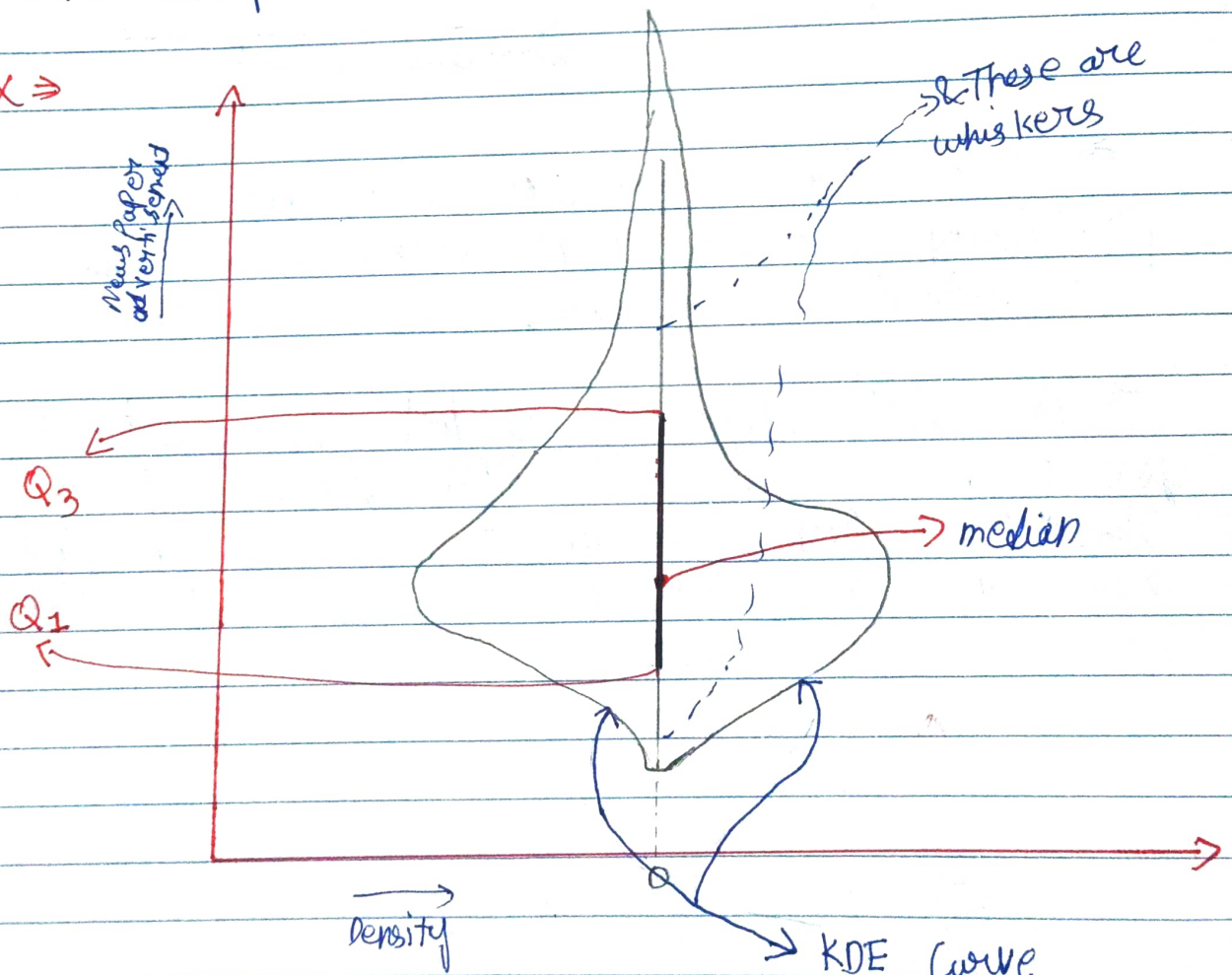


Visualizing data: Violin plot's

⇒ Violin plot's have the information of both "Kernel Density" & Box plot's

⇒ To make violin plot's you just have to take KDE curve turn it 90° to the right & place it by the box plot side

ex ⇒



"Notice both of these curves are same it's just for the fact that i wasn't able to draw them properly"

Visualizing Data :- QQ Plots

Checking for normality in data is frequently done in data sciences,

Some models assume normally distributed data like,

- ⇒ Linear regression
- ⇒ Logistic regression
- ⇒ Gaussian Naive Bayes
- etc,

But how can we tell if a data follows normal distribution or not.

⇒ The QQ plots or "Quantile Quantile" plot helps in telling us if the given data is normally distributed or not

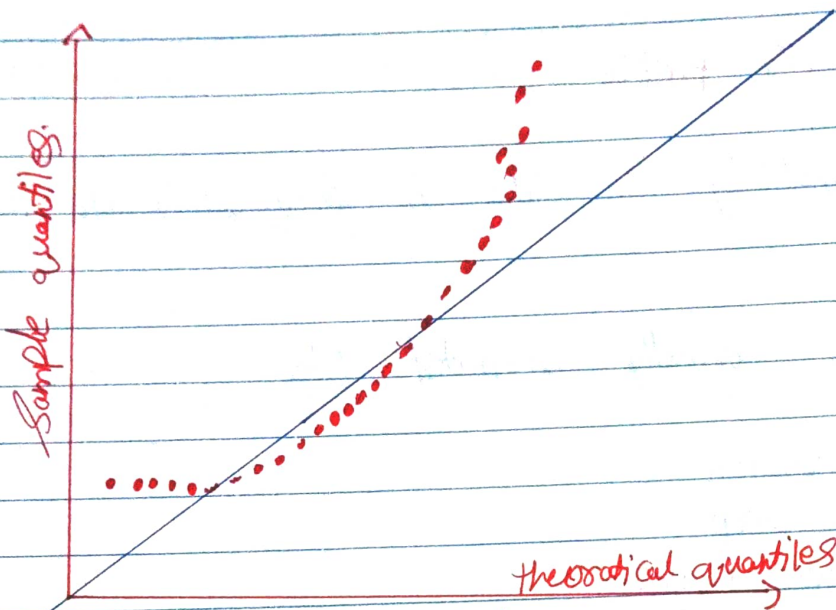
⇒ QQ Plots ("compare quantiles")

Step ① ⇒ Standardize your data
$$\left(\frac{X - \mu}{\sigma} \right)$$

Step ② ⇒ Compute quantiles

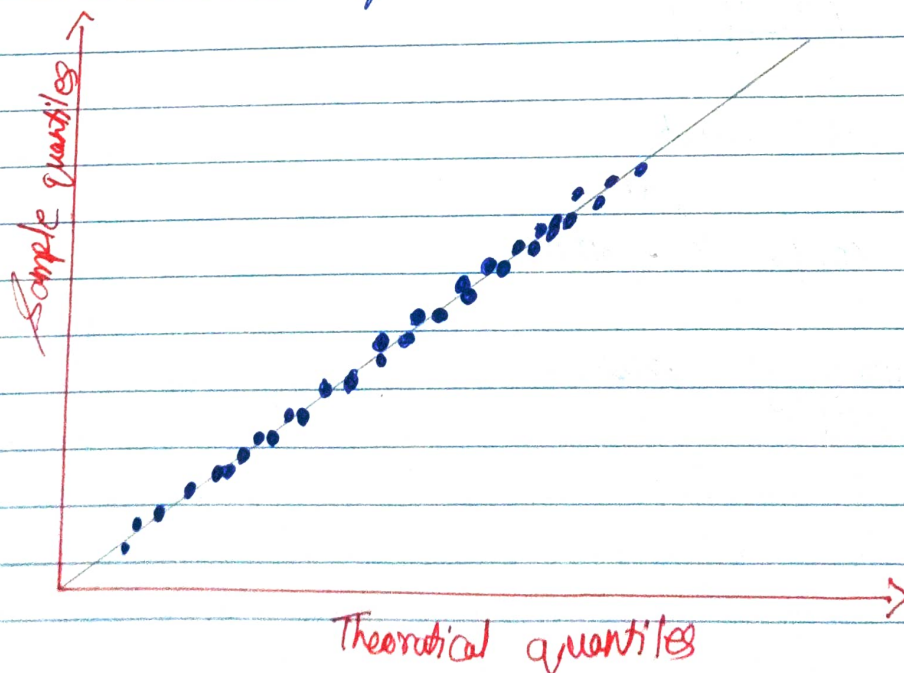
"If you have 10 data points then you compute 10 quantiles".

Step ③ ⇒ Compare to gaussian quantiles



& if it creates data ~~graph~~ a scattered plot,
 "If your data was normally distributed in the upper graph the dots should align with this line, since they are not aligned it's say that data is not normally distributed
 "It shows that the data is skewed.

But if you end up getting a scattered plot like the one below then your data is normally distributed. The dots are aligned more or less with the line



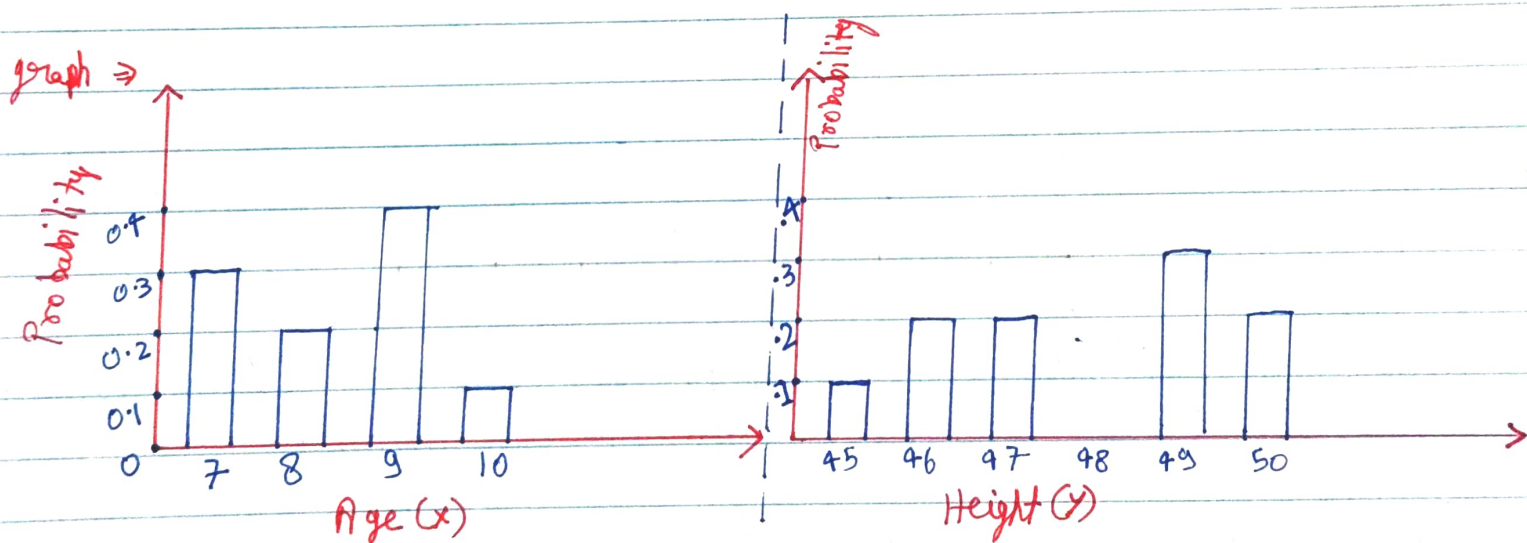
Joint Distribution (Discrete)

⇒ This uses the concept of probability distribution for ~~Two~~ ^{One} variables. & for that let's see,

Ex ⇒ let's say we have two data of children's one for Age & another for height.

Age (year)	Count
7	3
8	2
9	4
10	1

Height (in)	Count
45	1
46	2
47	2
48	0
49	3
50	2



Q ⇒ what is the probability that a child is 9 years old & 49 inches tall?

$$P_{xy}(9, 49) = P(x=9, y=49) = \frac{3}{10} \Rightarrow 0.3$$

More generally the joint probability of two discrete variables is denoted by.

$$P_{xy}(x,y) = P(X=x, Y=y)$$

If we organize the data properly, we can calculate probabilities easily. so let's construct the joint counts for two variables

Age (years)	7	7	7	8	8	9	9	9	9	10
Height (in)	45	46	46	47	47	49	49	49	50	50

	45	46	47	48	49	50
7	1/10	2/10	0	0	0	0
8	0	0	2/10	0	0	0
9	0	0	0	0	3/10	1/10
10	0	0	0	0	0	1/10

Divide the numbers by 10 to get the probability.

& now we have the probability for each possible combination of data "x & y" & it's joint distribution.

So, what is the prob that a child is 8 & 48 inches tall?

⇒ looking at the joint distribution table we can say 0.

$$P_{xy}(8, 48) = P(X=8, Y=48) = 0$$

$$\Rightarrow P_{xy}(7, 46) = P(X=7, Y=46) = \frac{2}{10}$$

Thus for independent discrete random variables:

$$P_{xy}(x,y) = P(X=x, Y=y) = P(x) \cdot P(y)$$

Joint distribution ("Continuous Variables")

Let's say,

X

"waiting time before a
call is picked up"
[0-10] minutes

Y

"Customer satisfaction rating"
[0-10]

"Both variables are
continuous"

$x \Rightarrow 2.4$ minutes

$x \Rightarrow 1.5$ minutes

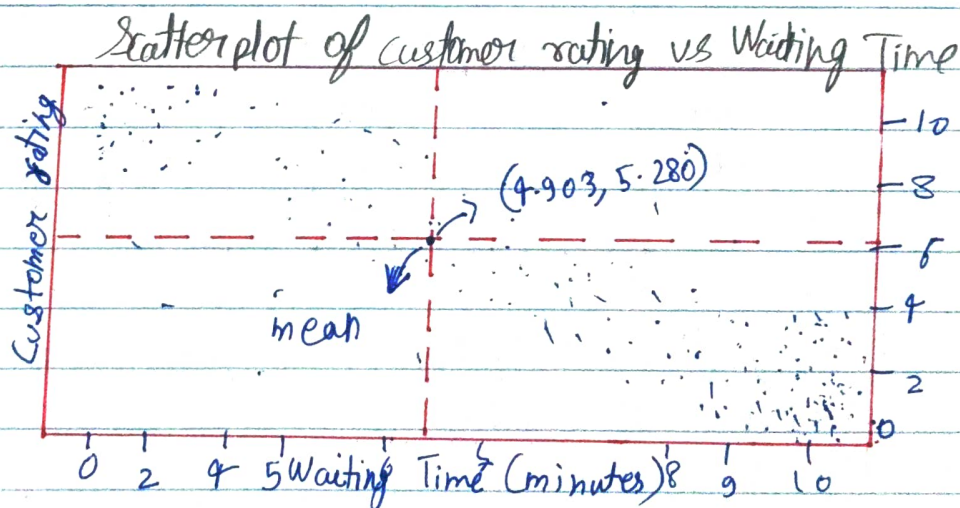
$y \Rightarrow 0.0$

$\Rightarrow 5.7$

Expected value of this
plot.

$$E[X] = 4.903$$

$$E[Y] = 5.280$$



Marginal & Conditional distribution

⇒ Imagine that you have a ~~marginal~~ joint distribution of two variables & from that distribution you want to find out the probability of a single variable, then we do Marginal distribution.

Marginal distribution

Ex ⇒ Joint distribution of ages & height from previous topics

Ages (years) 7 7 7 8 8 9 9 9 9 10
Height (in) 45 46 46 47 47 49 49 49 50 50

		Height (Y)					
		45	46	47	48	49	50
Age (X)	7	$\frac{1}{10}$	$\frac{2}{10}$	0	0	0	0
	8	0	0	$\frac{2}{10}$	0	0	0
	9	0	0	0	0	$\frac{3}{10}$	$\frac{1}{10}$
	10	0	0	0	0	0	$\frac{1}{10}$

⇒ Marginal distribution is a distribution of one variable while ignoring the other.

⇒ So to find the marginal distribution of height sum the joint probability distribution over all values of age.

for height, 45 ⇒ $\frac{1}{10} + 0 + 0 + 0 \Rightarrow \frac{1}{10}$ 50 ⇒ $0 + 0 + \frac{1}{10} + \frac{1}{10} \Rightarrow \frac{2}{10}$
46 ⇒ $\frac{2}{10} + 0 + 0 + 0 \Rightarrow \frac{2}{10}$ 49 ⇒ $0 + 0 + \frac{3}{10} + 0 \Rightarrow \frac{3}{10}$

In general we can say :-

$$P_y(y_j) = \sum_i P_{xy}(x_i, y_j)$$

So, to find the marginal distribution of Age sum the joint probability distribution over all values of age

$$\text{for age, } 7 = \frac{1}{10} + \frac{2}{10} + 0 + 0 + 0 + 0 \quad | \quad 8 = 0 + 0 + \frac{2}{10} + 0 + 0 + 0$$

$$= \frac{3}{10} \quad \quad \quad = \frac{2}{10}$$

$$P_x(x_i) = \sum_j P_{xy}(x_i, y_j)$$

Conditional distribution :-

⇒ Taking the dataset of Age's & Height

7 7 7 8 8 9 9 9 9 10
45 46 46 47 47 49 49 49 50 50

⇒ Here we want observe one variable given that the value of the other one is known.

ex⇒ If age = 9, what is the distribution across height ~~height~~ variable?

⇒ all you have to do is take a slice take the 9 variable & leave rest of them.

	45	46	47	48	49	50
7	$\frac{1}{10}$	$\frac{2}{10}$	0	0	0	0
8	0	0	$\frac{2}{10}$	0	0	0
9	0	0	0	0	$\frac{3}{10}$	$\frac{1}{10}$
10	0	0	0	0	0	$\frac{1}{10}$

		Height (y)					
		45	46	47	48	49	50
Age (x)	9	0	0	0	0	3/10	1/10

Here,

$$P_{Y|X=9}(y) = P(Y=y | X=9)$$

However, The given data in the table above is not accurate because it isn't normalized, so to normalize it we need to divide it by the sum of values in this case it is $0+0+0+0+3/10+1/10 = 4/10$. Because we need the probabilities to sum up to 1.0, That's why. So normal distribution will look like

$$\Rightarrow 0/4/10, 0/4/10, 0/4/10, 0/4/10, (3/10)/4/10, (1/10)/4/10$$

↓

		Height (y)					
		45	46	47	48	49	50
Age (x)	9	0	0	0	0	3/4	1/4

In general, for Discrete

$$P_{Y|X=n}(y) = \frac{P_{XY}(n, y)}{P_X(n)}$$

Joint PDF of x & y

Conditional PDF of Y

Marginal distribution of X

This is done to normalise

For Continuous,

$$f_{Y|X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)}$$

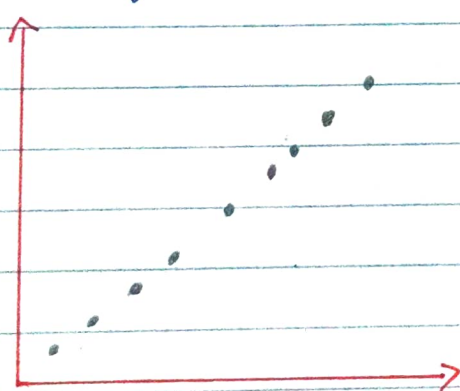
Conditional PDF of Y

Joint PDF of X & Y

marginal distribution of X.

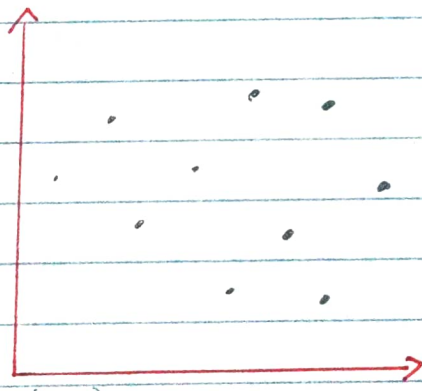
Covariance

Covariance tells us the relation between two variables.
meaning $\Rightarrow$ if X is increasing then Y is increasing, decreasing or behaving random & if X is decreasing then Y is increasing, decreasing or random.



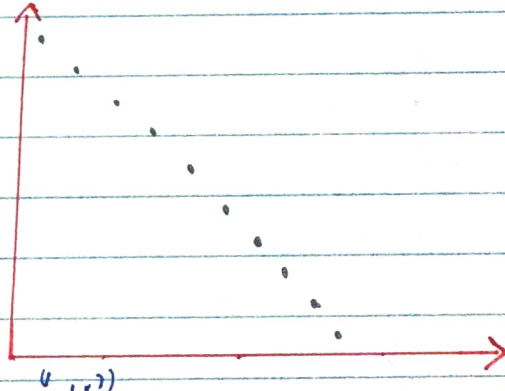
"Y" is increasing with increase in "X".

Covariance > 0



"Y" is behaving randomly with increase in "X".

Covariance ≈ 0



"Y" is decreasing with increase in "X".

Covariance < 0

The formula for covariance is given by :-

$$\text{COV}(X,Y) = \sum XY \Rightarrow \text{Standardizing} \Rightarrow$$

$$\Rightarrow \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n} \rightarrow \text{final formula} \left\{ \begin{array}{l} \text{Not taking average} \\ \text{applied only when probabilities are same.} \end{array} \right.$$

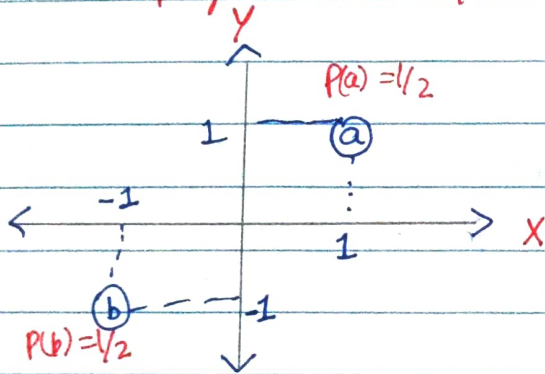
Covariance of a probability distribution:-

Consider X & Y are playing a game to either win or lose a dollar.

Game 1

a: Both players win \$1 each

b: both players lose \$1 each



$$E[X_1] = 0$$

$$E[Y_1] = 0$$

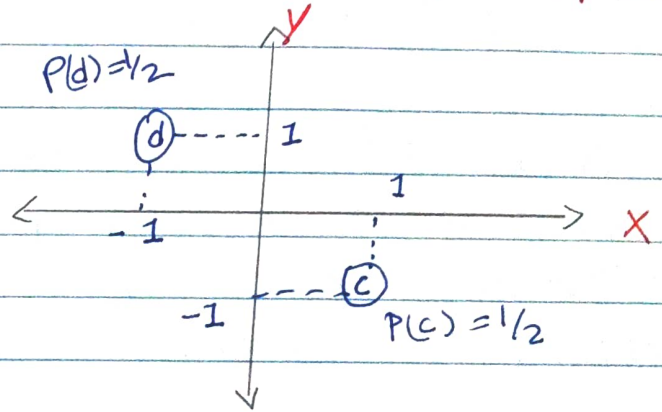
$$\text{Var}(X_1) = 1$$

$$\text{Var}(Y_1) = 1$$

Game 2

c: X wins \$1 & Y loses \$1

d: X loses \$1 & Y wins \$1

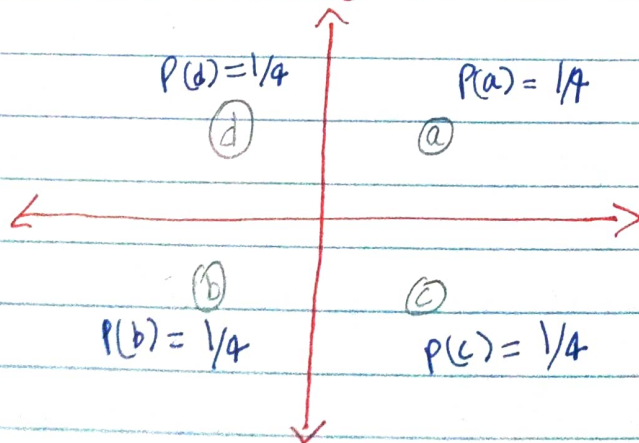


$$E[X_2] = 0, E[Y_2] = 0$$

$$\text{Var}(X_2) = 1$$

$$\text{Var}(Y_2) = 1$$

Game 3



$$E[X_3] = 0, E[Y_3] = 0$$

$$\text{Var}(X_3) = 1$$

$$\text{Var}(Y_3) = 1$$

⇒ How similar are these 3 games for player X & player Y?

These games are same in expectancy for (X & Y) variables & Variance of X & Y variable.

But Obviously looking at these 3 games, you can tell these games are different.

& you can tell the difference if you look at both Players at same time.

Game 1,
$$\text{Cov}(x, y) = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n}$$

$$\text{Cov}(x, y) = 1$$

$\text{Cov} > 0$ meaning with increase in X, Y increases.

So, the more player 1 wins the more player 2 wins

Game 2,

$$\text{Cov}(x, y) = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n}$$

$$\text{Cov}(x, y) = -1$$

$\text{Cov} < 0$ meaning with increase in X, Y decreases

So, either one is happy or other one is sad.

Game 3,
$$\text{Cov}(x, y) = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n}$$

$$\text{Cov}(x, y) = 0$$

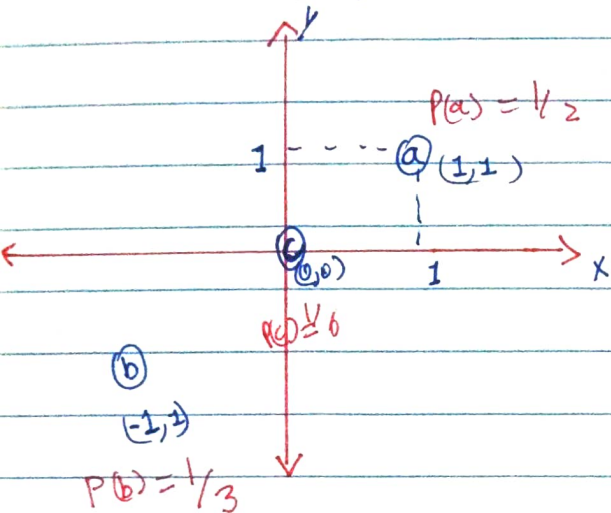
$\text{Cov} = 0$, So there is no pattern among these

Game 4, In game 4, we have different prob for each cases.

a: Both players win \$1 each $P(a) = 1/2$

b: Both Players loose \$1 each $P(b) = 1/3$

c: Neither players wins or lose anything $P(c) = 1/6$



$$E[X_4] = \frac{1}{2} \times 1 + 0 \times \frac{1}{6} + -1 \times \frac{1}{3}$$

$$\Rightarrow \frac{1}{2} - \frac{1}{3} \Rightarrow \boxed{\frac{1}{6}}$$

$$E[Y_4] = \frac{1}{6}$$

$$\text{Var}(X_4) = \sum_{i=1}^n (x_i - \mu_x)^2 \cdot P(x_i)$$

$$\text{Var}(X_4) = \frac{1}{2} \left(1 - \frac{1}{6}\right)^2 + \frac{1}{6} \left(0 - \frac{1}{6}\right)^2 + \frac{1}{3} \left(-1 - \frac{1}{6}\right)^2$$

Similarly,

$$\text{Var}(Y_4) = 0.806$$

$$\boxed{\begin{aligned} \text{Var}(X_4) &= 0.806 \\ \text{Var}(Y_4) &= 0.806 \end{aligned}}$$

$$\text{Cov}(x, y) = \sum_n (x_i - \mu_x) (y_i - \mu_y)$$

The above formula is for equal probabilities to get for Unequal probabilities, Just multiply by the probabilities, So, we don't need (x_i) , so

$$\text{Cov}(x, y) = \sum_n P_{xy}(x_i, y_i) \cdot (x_i - \mu_x) (y_i - \mu_y) \quad - (1)$$

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

So, according to (i),

$$\text{Cov}(X, Y) = \frac{1}{2} (1 - \frac{1}{6})^2 + \frac{1}{6} (0 - \frac{1}{6})^2 + \frac{1}{3} ((-1) - \frac{1}{6})^2$$

$$\boxed{\text{Cov}(X, Y) = 0.806}, \text{ so with increase in } X, Y \text{ increases.}$$

Covariance Matrix

Imagine you have a very large dataset. Let's say of 5 variables, then you have to find the Variance & Covariance & to put it in the matrix.

⇒ It has variance in the Diagonal & Covariance in Anti-diagonal

$$\begin{array}{cc} & \begin{matrix} X & Y \end{matrix} \\ \begin{matrix} X \\ Y \end{matrix} & \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}(Y) \end{bmatrix} \end{array}$$

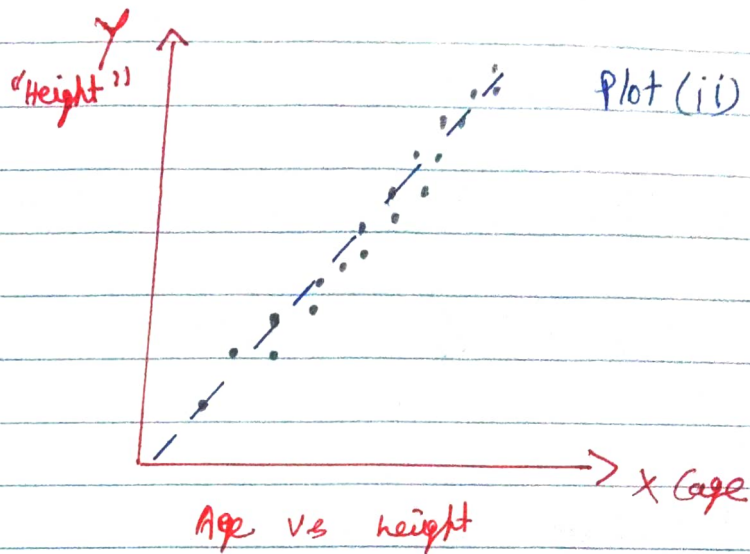
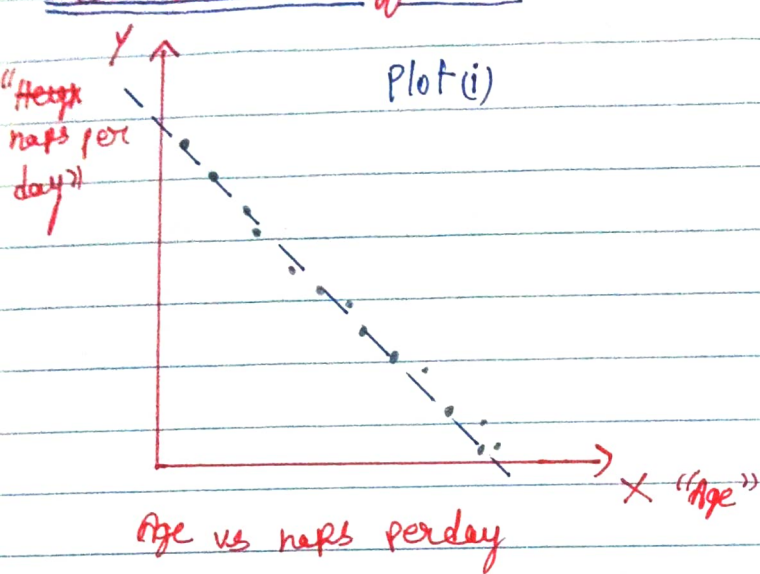
for more than 2 variables.

$$\Sigma = \begin{array}{ccc} & \begin{matrix} X & Y & Z \end{matrix} \\ \begin{matrix} X \\ Y \\ Z \end{matrix} & \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) & \text{Cov}(X, Z) \\ \text{Cov}(X, Y) & \text{Var}(Y) & \text{Cov}(Y, Z) \\ \text{Cov}(X, Z) & \text{Cov}(Y, Z) & \text{Var}(Z) \end{bmatrix} \end{array}$$

Here, Sigma means Covariance matrix

"Covariance matrix"

Correlation Coefficient



$$\text{Var}(x) = 9.17$$

$$\text{Var}(y) = 7.57$$

$$\text{Cov}(x, y) = -7.45$$

We know how the covariance tells us how two variables are related,

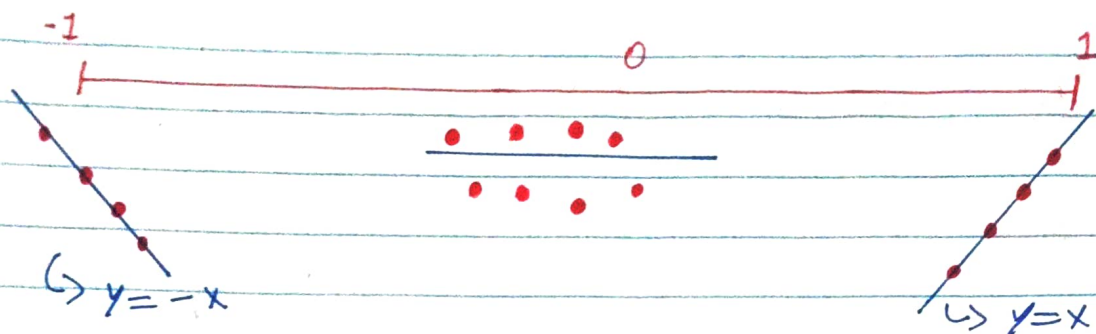
but in the two plots above notice that one number is bigger than the other number forget about the +ve or -ve signs,

so does that mean the correlation is stronger for the one being 17 & weaker for being -7.45, meaning when x grows y grows more or something,

but it's not true so how can you tell the actual correlation between the two,

& here correlation coefficient comes into play,

The correlation coefficient is the number that lies between -1 & 1,



$$\text{Correlation Coefficient} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}}$$

Correlation coefficient is just standardized covariance, meaning dividing covariance by standard deviation of both X & Y .

for, plot (i)

$$\text{The Correlation Coefficient} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}}$$

$$= \frac{-7.45}{\sqrt{9.17} \cdot \sqrt{7.57}}$$

$$= \boxed{-0.894}$$

for, plot (ii)

$$\text{The Correlation Coefficient} = \boxed{0.893}$$

So, from the correlation coefficient of both plot (i) & (ii), we can say that, for both plots the increase & decrease are happening in quite the same way, correlation coefficient is useful in statistics when you're comparing pairs of variables.

Multivariate Gaussian Distribution.

Two variables

H: Height of an adult in inches, $H \sim N(\mu_H, \sigma_H)$

W: Weight of an adult in pounds. $W \sim N(\mu_W, \sigma_W)$

If W, H were independent

$$f_{HW}(h, w) = f_H(h) \cdot f_W(w)$$

if the two variables were independent then the joint PDF would be the product of the marginal PDFs.

$$\Rightarrow \frac{1}{\sqrt{2\pi} \cdot \sigma_H} e^{-\frac{1}{2} \frac{(h - \mu_H)^2}{\sigma_H^2}} \cdot \frac{1}{\sqrt{2\pi} \cdot \sigma_W} e^{-\frac{1}{2} \frac{(w - \mu_W)^2}{\sigma_W^2}}$$

$$\Rightarrow \frac{1}{2\pi \sigma_H \sigma_W} e^{-\frac{1}{2} \left(\frac{(h - \mu_H)^2}{\sigma_H^2} + \frac{(w - \mu_W)^2}{\sigma_W^2} \right)}$$

$$\Rightarrow \frac{1}{2\pi \sigma_H \sigma_W} e^{-\frac{1}{2} \left(\frac{(h - \mu_H)^2}{\sigma_H^2} + \frac{(w - \mu_W)^2}{\sigma_W^2} \right)}$$

$\Rightarrow$ In the Independent case the distribution is perfectly symmetric & in the Dependent case you see that the distribution is not symmetric which is also the True case.

& the thing causing the deformation here is Covariance btw two variables

The only difference between Vari Dependent & Independent cases is that in Independent there is no Covariance in dependent there is covariance.

$$\Sigma = \begin{matrix} \text{In dependent} \\ \begin{bmatrix} \text{Var}(x) & 0 \\ 0 & \text{Var}(y) \end{bmatrix} \end{matrix} \quad \Sigma = \begin{matrix} \text{Dependent} \\ \begin{bmatrix} \text{Var}(x) & \text{Cov}(xy) \\ \text{Cov}(xy) & \text{Var}(y) \end{bmatrix} \end{matrix}$$

In general, Multivariate gaussian looks like -:

$$f_X(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2} \cdot |\Sigma|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu)^T \cdot (\Sigma^{-1}) \cdot (x-\mu)}$$

Σ = Covariance Matrix "Spread of the bell"

$|\Sigma|$ = determinant of the covariance matrix

$$x = [x_1 \ x_2 \ \dots \ x_n]$$

$$\mu = [\mu_1 \ \mu_2 \ \dots \ \mu_n] = \text{mean Vector}$$

Σ^{-1} = Inverse of the matrix $\Rightarrow$ it is multiplied to standardize & rescale the variances & Co-variances

$\Rightarrow$ while dealing with multivariable distribution's all scalar values are replaced with vector & scalar variances are replaced with covariance matrix.

"The video is there in downloads if you don't understand it watch the video it will be clear because i know where the doubt can occur."